MSc Dorien Huysmans

Ambulant Stress Detection in Patients with Stress Complaints

Master's Thesis

to achieve the university degree of

Master of Science

Information and Computer Engineering

Graz University of Technology

Supervisor: Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Institute of Interactive Systems and Data Science

Graz, February 2017

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Graz, 09.02.2017 Dorien Huysmans

Date Signature

Preface

This master's thesis marks the finishing line of two fantastic years in Graz, Austria and six interesting months at imec in Leuven, Belgium.

First of all, I want to thank my supervisor at imec, *Elena Smets*, for guiding me through, providing materials, carefully reading and improving my text and delivering inspiration to attack the encountered problems.

I want to thank my other supervisor at imec as well, *Walter De Raedt*, for the support and making this project possible.

I would like to thank my supervisor at TU Graz, *Denis Helic*, for his voluntary participation, support and advice for my thesis.

Furthermore, I thank the jury for carefully reading the thesis and being present at the defence.

Definitely, a big thank you to my parents, for allowing me to take another master abroad and for their endless support.

Thank you Fiona, for helping me with translations to German.

Abstract

This thesis demonstrates the potential and benefits of unsupervised learning with *Self-Organizing Maps* for stress detection in laboratory and free-living environment.

The general increase in pace of life, both in the personal and work environment leads to the intensification and amount of work, constant time pressure and pressure to excel. It can cause psychosocial problems and negative health outcomes. Providing personal information about one's stress level can counteract the adverse health effects of stress. Currently the most common way to detect stress is by the means of questionnaires. This is time consuming, subjective and only at discrete moments in time. Literature has shown that in a laboratory environment physiological signals can be used to detect stress in a continuous and objective way. Advances in wearable technology now make it feasible to continuously monitor physiological signals in daily life, allowing stress detection in a free-living environment.

Ambulant stress detection is associated with several challenges. The data acquisition with wearables is less accurate compared to sensors used in a controlled environment and physical activity influences the physiological signals. Furthermore, the validation of stress detection with questionnaires provides an unreliable labelling of the data as it is subjective and delayed. This thesis explores an unsupervised learning technique, the *Self-Organizing Map* (SOM), to avoid the use of subjective labels.

The provided data set originated from stress-inducing experiments in a controlled environment and ambulant data measured during daily-life activities. Blood volume pulse (BVP), skin temperature (ST), galvanic skin response (GSR), electromyogram (EMG), respiration, electrocardiogram (ECG) and acceleration were measured using both wearable and static devices.

First, a supervised learning with *Random Decision Forests* (RDF) was applied to the laboratory data to provide a gold standard for unsupervised learning outcomes.

A classification accuracy of 83.04% was reached using ECG and GSR features and 76.89% using ECG features only. Then the feasibility of the SOMs was tested on the laboratory data and compared a posteriori with the objective labels. Using a subset of ECG features, the classification accuracy was 76.42%. This is similar to supervised learning with ECG features, indicating the principal functioning of the SOMs for stress detection. In the last phase of this thesis the SOM was applied on the ambulant data.

Training the SOM with ECG features from the ambulant data, enabled clustering from the feature space. The clusters were well separated with large cohesion (average silhouette coefficient of 0.49). Moreover, the clusters were similar over different test persons and days. According to literature the center values of the features in each cluster can indicate stress and relax phases. By mapping test samples on the trained and clustered SOM, stress predictions were made. Comparison against the subjective stress levels was however poor with a root mean squared error (RMSE) of 0.50.

It is suggested to further explore the use of *Self-Organizing Maps* as it solely relies on the physiological data, excluding subjective labelling. Improvements can be made by applying multimodal feature sets, including for example GSR.

Contents

Abstract **v**

1 Introduction **1**

2 Literature Study **3**

2.1 Background 3

2.1.1 Definition of stress 3

2.1.2 Physiological reaction to stress 4

2.2 Methods for stress measurement 6

2.3 Physiological signals and sensors 8

2.4 Previous experiments on stress detection 11

2.4.1 Physiological features for stress detection 11

2.4.2 Physical activity confounds 13

2.5 Algorithms applied for stress detection 15

2.6 Validation of stress detection 16

2.7 Conclusion 19

2.8 Thesis objectives 19

3 Context and data set **21**

3.1 imec 21

3.2 Sensors and experiments 22

3.2.1 Sensors 22

3.2.2 Test persons 24

3.2.3 Experimental protocol 24

3.3 Goal and general approach for conducted experiments 26

4 Supervised learning **29**

4.1 The Random Decision Forest model 29

4.1.1 Framework 30

4.1.2 Implementation 32
4.2 Physiological signals 33
4.3 Feature set of Galvanic Skin Response 33
4.4 Labelling of feature set 34
4.4.1 Objective labelling 34
4.4.2 Subjective labelling 35
4.5 Classification 37
4.5.1 Objective classification 39
4.5.2 Report-based classification 40
4.6 Controlled environment with wearables 41
4.6.1 Empatica 42
4.6.2 Health patch 44
4.6.3 Combined model based on GSR Nexus and ECG Health patch 48
4.7 Conclusion 50

5 Unsupervised learning 51
5.1 The Self-Organizing Map 52
5.2 Clustering 55
5.3 Performance 59
5.3.1 Clustering Performance 59
5.3.2 Testing Performance 59
5.3.3 Optimal grid configuration 61
5.4 Stress detection in ambulant data 61
5.4.1 Preprocessing of ambulant data 62
5.4.2 Correlation of signals 65
5.4.3 Influence of day and night physiology 65
5.4.4 Cohesion and separation of clusters 69
5.4.5 Similarity of clusters 74
5.4.6 Defining stress and relax clusters 78
5.4.7 Detection of *stress* and *relax* intervals 79
5.4.8 Validation 79
5.5 Conclusion 80

6 Discussion **83**
6.1 Features and classification performance in supervised versus unsupervised learning . . . 83
6.1.1 Laboratory data . . . 83
6.1.2 Ambulant data . . . 84
6.2 Challenges of wearables for ambulant data collection . . . 85
6.2.1 Challenge of multi-modal signals . . . 85
6.2.2 Challenge of qualitative data . . . 86
6.3 Outcome clustering . . . 87
6.4 Validation on ambulant data . . . 88
6.5 Future work . . . 89

Bibliography **91**

List of Figures

2.1 Stress models 5
2.2 Autonomic Nervous System 7
2.3 Electrocardiogram sample 9
2.4 Skin conductance graph 10

3.1 Sensors 23
3.2 Experimental protocol 25

4.1 Decision Tree 30
4.2 Forest probability distribution. 32
4.3 Objective labelling 35
4.4 Subjective stress level 37
4.5 Subjective labelling 38
4.6 Objective classification 39
4.7 Feature importances after classification with GSR and temperature 40
4.8 Report-based classification 41
4.9 Synchronisation Nexus and Empatica. 43
4.10 Synchronisation Nexus and Health patch. 44
4.11 Detailed view ECG Health patch 45
4.12 Features of ECG Health patch 47
4.13 ECG feature importance 47
4.14 Classification ECG Health patch 48
4.15 Classification with GSR and ECG TP05 49

5.1 Training of SOM 54
5.2 Clustering of SOM 56
5.3 Cluster boxplots 57
5.4 Reference cluster boxplot 58
5.5 Algorithmic performance 60

5.6 Preprocessing Health Patch signal 63
5.7 Preprocessing Health Patch signal - low data percentage retained 64
5.8 Graphs of ECG features 66
5.9 Graphs of normalised ECG features 67
5.10 Correlation heat map 68
5.11 Day and night physiology correspondence with SOM 69
5.12 Silhouette coefficient 70
5.13 Silhouette coefficient *LFHF* 72
5.14 Silhouette coefficient feature subset *mean HR* and *RMSSD* . . . 73
5.15 Trained SOM with clustering 74
5.16 Similiarity of boxplots 76
5.17 Undesirable clustering 77
5.18 Cluster of extreme values 78
5.19 Prediction with questionnaire stress level 80

List of Abbreviations

ANS	Autonomic Nervous System
BVP	Blood Volume Pulse
ECG	Electrocardiography
EMA	Ecological Momentary Assessment
EMG	Electromyography
ESM	Experience Sampling Method
GSR	Galvanic Skin Response
HF	High Frequency
HPA	Hypothalamic–Pituitary– Adrenocortical
HRV	Heart Rate Variability
IBI	Inter Beat Interval
KNN	K-Nearest Neighbor
LEAS	Levels of Emotional Awareness Scale
LF	Low Frequency
LF/HF	Low-to-High Frequency Ratio
LOO	Leave-One-Out
NN	Neural Network
PNS	Parasympathetic Nervous System
RDF	Random Decision Forest
RMS	Root Mean Square
RMSE	Root Mean Squared Error
RMSSD	Root Mean Square Successive Difference of NN intervals
SC	Skin Conductance
SCL	Skin Conductance Level
SCPh	Phasic Skin Conductance
SCRR	Skin Conductance Response Rate
SDNN	Standard Deviation of NN intervals
SNS	Sympathetic Nervous System
SOM	Self-Organizing Map
SVM	Support Vector Machine
TP	Test Person

1 Introduction

According to the report from the European Commission in 2010, *stress, depression and anxiety* were the most serious work-related health problems among 14% of workers. There is a general increase in pace of life, both in the personal and work environment. This leads to the intensification and amount of work, constant time pressure and pressure to excel. If poorly managed, this can lead to psychosocial problems and negative health outcomes [1].

The adverse health effects of stress are present both in the short term as in the long term. Short periods of exposure to stress can cause sleep disturbance, changes in mood, fatigue, headaches and stomach irritability [1]. Chronic stress may lead to a weakened immune system and thereby contributing to the course of inflammatory diseases such as multiple sclerosis, rheumatoid arthritis and coronary heart disease [2, 3]. Other mental and physical health outcomes are depression, suicide attempts, sleep problems, back pain, chronic fatigue, digestive problems and high blood pressure. Also mental strain and reduction of quality of life may be experienced. Workplace stress can even influence the quality of relationships within the family [1].

Work-related stress eventually translates itself to a corporate financial burden. Employees have to take time off work or leave employment. The company undergoes a loss of productivity and an increased level of absenteeism and health care costs. In an EU-funded project in 2013, these costs to Europe were estimated to add up to €617 billion annually [1].

Currently the most common way to detect stress is by the means of questionnaires. This is time consuming, subjective and only at discrete moments in time. Literature has shown that in a laboratory environment physiological signals can be used to detect stress in a continuous and objective way. Advances in wearable technology made it feasible to continuously monitor our physiological signals in daily life. Together with increased knowledge of computational modelling, novel systems are being designed to detect stress in real-time. These provide feedback to its user and informs about his or her stress level [4]. Situations of stress could be unveiled, allowing proper counteracting.

Ambulant stress detection is however associated with several challenges. The data acquisition with wearables is less accurate compared to sensors used in a controlled environment and physical activity influences the physiological signals. Furthermore, the validation of stress detection with questionnaires provides an unreliable labelling of the data as it is subjective and delayed. This thesis explores an unsupervised learning technique, the Self-Organizing Map (SOM), to avoid the use of subjective labels.

2 Literature Study

2.1 Background

2.1.1 Definition of stress

Stress is both a physiological and mental state. Myrtek et al. [5, 6] studied the correspondence of both aspects and state that the perceived level of stress not necessarily coincides with the physiological measured level of stress. Most of the emotional arousal, indicated by physiological changes, do not reach the level of consciousness. They also found that emotion perception is largely determined by personality dimensions of individuals, such as being *emotional* or*cool*. However research about emotions is difficult as inference about emotions are often made from subjective reports, implying that subjects can only report on emotions they are aware of. In this perspective, Lane et al. [7] developed the Levels of Emotional Awareness Scale (LEAS). Its purpose is to assess emotional awareness. Higher scores on the LEAS correspond to a greater ability of recognising personal emotions. Verkuil et al. [8] suggest that persons with lower emotional awareness might have to rely on different indicators of stress compared to persons with higher emotional awareness.

Different models capturing work related stress have been developed over the last decades. One of these theoretical frameworks is the Demand- Control Model developed by Karasek and Theorell [9]. Two external variables define the model and affect the health of well-being of employees. These are the *psychological job demands* and the *job control* (Fig. 2.1a). The *psychological job demands* are the actual psychological stressors in the working environment such as time pressure, working pace and complexity of the task. The *job control* is the ability of the employee to control its tasks and his behaviour: how are the tasks executed, timing, sequence, et cetera. This control acts on the psychological stressors to keep them within an acceptable range. Jobs requiring a high demand, though with options for control are experienced to be stressful because they limit the employee's autonomy while delivering a constant pressure.

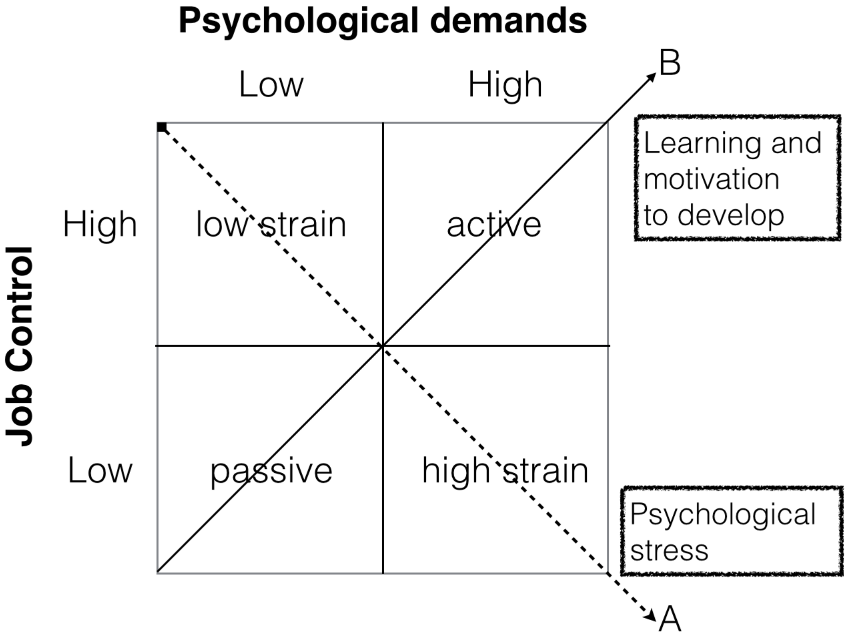
The Effort-Reward Imbalance model developed by Siegrist [3] focusses on the rewarding system of the job as an actor against stress. The *job-related efforts* are seen as part of a social exchange system and are balanced by *occupational rewards*(Fig. 2.1b). Efforts such as time pressure and physical labour could be rewarded by money, appreciation or career prospects. The model claims that stressful experiences at work are evoked by conditions of *high cost-low gain*, typically due to low qualification or overcommitment in people striving for approval.

Both models target the harmful factors at work and complement each other. Low control and low reward are considered to be equally stressful parameters in jobs requiring high effort. They elicit stress with long-term health consequences [3].

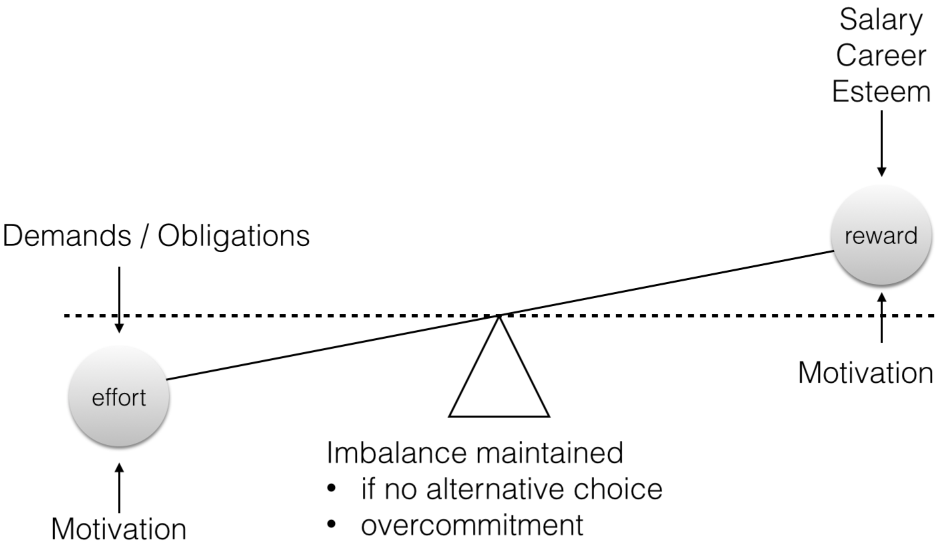
2.1.2 Physiological reaction to stress

The Autonomic Nervous System (ANS) plays a central role in stress reactions (Fig. 2.2). The ANS is divided in two branches: the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). Most tissues are innervated by both branches in which they have an opposing effect. The SNS activates the *fight-or-flight* mechanism of our body and prepares for action. The PNS predominates in relaxing conditions, regulates basic body functions and brings the body back to a rest state. The sympathethic system aims to increase the delivery of well-oxygenated blood to skeletal muscles for activation. Hormones epinephrine, norepinephrine and cortisol are released into the blood stream. This causes heart rate to rise to increase the flow of blood. The respiration rate rises to increase the uptake of oxygen from the atmosphere and release carbondioxide. Sweating enables thermoregulation during these conditions. Also the diameter of the pupil dilates and the lens adapts for distant vision. The PNS in its turn lowers heart rate and adapts the eye to its resting state. The body gets focussed back onto digestion [12].

A stressor activates the sympathethic system. After termination of the stressor, the production of cortisol ceases and the balance between both SNS and PNS is restored. However a prolonged presence of the stressor, such as chronic stress



(a) Demand-Control Model developed by Karasek and Theorell. Figure adapted from [10]



(b) Effort-Reward Imbalance Model developed by Siegrist. Figure adapted from [11]

Figure 2.1: Stress models

experienced in working situations, can cause an overload and eventually exhaustion of SNS and its balance with PNS [13]. It has been found that panic disorder patients experience higher baseline levels of skin conductance (related to sweating) and norepinephrine secretion [14]. During stress exposure these levels will increase further as compared to healthy subjects.

2.2 Methods for stress measurement

Stress occurs at different levels in the body, thus different sensing modalities exist to measure different aspects of stress.

A method of measuring the perceived level of stress is self-reporting. In [16] test persons filled in a Perceived Stress Scale questionnaire right before the actual stress experiment. The lab study of [17] used an anxiety questionnaire to define three degrees of stress: no stress, low and high. Some self-reports also aim to study how people feel and what they do in their daily lives. This includes daily diary studies, interview methods [18] and Experience Sampling Method (ESM) or Ecological Momentary Assessment (EMA) [19, 4]. A disadvantage of a diary study is the reliance on memory of the test persons. During an ESM study, test persons are prompted at random times to immediately give feedback. However, these prompts may be experienced as highly interruptive and become a source of stress itself [20]. Other disadvantages are its subjectiveness and the discontinuity of measurements.

An objective measure for acute stress are biochemical parameters. These are epinefrine, norepinerfrine and cortisol activitation in blood. Cortisol is a hormone regulated by the hypothalamic–pituitary– adrenocortical (HPA) axis. Physiological stressors such as arithmetic tasks and public speaking can increase cortisol levels. However, the effect of these stressors on cortisol activity is inconsistent over literature [17]. There are different characteristics of cortisol activation to take into consideration during assesment. First, cortisol has a circadian rhytm during which cortisol levels greatly vary. Second, the increase in cortisol level after the onset of a stressor experiences a delay as the HPA axis needs to be activated first. Third, the type of cortisol measured depends on the method of assesment. Plasma samples both include cortisol bounded to protein as unbounded cortisol. Salivary samples only reflect unbounded cortisol. Also, the reaction to the venipuncture

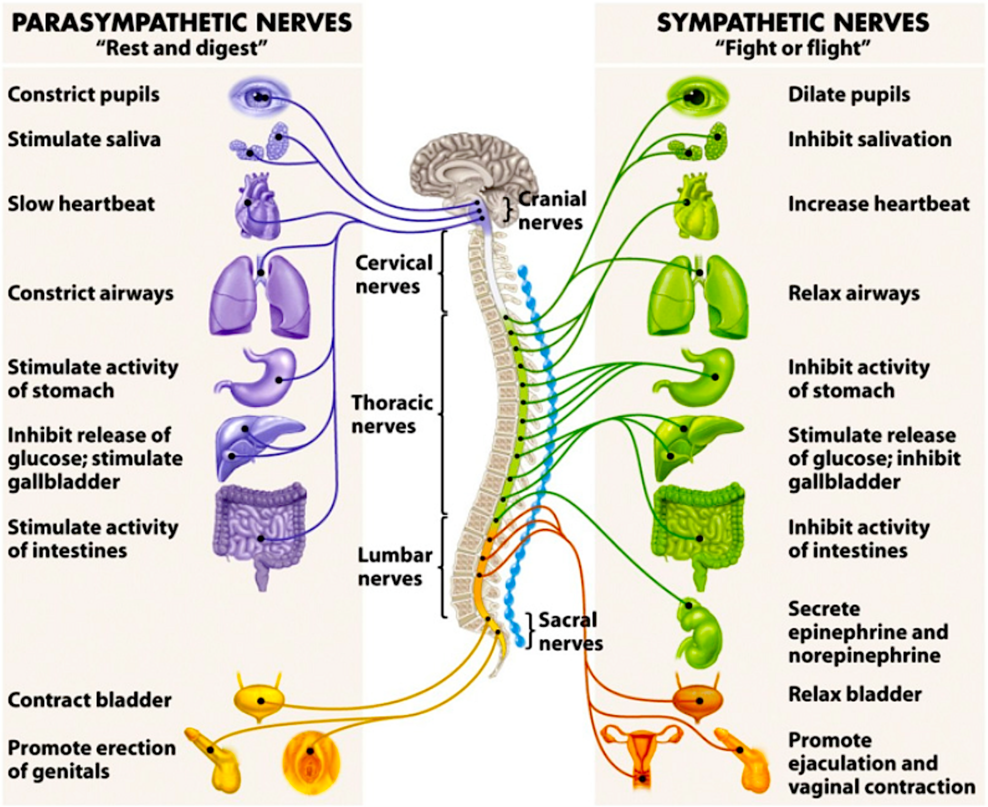

PARASYMPATHETIC NERVES
"Rest and digest"
Constrict pupils
Stimulate saliva
Slow heartbeat
Constrict airways
Stimulate activity of stomach
Inhibit release of glucose; stimulate gallbladder
Stimulate activity of intestines
Contract bladder
Promote erection of genitals
Cranial nerves
Cervical nerves
Thoracic nerves
Lumbar nerves
Sacral nerves
SYMPATHETIC NERVES
"Fight or flight"
Dilate pupils
Inhibit salivation
Increase heartbeat
Relax airways
Inhibit activity of stomach
Stimulate release of glucose; inhibit gallbladder
Inhibit activity of intestines
Secrete epinephrine and norepinephrine
Relax bladder
Promote ejaculation and vaginal contraction

Figure 2.2: The Autonomic Nervous System is divided in two branches: the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). The SNS activates the *fight-or-flight* mechanism of our body and prepares for action. The PNS predominates in relaxing conditions, regulates basic body functions and brings the body back to a rest state. Figure from [15]

for plasma samples could cause a reaction in cortisol levels [21]. Moreover, it as intrusive method which is not suitable for frequent measuring.

Methods that collect stress-related information continuously and without interference are based on measurement of physiological signals. In [22] a set of different sensors was used to detect stress during real-world driving tasks. In correspondance with the driver's task performance, the information could serve a system to reduce the driver's stress level. The study relied on the following physiological sensors: electrocardiogram (ECG), electromyogram (EMG), skin conductivity (or galvanic skin response (GSR)) and respiration. These physiological signals were also used in [16, 23, 24]. Another frequently measured signal is blood volume pulse (BVP) [25]. Not so widely used is the measurement of the pupil diameter and skin temperature, applied by J. Zhai and A. Barreto in [25, 26]. As activity can alter physiological signals [27], the study in [28] combined the measurement of ECG and GSR with accelerometer information.

2.3 Physiological signals and sensors

Many studies have shown that a combination of the physiological parameters that are influenced by the SNS during stress is suitable for stress detection. First, two important physiological signals are described in more detail, i.e. Heart Rate Variability and the Galvanic Skin Response. Then it is explained how different physiological signals can be measured.

Heart Rate Variability (HRV)

An electrocardiogram (ECG) records the electrical activity of the heart. A sample from a typical ECG is depicted in Fig. 2.3. The P wave is associated with the contraction of the atria. The QRS complex is associated with the contraction of the ventricles. The T/U waves are associated with the repolarization of the ventricles [29]. The R-R distance (or R-R interval) is the time between two R peaks and are used in the calculation of the heart rate. ECG signals from different individuals can exhibit personalized traits such as the relative timing of the peaks

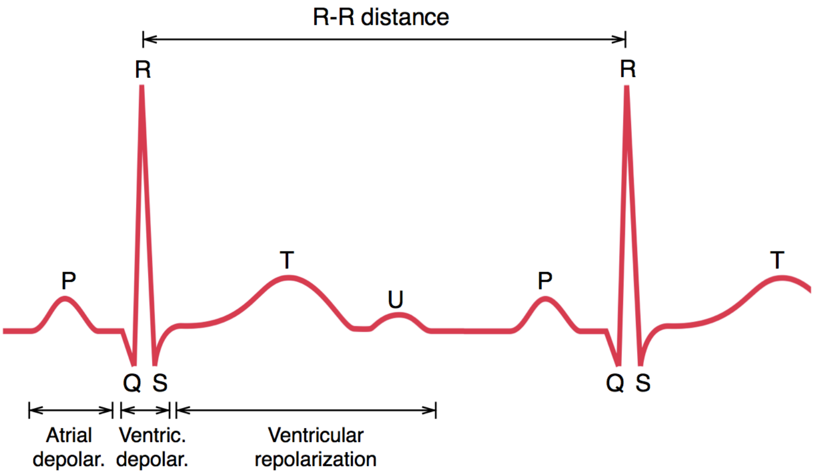


Figure 2.3: Sample from a typical normal ECG. The P wave is associated with the contraction of the atria. The QRS complex is associated with the contraction of the ventricles. The T/U waves are associated with the repolarization of the ventricles. The R-R distance is the time between two R peaks. [29]

but can also exhibit responses to stress and activity. Heart rate variability is the beat-to-beat variation in the R-R interval [28].

Galvanic Skin Response (GSR)

GSR is a measure for the electrical resistance of the skin. The physiological response to a sudden stimulus causes the resistance of the skin to vary. The skin conductance is proportional to sweat secretion. The density of sweat glands is highest at the hands and feet [30]. Under stress conditions the sweat glands are activated, causing an individual to sweat and the skin conductance to rise [28]. The sweat glands and skin blood vessels are exclusively innervated by the SNS. Thus skin conductance is the ideal measure for sympathethic activation and stress reaction. The slowly changing part of the SC signal is the skin conductance level (SCL) and is a measure of psychophysiological activation. The fast changing part of the SC signal is the skin conductance response (SCR) and is the reaction to a sudden stimulus, depicted in Fig. 2.4 with typical features [31].

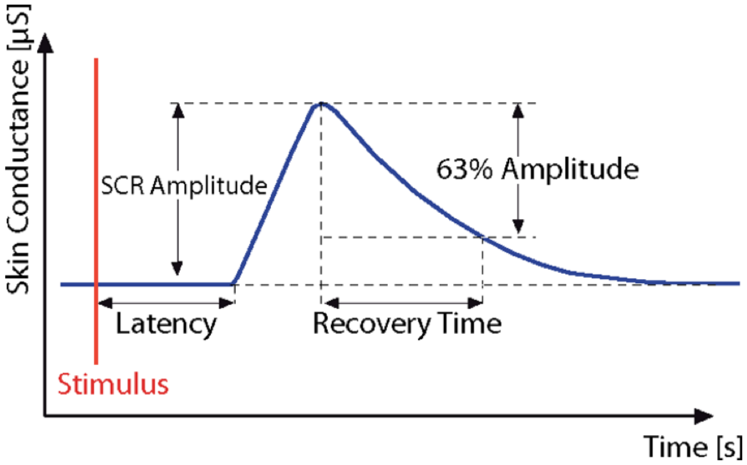


Figure 2.4: Ideal skin conductance response with typical features. [31]

Sensors

Some of the following studies focus on fixed sensor set-ups for stress detection, newer studies will aim for wearable sensor systems.

EMG is most commonly measured at the upper trapezius muscles of the shoulders [22, 16]. Skin conductance (SC) can be measured at different locations, such as palm of the hand or sole of the foot as they contain the highest density of sweat glands. If measuring the SC at the palm of the left hand, an electrode is placed on the first and the middle finger [22]. In [16] a wireless hand sensor measured the change in current after applying a voltage of 0.5V DC across the palm. For wearable applications, the SC is measured at the wrist. Blood volume puls is typically measured by photoplethysmography [25]. Respiration is related to chest cavity expansion and measured by an elastic sensor strapped around the test persons's diaphragm. In [22] the chest belt was connected to an analog-to-digital converter, while [16] used a wireless chest belt. This wireless chest belt was also used for measurement of one lead electrocardiography (ECG). Furthermore, respiration and blood volume pulse can be extracted from the ECG signal.

2.4 Previous experiments on stress detection

A vast number of studies have been carried out investigating stress and stress detection. They address questions such as: Which physiological signals are altered by stress? What are important features for stress detection? How to detect stress in an office-like or more broad real-life situation? How does activity hinder stress detection? Which sensors are useful? How to set up a wearable sensors system?

Earlier papers focus on the relation between stress and changing physiological parameters. In [32], test persons were exposed to psychologically challenging tasks of increasing difficulty. They found that cardiac activity was sensitive to these variations in difficulty, as easy conditions elicited lower cardiac activity compared to harder tasks. The study also reported that oxygen consumption and carbon dioxide production did not vary under different stress conditions. Research done by Boucsein has extensively investigated the effect of stress on GSR [30]. As more knowledge was gathered which physiological signals vary with stress, researchers started to focus on specific features to discriminate stress.

2.4.1 Physiological features for stress detection

The Heart Rate Variability (HRV) under stress and its derived features has been specifically explored by many studies [33, 34, 13, 35] and is still topic of ongoing research [8]. HRV measures the variance in time between consecutive heart beats and reflects the ANS activity [36].

The study [33] focussed on stress induced by working activities as participants were ambulatory monitored during two working days and one nonworkday. They received the Effort-Reward Imbalance questionnaire (sec. 2.1.1) which is developed to measure perceived chronic work stress. The blood pressure (BP) of participants was measured during daytime and their heart rate (HR) and vagal tone was recorded over 24h. Vagal tone was calculated based on the root mean square of successive differences in interbeat intervals (RMSSD). The researchers found three characteristics of high work stress, being an increased HR reactivity (HR during work minus HR during sleep), an increased systolic BP and a lower vagal tone. Moreover, subjects enduring chronic work stress have an increased systolic BP during 24 hours, not simply caused by an increased BP during work time. Also vagal tone stayed at a low level both during the night as during a non working day. Interestingly, resting level of HRV and levels of emotional awareness are proportionally related to with the ability to report stress.

The research of [34] was executed in a lab environment where participants had to perform standardized computer work, including stress sessions. It was reported that there was a reduction in the high-frequency (HF, 0.15 to 0.4 Hz) component of HRV, a stable low-frequency (LF, 0.04 to 0.15 Hz) component and an increase in low-to-high frequency ratio (LF/HF) during a stress situation. It also confirmed an increase in BP during stress sessions. These results were confirmed by another lab study [13] where the following frequency domain features were implemented: peak frequency and power of very low frequency bands (VLF, 0 to 0.04 Hz), low frequency bands and high frequency bands. The LF/HF ratio was found to increase during mental tasks however insignificant.The implemented time domain features were mean and standard deviation of RR intervals and HR, root-mean-square (RMS), the number of consecutive RR intervals that differ more than 50ms (NN50) and the proportion of NN50 (pNN50). The pNN50 was found to be significantly higher in the rest condition than during the mental task.

Melilo et al. [35], monitoring real-life stress, investigated nonlinear features of HRV for automatic stress detection. They proposed the use of Poincaré Plot measures and Approximate Entropy. A recent publishing [8] focussed on capturing prolonged additional reductions in HRV in daily life and verifying if these periods were related to stress. Prolonged periods of low HRV are considered to be harmful for health as they may contribute to lower resting levels of HRV. On the other hand, physical activity also reduces HRV, however only temporarily. Therefore, the study compensates for physical activity, to be able to detect periods of reduced HRV, not caused by physical activity. This is an important contribution as many other ambulatory studies simply exclude high activity data, according to this paper. Periods of additional HRV decrease where determined by comparing the expected and actual RMSSD for each individual and period.

An early study of Boucsein proved that skin conductance level (SCL) and skin conductance response (SCR) are valid indicators for stress [30]. Setz et al. used a wearable GSR device to monitor skin conductance (SC) while eliciting office-like stress by a mathematical task and psychological stress.They found that the distributions of the peak height of the SCL and the SCR peak rate carry information about a person's stress level [31]. Wijsman et al. [16] successfully applied the SC responses rate (SCRR) and the signal power of the SC signal (SCDIFF2).

Many other papers studying stress detection preferred a multimodal approach. Multiple physiological signals are examined simultaneously. The research by Zhai and Barreto [25] is a lab study with stress inducing tests. Participants are attached to several sensors to enable feature calculation from BVP, GSR and pupil diameter. From BVP, its amplitude was calculated together with mean and standard deviation of the BVP period or interbeat interval (IBI) and the LF/HF ratio. The IBI is a useful measure for HRV and its analysis in frequency domain is reliable with smaller sets of data. GSR analysis relied on its mean value and the number, amplitude,rising time and energy of skin conductance responses. The pupillographic activity was solely based on the mean value of pupil diameter, which was expected to increase during stress periods. The study by Healey and Picard [22] aims to detect stress during real-world driving tasks. The participant is situated outside of a laboratory environment, though attached to several sensors inside the car. The proposed statistical features are mean and variance for EMG, respiration, heart rate and skin conductivity. Besides those general features, four spectral power features were defined for respiration by summing the energy in four different energy bands. Skin conductivity was further examined by the number of SC responses in a window, the magnitude and duration of SC responses and the sum of the area under the SC response. One specific HRV feature was used, the low-to-high frequency ratio LF/HF.

An overview of widely used features are presented in table 2.1 for ECG and table 2.2 for GSR.

2.4.2 Physical activity confounds

Previous research focused on detection of mental stress from subjects at rest. Advances in mobile computing and wearable sensors allow development of a real-life ambulant stress monitoring system. First studies with wearables were executed with participants at rest [31, 16, 24, 23]. However daily-life stress detection comes

Table 2.1: Overview of widely used features for ECG

Feature	Explanation
HR	mean heart rate
SDNN	standard deviation of all normal RR intervals (i.e. NN intervals)
RMSSD	root-mean-square successive difference of all normal RR interval
pNN50	proportion of number of consecutive RR intervals that differ more than 50ms
VLF	very low frequency HRV (power in the 0-0.04 Hz band)
LF	low frequency HRV (power in the 0.04-0.15 Hz band)
HF	high frequency HRV (power in the 0.15-0.4 Hz band)
LFHF	ratio of low and high frequency of HRV

Table 2.2: Overview of widely used features for skin conductance

Feature	Explanation
SCL	mean SC level
SCPH	signal power in a phasic SC signal
SCRR	mean number of SC responses per window
SCDIFF2	signal power in second difference from SC signal
SCR peak rate	mean number of SCR peaks per window
SCL peak height	average height of SC peaks

along with the major obstacle that physiological responses caused by stress can be masked by responses originating from physical activity.

This difficulty is addressed in [28] in which an activity-aware stress detection system is presented. An accelerometer was placed at the waist to record sitting, standing and walking activities. Apart from established HRV and GSR features, twelve accelerometer features were calculated. Accelerometer data was proven to be necessary to improve mental stress detection in a mobile environment. Physiological signals tend to be user-dependent, therefore the training stage should rely on personalized data as well. Interestingly, GSR features were relative independent of the recorded activities. The study was however only limited to three specific activities.

Some very recent studies deal with physical activity by excluding these data intervals. One of these, is the study by Hovsepian et al., addressing many issues regarding ambulant stress detection [37]. They exclude one-minute frames of moderate-to-high physical activity by applying a simple threshold on information from a 3-axis on-body accelerometer. The authors of study [38] remove periods with moderate to high physical activity plus the estimated time for recovery of the activity.

In [17], the type of activity is recognised first (sitting, walking, running and cycling). With the use of a base stress detector and the recognised activity, a context based detector with contextual features is applied. The context based detector is designed to distinguish between periods of actual stress and other situations causing physiological arousal.

2.5 Algorithms applied for stress detection

Most stress detection mechanisms in literature rely on a machine-learning pipeline where the extracted features are fed to a self-learning algorithm. These feature values are either labelled and it is known which values corresponds to states of stress or the data is unlabelled. The former is applied to train supervised algorithms, the latter to train unsupervised algorithms. The trained algorithm is then able to predict stress states for unseen features values.

Most studies in the field of stress detection rely on supervised algorithms. Many studies make use of a Support Vector Machine (SVM) for learning and classification [26, 25, 39, 37]. Also Linear Discriminant Analysis (LDA) is a widespread algorithm [22, 31]. A Random Decision Forest (RDF) has been applied by Sarker et al. [38] and a combination of an SVM and a RDF in [17]. Sharma and Gedeon [40] apply a Genetic Algorithm and an SVM for feature selection, followed by a Neural Network (NN) for classification. Feng-Tso et al. examined the classification performances between Decision Tree, a Bayesian Network and an SVM [28]. Wijsman et al. made a performance comparison between Bayesian classifiers, LDA and K-Nearest Neighbor (KNN) [16], of which the latter is an unsupervised technique.

In previous work on stress detection, executed in the lab, activities were often constrained to e.g. sitting, walking and biking [17, 28]. This restriction enabled the labelling of activity data and context-specific monitoring. However real-life activities cover a much broader range of activities. Ramos et al. [41] included physical activity in their analysis but moved away from labels for different activities. Instead they exploited the full distribution of the physiological responses. Different unsupervised algorithms were tested for optimal performance: Hidden Markov Models, Naive Bayes, SVM and Logistic Regression.

Another reason for using unsupervised techniques is subjectiveness and unreliablity or complete absence of the labels. The study by Bornoiu and Grigore [42] investigates stress detection by using electrodermal features. Their evaluation method relies on subjective observations from an expert observer, marking a recorded signal as stress or relax. This labelling lacked consistency. Therefore an unsupervised method was required and they proposed to use a Kohonen Neural Network, also known as Self-Organizing Map (SOM). This technique has been successfully applied in many others fields besides stress detection such as Brain Computer Interfaces (BCI) [43]. Medina [44] identified stress states from ECG signals using several other unsupervised learning methods. These are clustering algorithms (including K-means and Spectral Clustering) and clustering ensemble methods as well as dimensionality reduction techniques (Principal Component Analysis and Forward Sequential Search) and evolutionary algorithms.

2.6 Validation of stress detection

Validation is a necessary step when developing algorithmic pipelines. It requires a different approach when dealing with unsupervised algorithms or ambulant stress monitoring systems, compared to supervised algorithms and laboratory studies.

Supervised learning

Supervised classification methods can make use of several validation measures (overview of examples in table 2.1). Frequently used is the classification accuracy, which is the number of correct predictions divided by the total number of predictions made [28]. More detailed measures are sensitivity and specificity, based on *TP* the number of true positives, *FN* the number of false negatives, *TN* the number of true negatives and *FP* the number of false positives. Sensitivity can be seen as the *stress* detection rate and specificity as the *relax* detection rate. The average of both can be taken as the classification performance instead of the aforementioned classification accuracy. It delivers a more balanced outcome if one of the labels appears less frequent [45]. Closely related are the precision and recall. The precision is intuitively the ability of the classifier not to label negative samples as positive. The recall is seen as the ability of the classifier to find all the positive samples. The F1 score is the weighted average of precision and recall [46]. These numbers of *TP, FN, TN* and *FP* can be summarised in a confusion matrix [28].

Examples of validation measures for supervised learning:

$$\begin{aligned} Sensitivity &= \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{TN+FP} \\ Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ F1 &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \end{aligned}$$

Unsupervised learning

The aforementioned study by Medina [44] was performed in laboratory conditions, however using unsupervised techniques and therefore lacking labelling. Without labelling, standard validation techniques cannot be applied as there is no ground truth to compare the results with. Here, clustering results are combined and represented by a co-association matrix as measure of similarity between patterns. This co-association matrix is visualised in which quadrangular shapes will emerge if contiguous patterns belong to the same cluster. Validation is performed by inspection of these patterns, such as the degree of separation of clusters. Different algorithms will lead to different patterns. However extracted patterns are expected to be similar if a real structure is present in the analysed data.

Ambulant stress monitoring systems might have been validated in the lab first, but require a validation step in the field too. Validating ambulant stress detections is not a clearly defined process as there is no precise recording of what the participants' activities are. The only apparent way for validation are different types of questionnaires and diaries, filled in multiple times a day.

Kusserow et al. [47] monitored the participants by a diary of daily activities (e.g. working, transport, conversation) and mood-state questionnaires which had to be filled in as soon as possible after perceiving stress arousal. They obtained individual and daily-activity-specific stress-arousal characteristics. The authors calculated the probability $P(A|R)$ of being in an arousal phase A during a daily activity R and matched these probabilities to the questionnaires. However most questionnaires were completed randomly and could not be related to the estimated stress-arousal phases.

Adams et al. [20] used an ESM study 2.2, enabled by their specifically designed smartphone app *SESAME*. Participants received approximately every half hour a notification to fill in the self-report and were free to fill in additional self-reports. Researchers increased the prompt frequency to a maximal accepted rate to collect ground truth data. The sensed data was smoothed over one hour windows, giving a larger weight to data points close to the self-report time stamp. This data was compared to normalized self-report values. In practice, many experience-sampling responses were delayed due to practical reasons of the application or occupation of the participants, or participants did not respond at all to notifications. Moreover, periods of time associated with very high levels of stress are under-reported.

Hovsepian et al. [37] present a stress model *cstress* that wants to provide a gold standard for continuous stress measurements from wearable sensors. Regarding self-reports in a field study, they prompted participants at random 15 times a day to fill in a an Ecological Momentary Assessment 2.2. This EMA self-report serves as the ground truth for field validation. The *cstress* model compensates for the arbitrary lag between the occurrence of a stressor and its self-report logging. The perception of stress for minute i depends on the perception of stress in minute $i-1$ and, if present, the physiological stress arousal in minute $i-1$. On the field data, an accuracy of 72% was reached to predict the self-report. The *cstress* model was applied by Sarker et al. [38] in combination with questions about drugs and smoking cues. Additionally, they assessed the consistency of the self-reported responses by Cronbach's alpha measures.

2.7 Conclusion

The concept of stress is difficult to capture. It has both a psychological and a physiological aspect, of which both are complex and caused by multiple factors. The psychological part has been described by multiple models, of which two important models have been discussed, i.e. the Demand-Control Model and the Effort-Reward Imbalance Model. Physiologically, stress can be described by the activity and balance of the autonomic nervous system. As stress is multifaceted, different sensing modalities exist.

The focus of this thesis is on the physiological signals, of which heart rate and galvanic skin response are considered most relevant. Many studies have been executed on the physiologic reaction of stress and how to apply this knowledge for stress detection. Several studies especially focus on HRV and its features under stress. Fewer studies exclusively analyse GSR. However, GSR is frequently applied in multimodal sensoring systems, which not only includes HRV but also EMG and respiration.

The interest for stress detection shifted from laboratory conditions to ambulatory, enabled by the growth of wearable sensor technology. Wearables were first tested in lab studies, to slowly implement them into daily-life studies. This unveiled two major problems. First, physiological responses caused by stress can be masked by responses originating from physical activity. Some studies take the high-activity data into account, others simply exclude this data. Second, validation of the ambulant data requires a different approach compared to lab studies. Labelling of the physiological data is either nonexisting or subjective and possibly unreliable. The latter problem encourages the use of unsupervised stress detection algorithms.

2.8 Thesis objectives

This thesis explores an unsupervised learning technique, the *Self-Organizing Map* (SOM), to avoid the use of subjective labels.

In order to appropriately explore this alternative unsupervised technique, the presented thesis first focusses on established methods for stress detection. Literature has shown that in laboratory environment physiological signals can be used to detect stress in a continuous and objective way by applying supervised techniques. Therefore, the first objective of this research is to detect stress with good accuracy on the given laboratory dataset by application of the supervised learning method *Random Decision Forests* and the use of objective labels. The associated questions are which physiological signals to include and which features to derive for optimal stress detection. Furthermore, the quality of the wearable sensors during lab phase has to be reviewed first. If these signals lack quality already during lab phase, they cannot be expected to improve quality during ambulant phase. These outcomes are the gold standard for the subsequent unsupervised phases.

The second objective of this thesis is to explore the unsupervised learning technique, *Self-Organizing Map*, for stress detection. During the first phase of the exploration the objective is to test and validate the feasibility of the *Self-Organizing Map* on the laboratory dataset One of the most important aspects in this phase is the selection of the most performant subset of features for training of the SOM. Subsequent challenges are how to interpret the trained SOM and how to apply it correctly for stress detection.

In the second phase of the exploration the objective is to test the SOM on the ambulant dataset for stress detection. Ambulant stress detection is associated with several challenges. The data acquisition with wearables is less accurate compared to sensors used in a controlled environment and physical activity influences the physiological signals. Furthermore, the validation of stress detection by questionnaires is expected to be poor as it is subjective and delayed. Together with exploration of the unsupervised learning technique, the exploration of validation techniques is enforced.

3 Context and data set

This chapter presents the context in which this thesis is conducted and the provided data sets. The thesis is written in cooperation with imec, a research and innovation hub in nano-electronics, energy and digital technologies. The company and context of this thesis is presented in section 3.1. Two datasets were provided for which techniques for stress detection had to be developed. The sensors applied, participant group and experimental protocol are presented in section 3.2. The last section (sec. 3.3) gives an overview of how this thesis research is carried out using laboratory and ambulant datasets.

3.1 imec

Imec is the world-leading research and innovation hub in nano-electronics, energy and digital technologies. They provide a unique combination of widely acclaimed leadership in microchip technology and profound software and ICT expertise. Their world-class infrastructure is leveraged by a local and global ecosystem of partners across a multitude of industries. Groundbreaking innovations are made in application domains such as healthcare, smart cities and mobility, logistics and manufacturing, and energy [48].

Imec partners with different companies, start-ups and universities, bringing together close to 3,500 researchers from over 70 nationalities. Imec is headquartered in Leuven, Belgium and also has distributed R&D groups at a number of Flemish universities, in the Netherlands, Taiwan, USA, China, and offices in India and Japan. In 2015, imec's revenue (P&L) totaled 415 million euro and of iMinds which is integrated in imec as of September 21, 2016 52 million euro [48].

This thesis is a contribution for the Body Area Network (BAN) department of imec headquarters in Leuven. The data sets are part of Elena Smets' PhD research.

Table 3.1: Physiological signals measured by NeXus-10MKII.

NeXus-10MKII	
EMG	measures muscle tension on Trapezius descendens on non-dominant side
Respiration rate	measured by belt around waist
BVP	measured at index finger of non-dominant hand
Skin conductance	measured at middle and ring finger of non-dominant hand
Skin temperature	measured at pink of non-dominant hand

Her research examines if the stress level of employees can be reduced by giving qualitative feedback about this stress level, and as such increase their wellbeing and productivity.

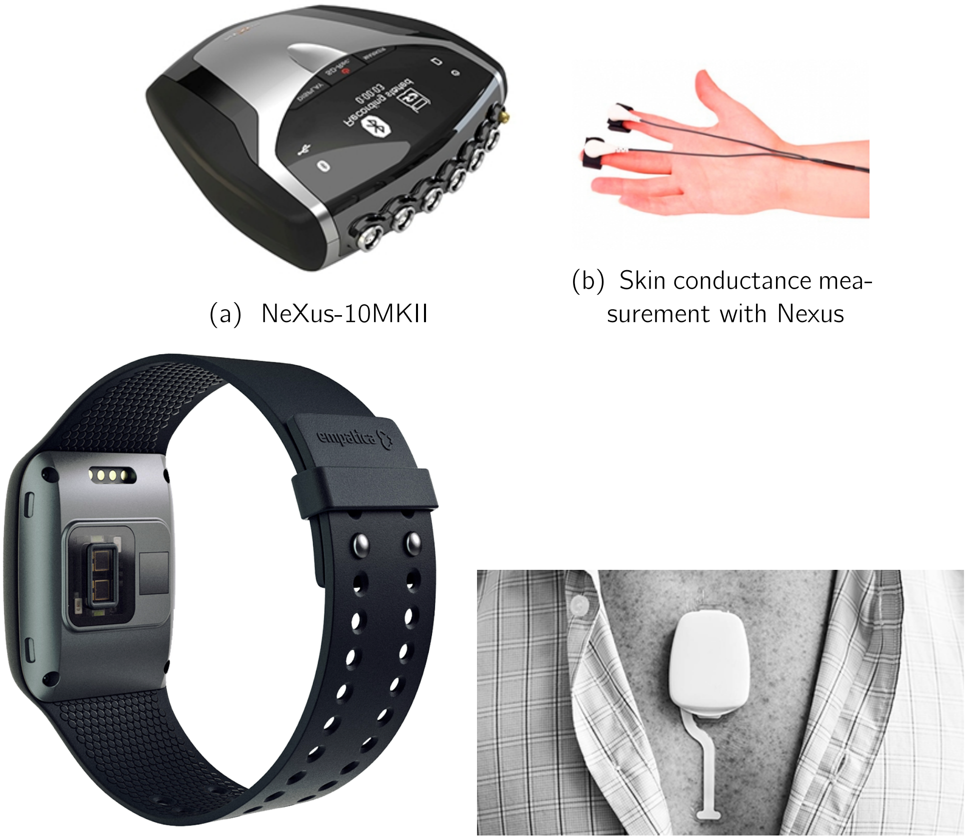
3.2 Sensors and experiments

3.2.1 Sensors

The full experiment consists of two subsequent phases: one lab study and one ambulant study. During these parts, three different sensors were used. These are Empatica E4, Health Patch, worn in lab and ambulant phase, and the NeXus-10MKII, exclusively worn in lab phase (Fig. 3.1).

NeXus-10MKII

The NeXus-10MKII (Mind Media BV, Herten, The Netherlands), later referred to as Nexus,is not wearable, though highly accurate (Fig. 3.1a, Fig 3.1b). Therefore this sensor can serve as a gold standard to compare measurements of other sensors. The following signals were measured: BVP, skin temperature, skin conductance, EMG and respiration (table 3.1).



(a) NeXus-10MKII

(b) Skin conductance measurement with Nexus

(c) Empatica

(d) Health Patch

Figure 3.1: Sensors of laboratory and ambulant experiments

Table 3.2: Physiological signals measured by Empatica E4.

Empatica E4	
Photopletysmography	measured on index finger of non-dominant hand
Skin conductance	measured at the wrist
Skin temperature	measured at the wrist
Acceleration	measured in three dimensions, measured at the wrist

Empatica E4

The Empatica E4 (Empatica, Milan, Italy) is a commercial wearable sensor for BVP, skin temperature, skin conductance and acceleration (table 3.2). It is worn as a wrist band (Fig 3.1c). The Empatica E4 is waterproof and has a memory of +36h.

Health Patch

The health patch is a wearable monitoring system developed by imec (Fig 3.1d). The sensor is a patch consisting of a sensor node and an electronic module to record the ECG signal, continuously for seven days. The battery nor the patch have to be replaced during these seven days.

3.2.2 Test persons

A group of 13 test persons has participated on the experiments (age = 38.5 ± 9.4). There were five male participants and eight female participants. Test subjects were recruited at a therapeutics center. All test subjects reported stress-related complaints, but were not diagnosed with any clinical disorder (e.g. depression or burnout). They were patients at risk, suffering from chronic stress.

3.2.3 Experimental protocol

The test protocol exists of two phases. The first phase is executed under controlled circumstances in the laboratory, followed by a second ambulant phase.

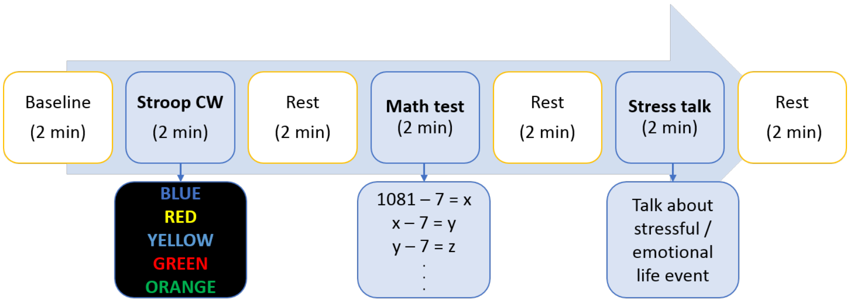


Figure 3.2: Experimental protocol.

Laboratory phase

The goal of the laboratory phase is to define a psychophysiological stress profile of the participants. The participants have to complete three different stress-inducing tests. During these tasks, the participant were monitored with the NeXus-10KMII, Empatica E4 and the Health Patch to measure ECG and the different physiological signals mentioned in table 3.1 and table 3.2. The experiment itself lasts for 14 minutes and is set up as seen in Fig. 3.2. Before the stress test, participants are asked to fill in a questionnaire about their current stress level on a scale of 1 (not at all) to 5 (very much). After the tests, a retrospective questionnaire is presented to determine their stress level between stress tests. A translated excerpt is shown in table 3.3.

Table 3.3: Retrospective questionnaire of stress tests

1	How stressed did you feel before the start of the tests?
2	How stressed did you feel during the tests?
3	How stressed did you feel during the *Stroop* test?
4	How difficult did you find the *Stroop* test?
5	How stressed did you feel during the *Calculation* test?
6	How difficult did you find the *Calculation* test?
7	How stressed did you feel during the *Speech* test ?
8	How difficult did you find the *Speech* test?
9	How relaxed did you feel during breaks?
10	How stressed do you feel at this moment?

During the *Stroop Color Word Test* words of colours are written in a different colour as the colour the word represent, e.g. the word *blue* is printed in red ink. The test person has to say the colour of the ink as correct and as fast as possible. The test person has to suppress the instinctive response of saying the colour the word represents. The correct answer is *red* in this example. Additional stress could be added when the test supervisor urges the test person to be faster or to say *wrong* when a mistake has been made.

The second test is a calculation test in which the participant gets a large number and continuously has to subtract the number 7. In the same manner as the Stroop test, additional stress can be added by the test supervisor.

During the stress talk test, the participant has to talk about a very stressful or emotionally negative event in his life. The participant has to recall his or her feelings related to this event. The test supervisor could ask questions such as *How did you feel?*.

Ambulant phase

The ambulant phase is immediately started after the laboratory phase and lasts four days. During this period the Empatica E4 and the Health Patch are worn by the participant and activated. The Empatica can be taken off over night. The Health Patch stays attached and continues recording during four days. Both sensors are waterproof, though participants cannot go swimming or bathing. During these four days a questionnaire for self-observation was filled in. The questionnaire includes information about activity, food intake, the level of stress and physical complaints. This information is given by the participant hourly.

3.3 Goal and general approach for conducted experiments

The laboratory and ambulant data sets will be used over different phases of research in this thesis. The goal is to explore the performance of an unsupervised learning technique for ambulant data.

The first phase consists of training a supervised model with the laboratory data and assessing its performance with objective labels (chapter 4). The outcome of this phase serves as the gold standard for subsequent phases. The second phase consists of training an unsupervised model with the laboratory data to investigate the feasibility of the unsupervised technique (chapter 5). Therefore, the ambulant data is applied for training of the unsupervised model (chapter 5, from sec. 5.4).

4 Supervised learning

This chapter presents the model based on supervised learning with *Random Decision Forests* (RDF) to detect stress in a controlled environment. The goal is to provide a gold standard for unsupervised learning outcomes of chapter 5.

The framework for classification is the *Random Decision Forest* model (RDF), explained in section 4.1. As the RDF model is a supervised classification method, it requires features and labels. A featureset for Temperature signals is already available. However, the featureset for GSR signals is further expanded (sec. 4.3). The combined feature sets are fed to the model. The labelling is split in an objective labelling (sec. 4.4.1) and a subjective labelling (sec. 4.4.2). The objective labelling is based on the alternation of *relax* or *stress* states in the experiment, while the subjective labelling is based on self-reported levels of stress. The output of the model is a classification into *relax* or *stress* states. Depending on the type of labelling, the output is an objective (sec. 4.5.1) or report-based classification (sec. 4.5.2). During the experiments in a controlled environment, test persons also wore wearable sensors Empatica and Health patch. In section 4.6 it is examined if these wearables deliver reliable and useful signals for *stress-relax* classification.

4.1 The Random Decision Forest model

Several algorithms exist to detect stress. The random decision forest model holds several advantages. First, the method is suitable for large feature sets as no feature selection is necessary. Second, RDF has a large robustness and generalisation power. The reason is that the outcome of the RDF is averaged out over an ensemble of decision trees, in which a certain degree of randomness is applied during training. The basis of the algorithmic pipeline is explained in this section and is based on the studies *Random Forests* by Breiman [49] and *Decision Forests:*

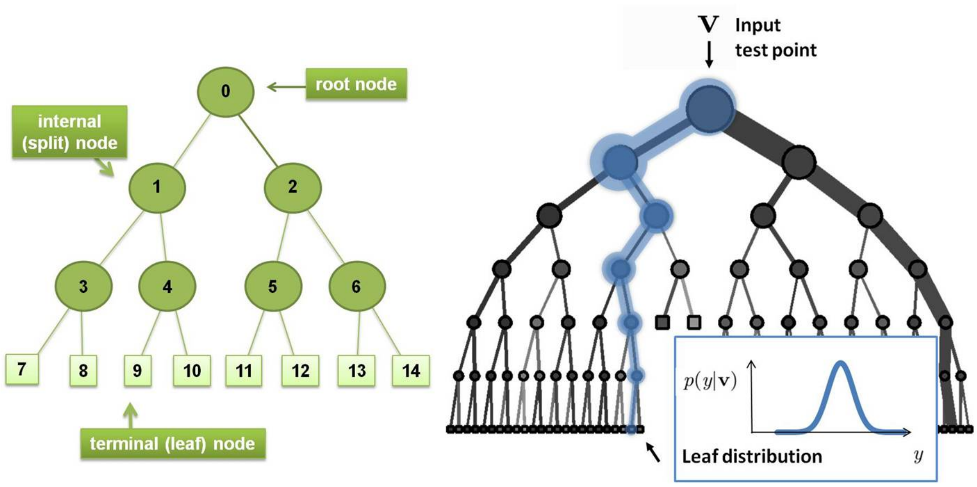


Figure 4.1: (Left) General structure of a decision tree. (Right) Testing of input data. Figures adapted from [50]

A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning by Criminisi, Shotton and Konukoglu [50] .

4.1.1 Framework

A random decision forest is a collection of decision trees. Each decision tree organises a series of questions called split functions which are embedded within the root and split nodes (Fig. 1 (Left)). The split function evaluates incoming data points $\mathbf{v}$ of the analysed signal. The user can define the features space $\mathcal{F}$ to characterise these data points. A data point is then denoted by a vector $\mathbf{v} = (x_1, x_2, ..., x_d) \in \mathcal{F}$, where x_i are its feature values. At each node, a feature value is compared with a node-specific threshold $\boldsymbol{\tau}$. Depending on the result, the data point is sent to the left or right child node. There, another feature and threshold is considered. The process continues until a leaf node is reached.

Training

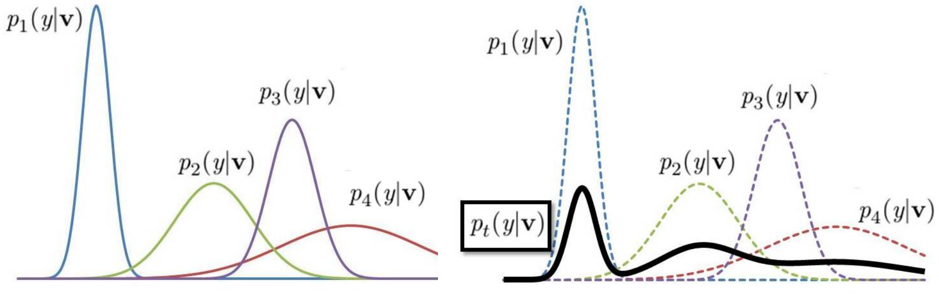
Forest trees are automatically constructed during forest training. A tree is built up incrementally from root to leaves using training data and an objective function. Training data consists of a set of data points $\mathcal{S}_0$ which are each associated with a label. A label is a discrete or continuous value adding information to the data point. During training, data is split to the left and right branch of a node. The objective function is measured at each node throughout the training procedure. Generally, the applied training objective function supports on the concept of information gain. It indicates how effective the input data and associated labels get distributed from parent node to child nodes. Tree training attempts to find those parameters that maximise the information gain at this node. Randomness is introduced by only presenting a subset $\mathcal{T}_j$ of the complete parameter space at each node optimalisation. Hence, the amount of randomness is determined by the ratio $|\mathcal{T}_j|/\ |\mathcal{T}|$.

During split node optimisation, nodes become leaves when one of the stopping criteria has been fulfilled. After conversion to leave nodes, each leaf node has received a set of training points. Training points in the same leaf are similar, meaning the feature values of these points are alike. They all passed the same split nodes and were evaluated with the same result by the split functions. Next, a label prediction model is derived from the labels associated with the data points within the leaf. The probability of a certain label at a leaf is proportional to the number of data points within this leaf associated with that label.

Testing

During testing, a test data point without a label is presented at the root (Fig. 4.1 (Right)). Based on its feature values, the split nodes guide the unseen data point to a leaf. This leaf contains training points with similar feature values as this new data point. It is reasonable that this test point should receive a label similar to the labels of the training points in this leaf. Thus the label of this test point is predicted with the label statistics of the indicated leaf.

Since each tree is trained independently and randomly, the trees are decorrelated. As a consequence, every tree has a different predictive outcome for the same test



(a)

(b)

Figure 4.2: Forest probability distribution. (a) Posterior probability distributions of four regression trees. (b) Forest distribution created by averaging all tree distributions.

data point. The posterior distributions of the identified leaves are summed and averaged out. In this way a single forest probability distribution is obtained from all the tree posteriors. For a continuous regression problem, the forest distribution is formulated as

$$p(y|\mathbf{v}) = \frac{1}{T}\sum_{t=1}^{T} p_t(y|\mathbf{v}) \quad , \tag{4.1}$$

where $p(y|\mathbf{v})$ is the forest distribution, $p_t(y|\mathbf{v})$ the posterior distribution of the tth tree and T the total number of trees in the forest. Figure 4.2 illustrates this principle of averaging probability distributions in the case of a continuous output variable y. For a test data point $\mathbf{v}$ the corresponding tree distributions are $p_t(y|\mathbf{v})$. Some leaves have distributions with a lower variance. Their prediction is more confident. The combined forest distribution is stronger influenced by more confident tree distributions. Hence the forest will follow the decision of more informative trees. This approach of decorrelation and combination assures the generalisation power and robustness of a random decision forest.

4.1.2 Implementation

The model of random decision forests is a framework suitable for a diversity of problems. To accommodate the framework to a certain type of problem, it requires specific properties of the data points. These are the features extracted from the data points and the labels associated with the data points. The set of features that is chosen and the type of labelling determine the outcome of the decision forest. The next sections present the features derived from the measured physiological signals and the labels used for classification.

4.2 Physiological signals

The NeXus 10-MKII (MindsMedia, Herten, The Netherlands) is able to monitor a number of different physiological signals. These include GSR, temperature, respiration rate, electromyography (EMG), blood volume pulse and heart rate. For data analysis in the controlled environment only GSR and temperature are taken into account. Other signals measured during experiments in the controlled environment are excluded because of the following considerations. Respiration rate and EMG are not measured in the ambulant phase, therefore analysis of these signals in the controlled environment will not contribute to the analysis in ambulant phase. Next, blood volume pulse does include information about heart rate and heart rate variability. However, an electrocardiogram (ECG) is more accurate as it is less prone to movement artifacts [51]. As ECG is only recorded by the wearable Health patch, BVP measured by the Nexus is not taken into account. Nexus also measures temperature for which a large feature set is already developed by Imec. Therefore temperature is included without further expanding the feature set. The remaining physiological signal is GSR. It is an important signal as it is less influenced by respiration than ECG [52]. Moreover, GSR seems least influenced by movement in comparison to other physiological signals [28].

4.3 Feature set of Galvanic Skin Response

As mentioned in section 4.1, the Random Decision Forest relies on features to build up a model. The feature set is based on both GSR and Temperature signals. An extensive feature set for Temperature is already developed by Imec. The feature set for GSR is extended in this section. A few GSR features were already provided by Imec (table 4.1). Another part of the GSR feature set was designed for this thesis (table 4.2), based on performant GSR features in [31]

Table 4.1: GSR feature set, implementation by Imec

SC L	mean SC level
SC PH	signal power in a phasic SC signal
SC RR	SC responses rate
SC DIFF2	signal power in second difference from SC signal
SC R	number of SC responses
SC MAG	sum of the magnitudes of SC responses
SC DUR	sum of the duration of SC responses
SC AREA	sum of the area of SC responses calculated as $(SCMAG \times SCDUR)/2$

Table 4.2: GSR feature set, additional to current feature set

Slope	slope of the regression line of the signal
Percentiles PH	percentiles of peak height: 25%, 50%, 75%, 85% and 95%
Mean PR	number of peaks per window (only peaks with height $\geq$ 50% taken into account)
Median PR	median of instant peak rates, instant peak rate calculated as $1/T$ with T number of seconds to previous peak (only peaks with height $\geq$ 50% taken into account)

4.4 Labelling of feature set

4.4.1 Objective labelling

During the stress experiment, two possible phases are alternated, a *Relax* phase and a *Stress* phase. In order to classify unseen data in these two phases, the training data is split and labelled according to the *Relax* and *Stress* phase. This results in a binary classification problem. As the labelling relies on the known alternation of *Relax* and *Stress*, it is named *objective* labelling in this thesis. This is the counterpart of *subjective* labelling, described in section 4.4.2.

In the same manner as for feature calculation, the data is divided into time

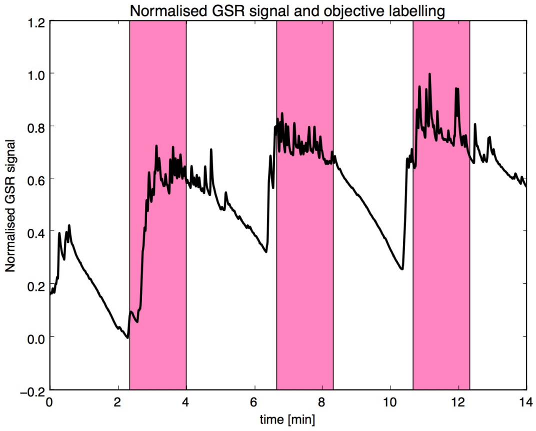


Figure 4.3: Objective labelling. White background denotes *Relax* phase and pink background denotes *Stress* phase. Black curve is the normalised GSR signal.

windows. For every time point, it is known whether the test person is in a relax or stress phase of the stress experiment and the window is labelled accordingly. The labelling can be seen in Fig. 4.3. A white background denotes *Relax* phase and pink background denotes *Stress* phase. The black curve is the normalised GSR signal. The first pink area corresponds to the *Stroop* test, the second to the *Calculation* test and the last pink area is the *Speech* test.

4.4.2 Subjective labelling

Subjective labelling is based on the perceived level of stress by the test persons during the stress experiment. Participants filled in a questionnaire after perfoming the stress tests. They indicated how stressed or relaxed they felt before, during and after the tests. A discrete scale of 1 to 5 is used where 1 indicates 'not at all' and 5 indicates 'very much'. An example of the questionnaire is given in table 4.3, with TP XY as a fictious test person.

Every test person fills out the questionnaire differently, with rather moderate or rather extreme values. Therefore the stress levels are normalised over every test

Table 4.3: Questionnaire after stress tests phase 1 with normalised stress level

	TP XY	rating	normalised stress level
1	How stressed before the start	1	0.0
2	How stressed during tests	2	0.33
3	How stressed during the *Stroop* test	3	0.66
4	How difficult is the *Stroop* test	4	/
5	How stressed during the *Calculation* test	2	0.33
6	How difficult is the *Calculation* test	3	/
7	How stressed during the *Speech* test	3	0.66
8	How difficult is the *Speech* test	4	/
9	How well relaxed during breaks	2	1.0
10	How stressed at this moment	1	0.0

person to a value between 0 and 1.

The normalisation takes into account all the ratings regarding stress levels (questions 1,2,3,5,7 and 10). Questions regarding difficulty of the test are not taken into account as this complicates the normalisation. Question 9 'How well relaxed during breaks' examines level of relaxation and not stress. Therefore the level of stress during breaks is calculated as

$$\text{stress level} = 6 - \text{relaxation level} \quad . \tag{4.2}$$

If a test person is very relaxed during breaks and rates a break with 5 ('very much'), the level of stress is 1 ('not at all').

The normalised stress level is then calculated as

$$\text{normalised stress level} = \frac{\text{stress} - \text{stress}.min}{\text{stress}.max - \text{stress}.min} \quad . \tag{4.3}$$

Based on the subjective rating in the questionnaire, the different phases of the stress experiment are labelled. Unlike objective classification in section 4.4.1 it is possible to work with multiple levels of stress. The subjective stress levels are illustrated in Fig. 4.4. The red curve depicts the normalised level of stress for every time window and the black curve is the normalised GSR signal. The background colors only indicate the *Relax* and *Stress* phases which have no influence on the subjective labelling.

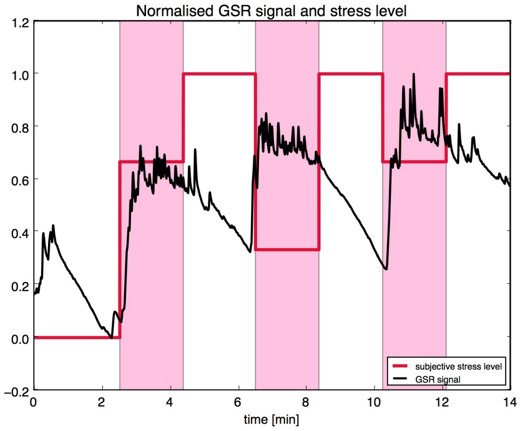


Figure 4.4: Subjective stress level. Red curve depicts the normalised level of stress for every time window. Black curve is the normalised GSR signal.

A binary labelling can be obtained by rounding the normalised stress levels to 0.0 or 1.0 and assigning these values to time windows as labels (Fig. 4.5). A three-level classification is obtained by rounding normalised stress levels to 0.0, 0.5 or 1.0. A four-level classification means rounding to level 0.0, 0.33, 0.66 or 1.0 and a five-level classification to 0.0, 0.25, 0.5, 0.75 or 1.0. With more possible levels, the difficulty of classification increases as classification needs to be more precise. This means the training set needs to be large enough and contain enough samples of every class to train the model well. As the given dataset with 13 test persons is fairly small, a binary classification is performed.

4.5 Classification

Classification is assigning samples from an unseen data set to one of the possible classes [53]. The quality of the model is validated by the performance of the classification outcome. Two models are tested for objective and report-based classification based on different labelling. They are validated by a leave-one-participant-out (LOO) cross validation scheme. This means the dataset is split in *n* folds, with *n* being the number of participants. The training set to train the RDF

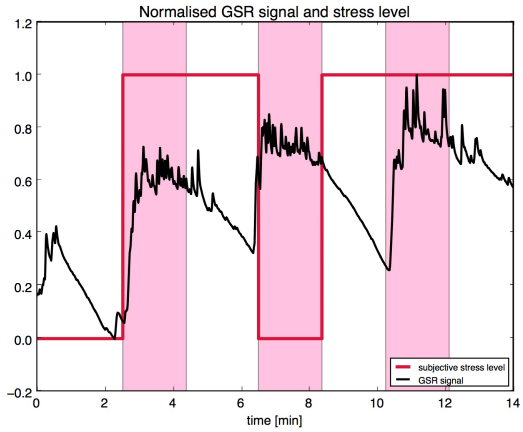


Figure 4.5: Subjective labelling. Red curve depicts the normalised level of stress for every time window, clipped to 0.0 or 1.0. Black curve is the normalised GSR signal.

Table 4.4: Definition of sensitivity and specificity

True positive	TP	number of *stress* samples classified as *stress*
False negative	FN	number of *stress* samples classified as *relax*
True negative	TN	number of *relax* samples classified as *relax*
False positive	FP	number of *relax* samples classified as *stress*
Sensitivity		TP/(TP+FN)
Specificity		TN/(TN+FP)

model consists of n-1 folds. Testing is performed on just one fold. This procedure is repeated *n* times, testing every single fold. For the dataset used in this thesis, one fold contains the data of one test person. The classification outcome of the LOO cross validation is the average of n testing procedures.

The implemented measure for classification performance of one test person is the average of sensitivity and specificity. Sensitivity is a ratio indicating how many *stress* samples are classified as *stress*. Specificity is a ratio indicating how many *relax* samples are classified correctly. The calculation of both measures is shown in table 4.4.

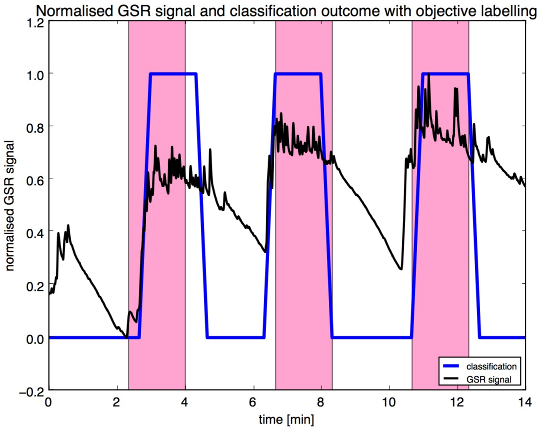


Figure 4.6: Objective classification. Blue curve represents the classification outcome. Black curve is the normalised GSR signal.

4.5.1 Objective classification

Objective classification is a classification procedure based on objective labels. First, the feature set of for GSR signals is used.

To maximise the classification performance, the window length and window shift are optimised. Therefore, the model is run with different sets of window length and window shift. The iteration over the parameter sets is repeated three times. Next, classification outcomes are averaged over these three iterations. The maximal classification performance is reached for a window length of 50s and a window shift of 20s. With these parameters, the averaged classification outcome of LOO cross validation only using GSR is 82.66%. The outcome of classification is depicted in figure 4.6. *Stress* and *relax* phases are indicated by background color, toghether with the normalised GSR signal. The blue curve represents the classification outcome.

The classification procedure is repeated in the same manner, though adding the Nexus temperature signals and its features. The calculated features are mean, median, standard deviation, maximum, absolute mean, skewness, kurtosis, variance, amplitude range, interquartile range and area of the signal. Window length of 50s and window shift of 20s are applied. The averaged classification outcome of LOO

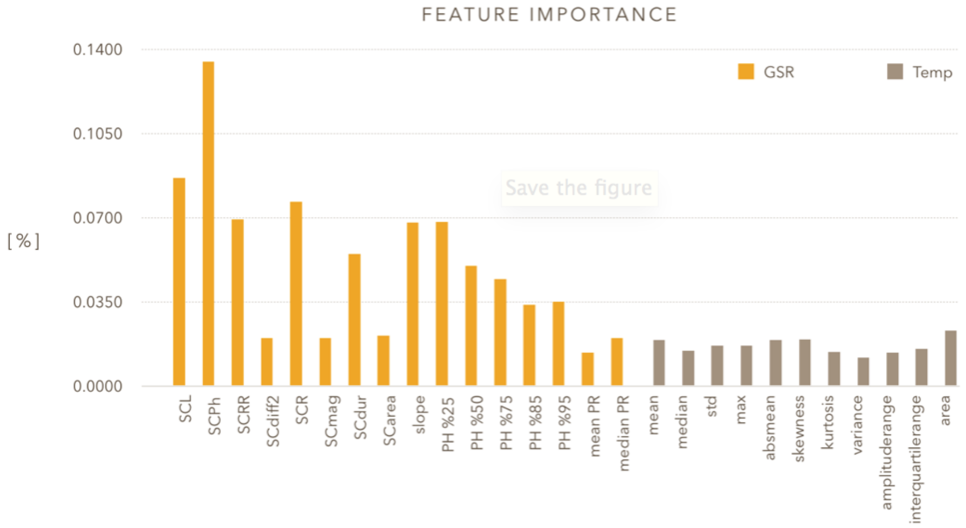


Figure 4.7: Feature importances during RDF training with features derived from GSR and temperature

cross validation is 82.36%. The performance is in the same range as when only using GSR features for classification. This result is explained by the fact that the RDF model mainly selects GSR features for classification. As an example, the importance of every feature during a classification test of one user is extracted. The sensitivity in this test is 94.44% and the specificity is 89.29%. Their average is the classification performance, 91.86%. The feature importances of all features of physiological signals GSR and temperature are displayed in Fig. 4.7. As adding information of the temperature signal does not increase classifcation performance, classification is based only on GSR signals.

4.5.2 Report-based classification

Report-based classification is based on the same feature set as for objective classification, though labels are subjective.

As mentioned in section 4.4.2 a two-level classification is chosen. Therefore the same performance measures as in section 4.5.1, sensitivity and specificity, can

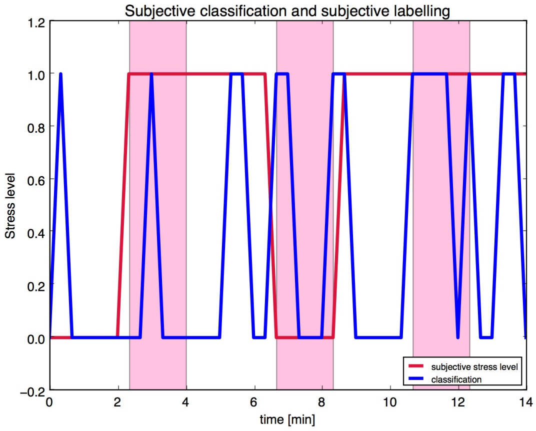


Figure 4.8: Report-based classification. Red curve is the subjective labelling. Blue curve is the classification outcome.

be used. Also the same window settings are used, a window length of 50s and a window shift of 20s. Classification performance reaches only 40.46% with averaging over LOO cross validation results. An example of report-based classification can be seen in Fig. 4.8. The red curve is the subjective labelling and the blue curve is the classification outcome. White and pink background indicate *Stress* and *Relax* phases.

4.6 Controlled environment with wearables

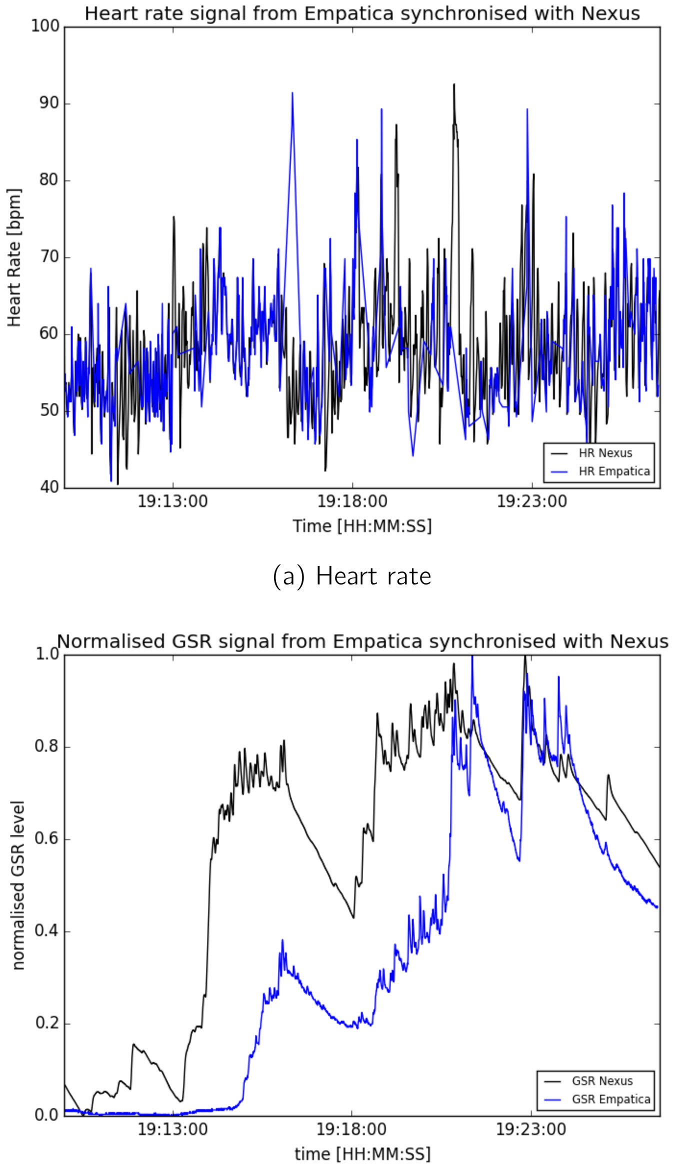
Previous section 4.5.1 showed that in a controlled environment with state-of-the-art material the developed model reaches a good classification performance of 82.66%. In this section it is investigated if the same model can be applied in a controlled environment on signals from wearable sensors. To be able to apply the Nexus-based classification model onto wearable sensors, the physiological signals from Nexus and wearable sensor need to be similar. Therefore their correlation is reviewed first.

The first subsection examines the correlation of the GSR signal from the Nexus sensor with the Empatica wristband. The second subsection examines the correlation between the heart rate signal derived from the Nexus sensor and from the Health patch.

4.6.1 Empatica

During the stress experiments, GSR signals from Nexus and Empatica are measured simultaneously. However, timestamps in the data logging from both sensors may differ as their internal clock is not synchronised perfectly. To compare GSR signals from Nexus and Empatica, the signals first need to be synchronised. Synchronisation cannot be performed by matching the signal appearances. The GSR signal from the Nexus sensor is measured at the fingertips while the GSR signal from the Empatica sensor is measured at the wrist. This means the signal differs as they are measures at different sweat glands. Therefore synchronisation is based on matching the heart rate signal from both sensors as heart rate is the same throughout the body, independent from the measurement location (Fig. 4.9a). From the synchronised heart rate signals, the timeshift needed for synchronisation of the GSR signals can be deduced (Fig. 4.9b).

The GSR signal from both sensors show some similarity. However, this is only the case for two test persons of the data set. The measurements by the Empatica wristband often lack quality, as for instance only a few datapoints are recorded. As measured GSR signals from Nexus and Empatica differ too much, the RDF model (sec. 4.1) trained on Nexus data cannot be applied onto Empatica signals. Moreover, the RDF model cannot be trained on GSR signals from Empatica as there is not enough qualitative data available. It is concluded that the RDF model developed in the controlled environment, cannot be applied on the Empatica data.



(a) Heart rate

(b) GSR

Figure 4.9: Synchronisation of Nexus and Empatica based on heart rate with resulting synchronised GSR signal.

4.6.2 Health patch

The synchronisation procedure is repeated for the wearable Health patch. As the Health patch measures ECG solely, the synchronisation is only performed for heart rate. Heart rate is the same throughout the body, thus it should provide good synchronisation possibilities. Adequate synchronisations were attained for seven test persons. The result for one of the test persons is depicted in Fig. 4.10. A close-up of an ECG signal can be seen in Fig. 4.11 .
The Nexus is a reliable and stable sensor and can be seen as the ground truth sensor for the wearables. As the Nexus and Health patch signal ressemble so well after synchronisation, the signal from the Health patch is of good quality. Thus, it is possible to use the Health patch signals for further data analysis.

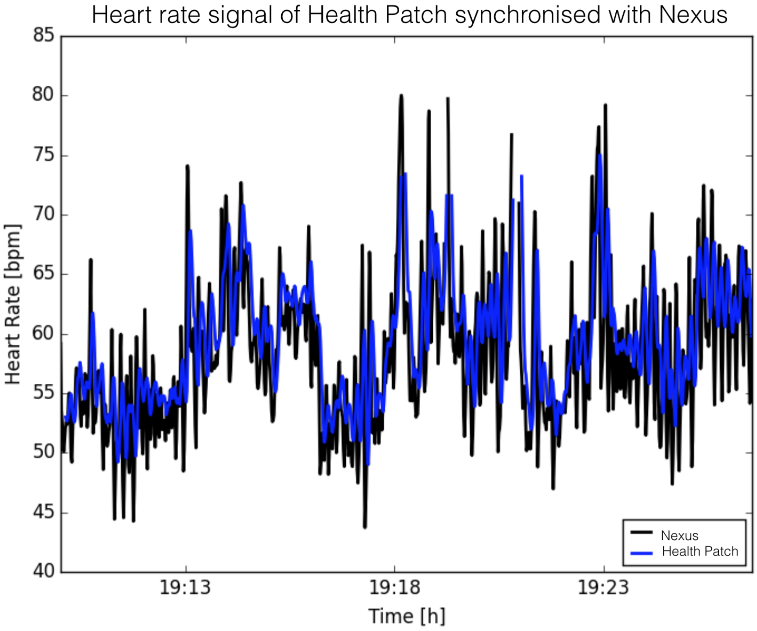


Figure 4.10: Synchronisation of Nexus and Health patch based on heart rate signal.

In order to classify on ECG signals, heart rate variability features in time and frequency domain are calculated. These features are listed in table 4.5. They are

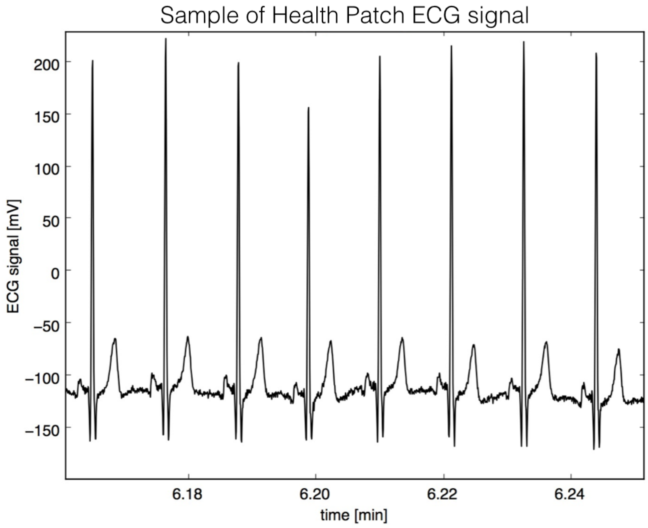


Figure 4.11: Detailed view on ECG signal of Health patch.

based on the *inter-beat interval* or so-called *RR interval* of the ECG signal. The peaks in the signal are called *R peaks* and thus the time interval between two peaks is the *RR interval*. The mean heart rate is calculated as

$$\text{mean HR } = \frac{60 \times f_{sampling}}{\text{average RR interval}} \quad \text{[bpm]} \ .$$

As derived in section 4.5.1 a window size of 50s and window shift of 20s is applied, giving optimal test results. The ideal window size and shift might differ for ECG signals, but is not taken into account as an adaptive window size and shift implies overfitting the data. Classification based on the ECG signal of seven test persons measured by the Health patch sensor reaches an average performance of 76.89%. The average is calculated out of three repeated LOO cross validations. The importance of every feature during training, of one LOO iteration, can be seen in Fig. 4.13. Based on the graph, the most important features are two time domain features. These are *mean heart rate* and *RMSSD*. The fact that these features are

Table 4.5: ECG feature set

Time domain features

mean HR	mean heart rate [bpm]
SDNN	standard deviation of all normal RR intervals (i.e. NN intervals) [ms]
RMSSD	root-mean-square successive difference of all normal RR intervals [ms]

Frequency domain features

LF HRV	low frequency HRV (power in the 0.04–0.15 Hz band)
HF HRV	high frequency HRV (power in the 0.15-0.4 Hz band)
LFHF HRV	ratio (LF HRV) / (HF HRV)

equally important is because of their high negative correlation. This can be visually confirmed in Fig. 4.12 and has a value of -0.9934 for TP01. As a comparison, the correlation between *mean heart rate* and *LFHF* is 0.3305. A classification example of one test person is depicted in Fig. 4.14 with the normalised mean heart rate signal. The signal has a clear increase during stress tests and decrease during relax phases.

Stress classification based on GSR signals of the Nexus sensor reached on average 82.66%, by training and testing on a dataset of 11 test persons. Seen the fact that only seven test persons are included in the ECG dataset, the ECG based classification result of 76.89% is promising. Smets et al. [45] had a similar experimental set-up and reported a maximum performance rate (for non-personalized models) of 79.2% using multiple physiological signals and RDF. The trained RDF model would get more robust when the input data set increases and thus higher classification results would be expected. Considering the current result as satisfying, a stress classification can be done based on ECG signals measured by Health patch. Therefore the classification procedure in the ambulant phase will also include ECG based features.

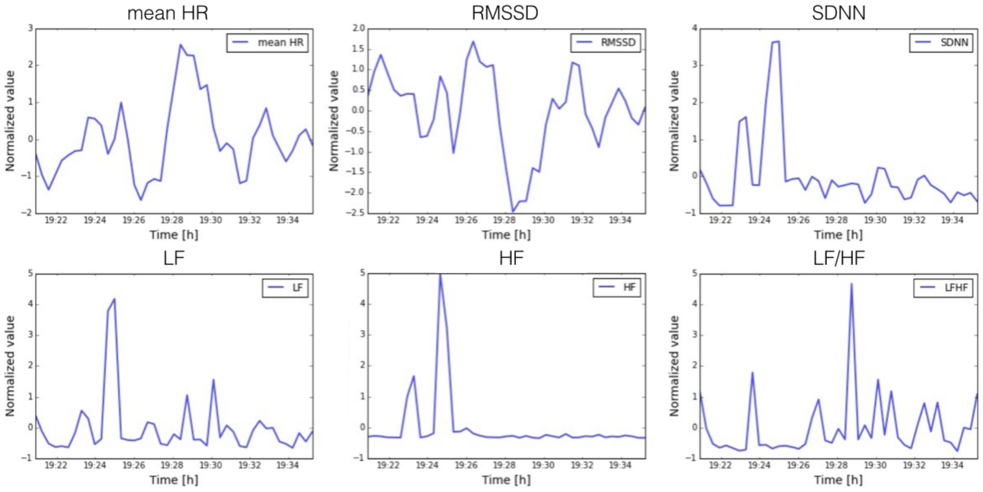


Figure 4.12: Normalized features of ECG Health patch.

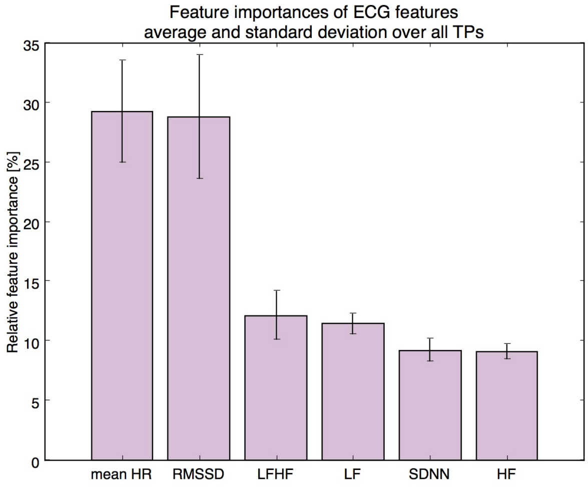


Figure 4.13: ECG feature importance during RDF training in ranked order.

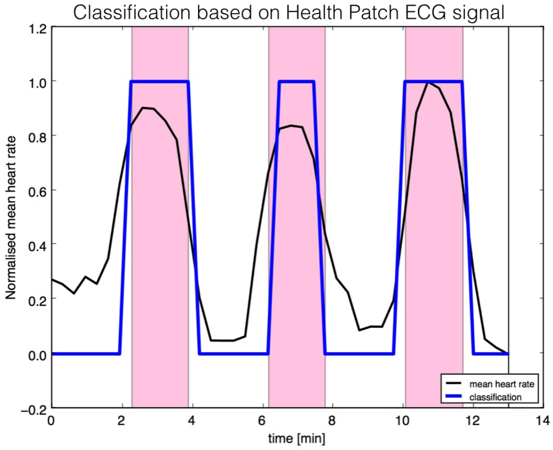
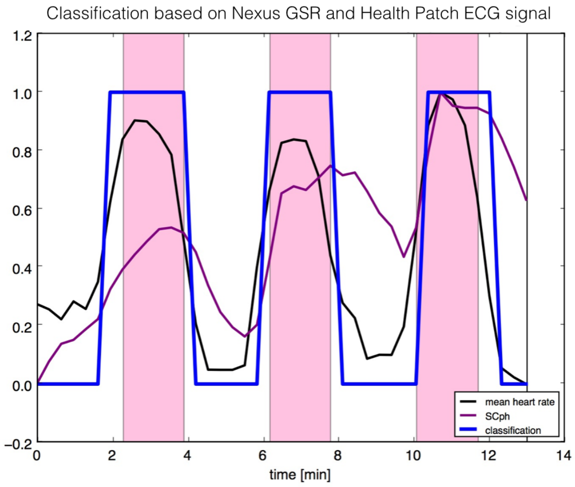


Figure 4.14: Classification on ECG signal of Health patch (blue) with normalised *mean heart rate* (black) as one of the most important features.

4.6.3 Combined model based on GSR Nexus and ECG Health patch

Classification based on the GSR signal of the Nexus sensor reaches an average performance of 82.66%. Classification based on the ECG signal of the Health patch sensor reaches a lower average performance of 76.89%. It is investigated if a combined model based on both GSR of Nexus and ECG of Health patch reaches a higher performance. A disadvantage of this study is that the dataset can only contain seven test persons, as the ECG signal of only seven test persons could be synchronised with the Nexus sensor. Nonetheless, an averaged classification performance of 83.04% is reached. This result outperforms the classification based on ECG features only, with an extra 6.15%. This means GSR features greatly contribute to the classification model. Moreover, the performance of the combined model is higher than the model based on solely GSR signals of Nexus. Thus ECG features contribute positively to the classification result. The difference in performance is on average less than one percent, though one has to take into account the relatively large difference in data set. Classification outcome for one test person and feature importances for the RDF model are displayed in Fig. 4.15. In this test case, the classification rate is high with 92.87%, a sensitivity of 94.44% and specificity of 91.30%. The most important feature is *mean heart rate* of the ECG signal. Second and third are *phasic skin conductance* and *skin conductance* *level*. The GSR features form the basis of the RDF model, with the mean heart rate as a crucial feature.



(a)

(b)

Figure 4.15: Classification of TP05 with GSR of Nexus and ECG of Health patch. Performance reached sensitivity of 94.4% and specificity of 91.3%.

4.7 Conclusion

The *Random Decision Forest* model has proven to be a suitable model as it does not require prior feature selection. Therefore all derived features could be fed to the model, without losing accuracy.

First, the input signals of Nexus were analysed by feature derivation and RDF training and testing. Outcome of objective classification with GSR features was 82.66%. A comparable result was found with GSR and temperature features, indicating temperature features did not add information to the model. This was confirmed by reviewing the feature importances during RDF training. The report-based classification performed poor with a performance of 40.46%. This confirms that there is a weak relation between physiological sensor measurement and subjective stress levels.

As the RDF model performed well with Nexus data and objective labels, the subsequent step was to apply the procedure onto wearable data. Not enough qualitative data of Empatica was available. Therefore only data of the Health Patch was analysed by several ECG features, which reached a classification performance of 76.89%.

Combining GSR features from Nexus and ECG features from Health Patch was found to be performant with a classifiation performance of 83.04%. GSR features contributed most to the classification model, aided by the ECG features *mean heart rate* and *RMSSD*.

The results found with supervised learning with RDF were successful and serve as the gold standard for unsupervised learning further on.

5 Unsupervised learning

Ambulant data is measured with wearable sensors during the daily life activities of test persons. This has consequences related to the acquired data and related to the gold standard of labels.

Wearable sensors are less accurate as they generally have a lower sampling rate as fixed bulky sensors. Also, wearable sensors might detach as a person is moving and measuring points are lost. Moreover, movement of the test person adds physical activity in the signal and influences the acquired data.

Another problem of the ambulant dataset is the given questionnaire as it only provides a course, unreliable labelling of the data. The questionnaire demands a score for stress and complaints every hour. First of all, the scoring is subjective. Second, hourly labelling is very unprecise. Third, it is possible a test person only fills in the questionnaire at the end of the day. Therefore, ambulant data requires a different approach for detection of stress as opposed to sensor data from a controlled environment.

Although many studies have already reported these issues regarding subjective labelling in an ambulant environment [20, 4], these problems are rarely addressed in the analyses.

In the current research an algorithmic pipeline based on the unsupervised learning algorithm *Self-Organizing Maps* is presented in order to rule out the subjective labels. This technique only relies on the physiological feature data, not on the corresponding labels.

The first stage is mapping the higher-dimensional feature space onto a two-dimensional grid, while preserving the topological relationships within the data. The algorithm is based on the Self-Organizing Map (SOM), explained in section 5.1. Measurement data is mapped by to algorithm to different areas on this grid. The second stage consists of exploring these areas and clustering them into defined regions, as seen in section 5.2. Next, the statistics of the clusters are analysed to locate regions in the grid that will correspond to *relax* and *stress* phases. To validate the functional concept of this algorithmic pipeline it is applied onto the dataset of the controlled environment. The outcome of the SOM and clustering is compared to the objective labels to calculate its performance for stress dection (section 5.3).

5.1 The Self-Organizing Map

Self-Organizing Maps represent higher-dimensional data as a globally ordered two-dimensional map. The SOM can be seen as an elastic grid of nodes fitted to the input signal space, while preserving the topological relationships of the signal space [54] .

Here, the input signal space is an n-dimensional feature space. Every node i is associated with a weight vector $w_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{in}]^T \in \mathbb{R}^n$. The input feature vector is $x_{stim} = [\xi_1, \xi_2, ..., \xi_n]^T \in \mathbb{R}^n$ [55]. The feature vector x_{stim} is mapped to the *best-matching node* c by comparing it with all weight vectors w_i. As a metric of similarity, the smallest Euclidean distance is searched:

$$c = \arg\min_i ||x_{stim} - w_i|| \tag{5.1}$$

During training, topological relationships of the input feature space are projected onto the two-dimensional SOM by adapting the weight vectors w_c. Nodes that are topographically close to the *best-maching node* c are also activated to learn from the same input x_{stim}. This results in a smoothing effect on the weight vectors of nodes in the neighbourhood and eventually leads to *global ordering* of the map. Input vectors are presented to the map in a random order. Given x_{stim} at time t, the update of the weight vector w_i of node i is as follows:

$$w_i(t+1) = w_i(t) + h_{ci}(t)[x_{stim}(t) - w_i(t)]. \tag{5.2}$$

The initial values of the $w_i(0)$ can be arbitrary.

The neighbourhood function $h_{ci}(t)$ can be defined in terms of the Gaussian function:

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}\right), \tag{5.3}$$

with $0 < \alpha(t) < 1$ the *learning-rate factor* and σ the width of the kernel, both decreasing monotonically in time. $r_c \in \mathbb{R}^2$ and $r_i \in \mathbb{R}^2$ are the location vectors in the SOM of nodes c and i, and with increasing $||r_c - r_i||$, $h_{ci} \to 0$.

Self-Organizing Map of lab data

To better visualize and understand the functionality of the SOM, the input feature space is two-dimensional, thus consisting of two features. The chosen features are those found most important for stress detection in Chapter 4. These are phasic skin conductance (SCph) derived from the Nexus GSR signal and mean heart rate from the Health patch ECG signal (Fig. 4.15b).

A leave-one-participant-out procedure is applied for training the SOM. The learning rate α in Eq. 5.3 is set to 0.05. Feature vectors from all TP, except one, are fed to the SOM until convergence. Weight vectors of nodes are adapted during training, meaning they shift position in the feature space. The trained SOM after a different number of training iterations is illustrated in Fig. 5.1. It is noted that every figure represents the training of a SOM that started from different random initialisation weights. Thus these figures are not part of the same training sequence. The node colour represents the node's position along one feature axis in feature space. For illustration the axis of *SCph* is chosen. The brighter red its colour, the more a node is shifted towards higher *SCph* values. The darker blue, the more a node is shifted towards lower *SCph* values. The evolution along the axis of *mean HR* is similar.

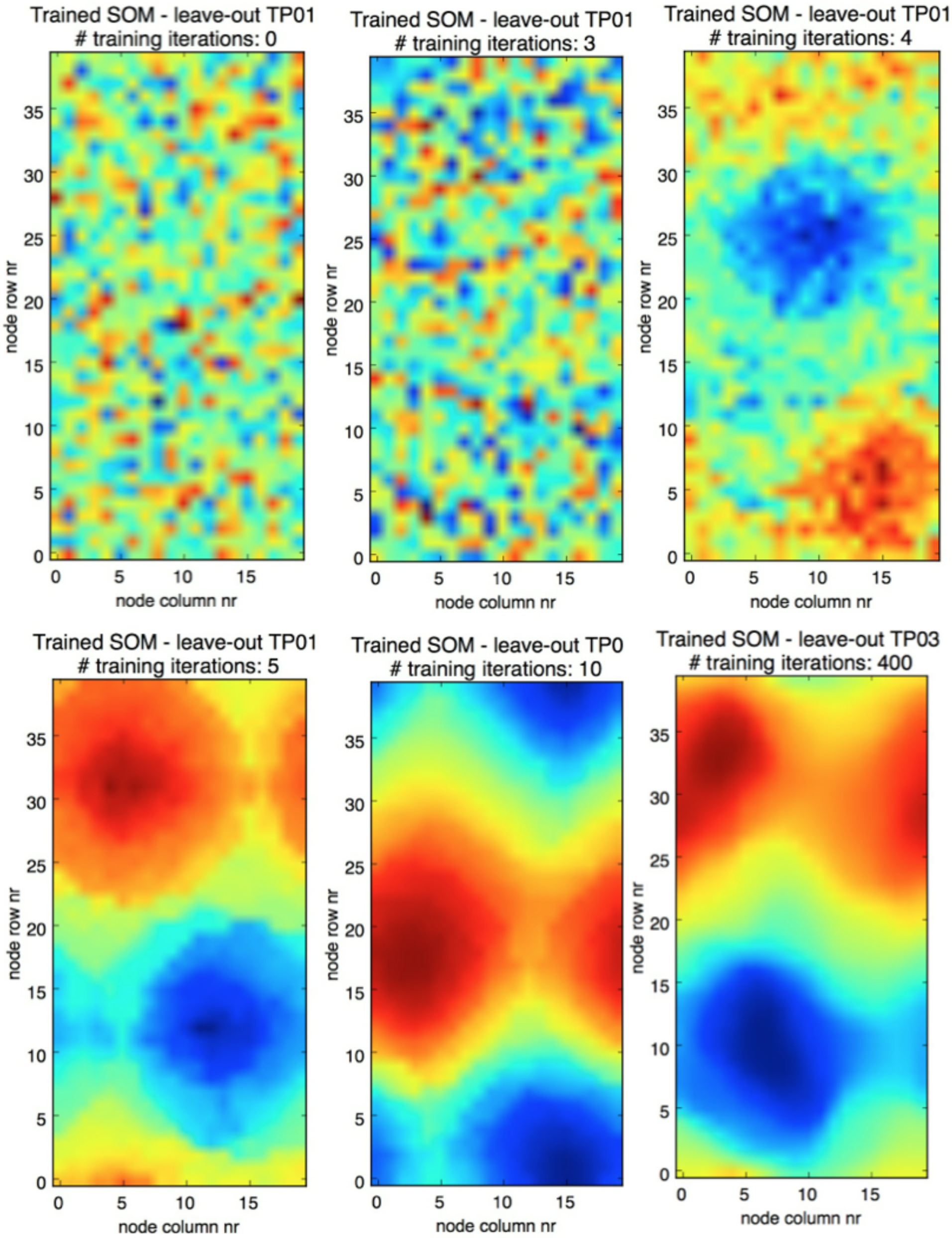


Figure 5.1: Training of SOM after different number of iterations.
Every figure started from a different set of initial random weights.
Smoothness of clustering increases with number of iterations.

Initially, node weights are random, thus nodes have random positions in feature space. At three iterations, several nodes became dark blue, meaning that they shifted towards lower *SCph* values. After four iterations, it can be seen that neighbouring nodes start pull each other to similar locations in feature space. Already at five iterations a clear pattern of two clusters emerges in the SOM. As declared before, the SOM becomes smoother with an increasing number of iterations.
The maximum number of iterations is set to 400, at which the SOM should be converged. The nodes of the SOM are settled at a location in feature space. The goal is that some nodes are driven to locations in feature space characterizing *stress* and others to areas characterizing *relax*. The next step is to outline these areas by clustering.

5.2 Clustering

Patterns have emerged in the SOM. The goal of clustering here is to outline these patterns. Every node of the SOM is assigned to a cluster by a clustering method based on Variational Bayesian Gaussian Mixture [56]. The advantage of this algorithm is the possibility to define a maximum number of clusters. The implementation is based on the scikit-learn package of Gaussian Mixture Models [46]. By varying the concentration parameter *weight_ concentration_ prior*, the effective number of active components (i.e. number of clusters) can be influenced. Setting this parameter to a low value will make the model put most of the weight on just a few components. The weights of remaining components are set very close to zero. Despite the automatic search for an optimal number of clusters, the main interest is to find a *stress* and a *relax* cluster. Therefore, the maximum number of active components is set to 2 and the *weight_ concentration_ prior* to a high value of 10^2. The algorithm is forced to find two clusters. In a later stage, the parameters could be relaxed to find multiple clusters, representing for example different levels of stress. Furthermore, the weights are initialized random (*init_params* to *random*). The *mean_precision_prior* is set to 10^{-2}. It will concentrate the means of each clusters around the mean of the node positions. Other parameters are left at default, such as the *covariance_prior* is *None* as no assumptions are made on the covariance of the clusters.

Examples of clustering outcomes on a trained self-organizing map can be seen in Fig. 5.2. Labels A (blue) and B (purple) represent the assigned cluster of the nodes. The form of the clusters correspond to the colors of the SOM, representing the distribution of feature values of *mean heart rate*.

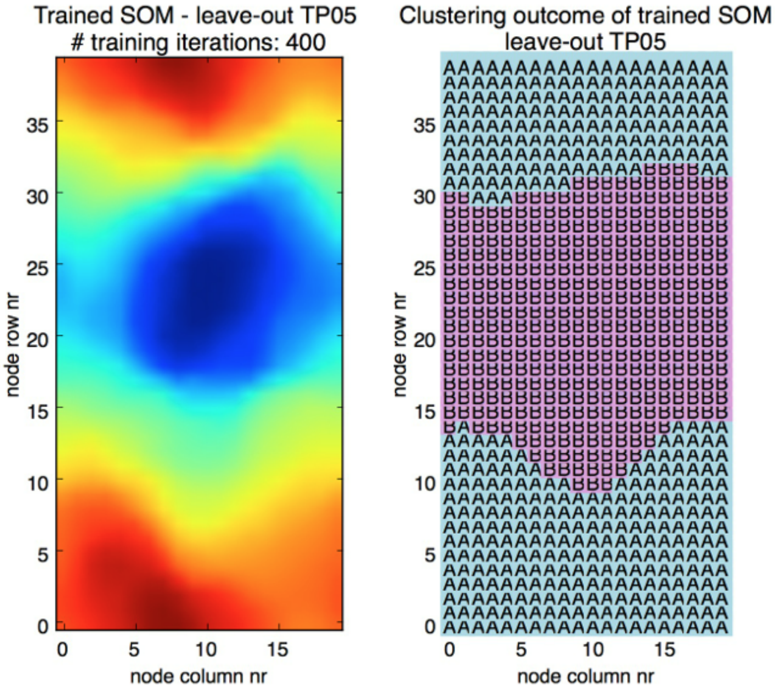


Figure 5.2: Clustering outcomes on trained SOM. Left: colors represent distribution of feature values on *mean heart rate*. Right: Labels A (blue) and B (purple) represent the assigned cluster of the nodes.

Cluster identification

Training of the Self-Organizing Map is unsupervised. Thus it is not known to which state a cluster belongs, *stress* or *relax*. To derive the state of a cluster, statistics from both clusters are analysed. The location in feature space, thus feature values, of all nodes within one cluster are gathered. These are represented in a boxplot. This is repeated for every cluster from each trained SOM (Fig. 5.3).

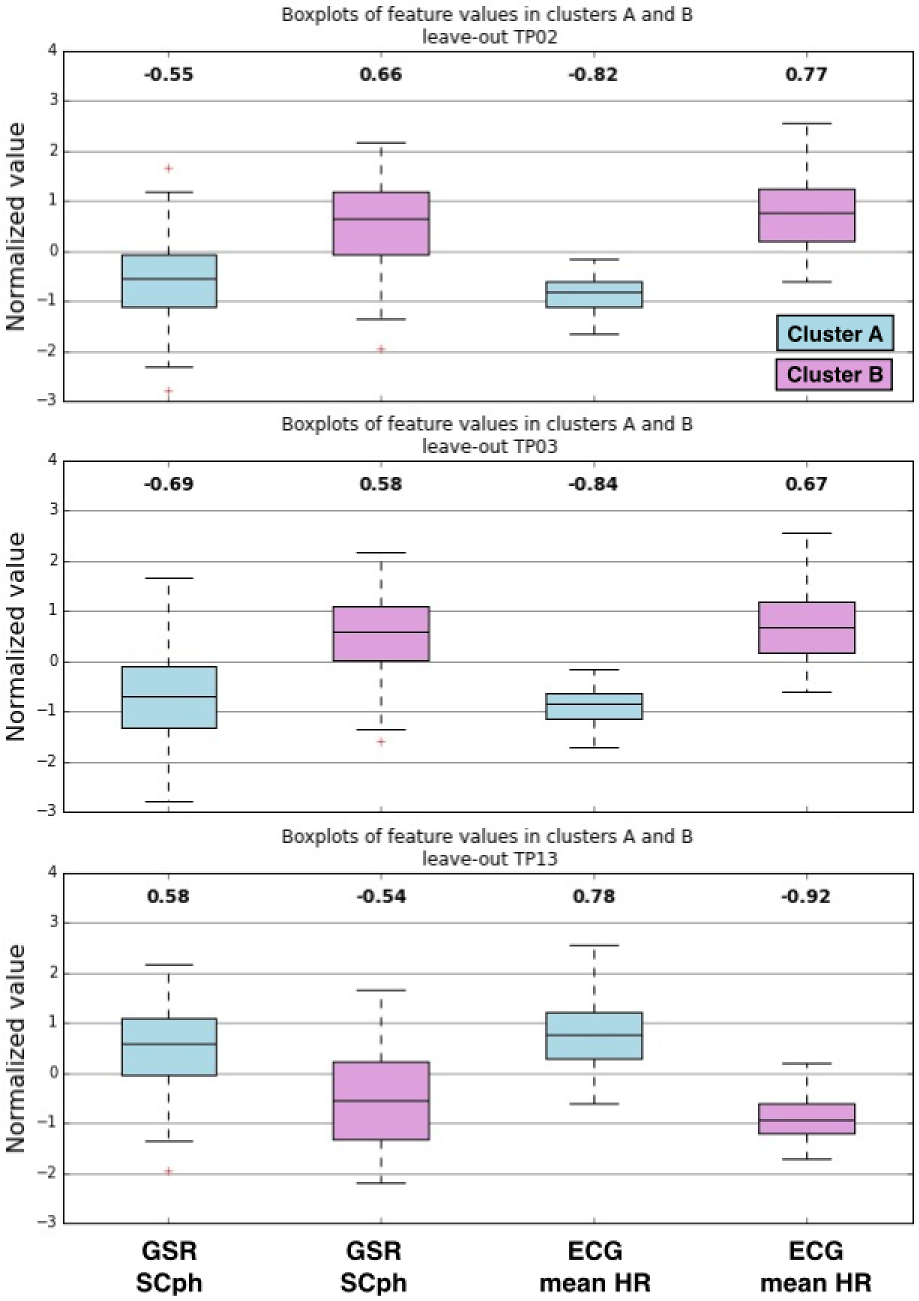


Figure 5.3: Boxplots of features values within clusters. Depending on the dataset, cluster A contains high *SCph* and *mean HR* values or it contains lower *SCph* and *mean HR* values. Vice versa for cluster B.

A clear pattern can be seen in between boxplots of different test persons. One cluster has high *SCph* and *mean HR* values, while the other cluster contains lower *SCph* and *mean HR* values. From Fig. 4.15 it can be derived that during stress tests both *SCph* values and *mean HR* are high. This prior knowledge is applied to define to which state the cluster boxplots of TP01 belong. Boxplot on the left is recognised by prior knowledge as corresponding to *relax* and boxplot on the right as *stress*. Cluster boxplots of TP01 is taken as a reference (Fig. 5.4). Cluster boxplots of other TPs are compared to these of TP01. Corresponding boxplots define corresponding states of the clusters. To determine corresponding clusters of TP01 and TPxx, the Root Mean Squared Error (RMSE) between the average of boxplots are compared. Corresponding clusters have a minimum RMSE between their boxplots.

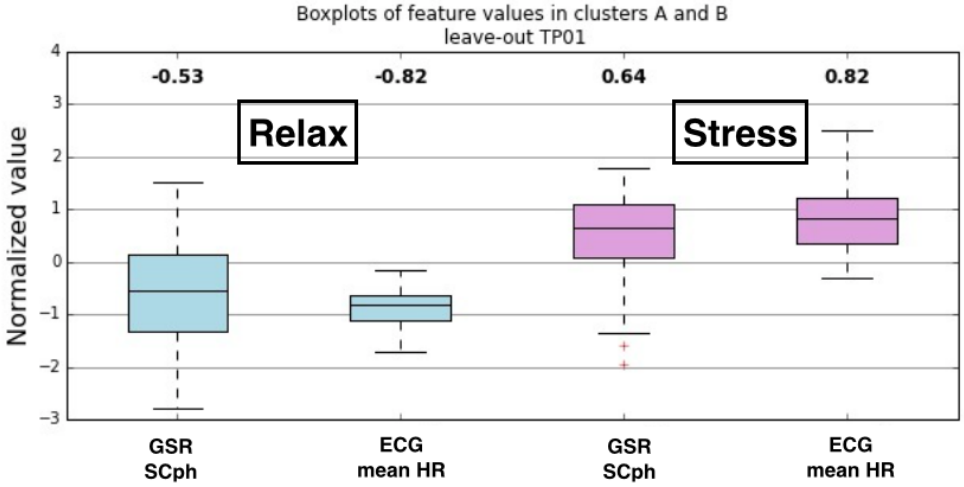


Figure 5.4: Boxplots of features values within clusters, from data of leave-out TP0. These boxplots are taken as reference boxplots to compare with. Boxplot on the left is recognised by prior knowledge as corresponding to *relax* and boxplot on the right as *stress*.

The next step is to calculate the performance of the SOM and clustering procedure. For simplification of this step, clusters are assigned such that cluster A of every SOM will be corresponding to the *relax* state.

5.3 Performance

To determine the feasibility of the algorithmic pipeline for stress detection, a performance measure has to be calculated, based on the feature set *GSR - SCph* and *ECG - mean HR*. First, it is examined how well the SOM has been clustered (section 5.3.1). Second, the testing performance is calculated on how well the algorithm can detect stress based on a clustered SOM (section 5.3.2). No labels have been used for training the SOM, nor clustering. The objective labels are introduced exclusively for performance calculations. The performance rate or classification performance is the average of sensitivity and specificity, as explained in table 4.4. A similar leave-one-participant-out (LOO) validation as in Chapter 4 is applied.

As the testing performance measure is established, it is possible to determine the optimal grid configuration of the Self-Organizing Map (section 5.3.1).

5.3.1 Clustering Performance

To calculate clustering performance, the objective labels of the training test persons are compared to the delineated clusters. Label '0' corresponds to a *relax* phase and label '1' corresponds to a *stress* phase (sec. 4.4.1). The left figure in Fig. 5.5 shows the clusters with overlayed labels of the training test persons. It is noted that cluster names A and B have switched compared to Fig. 5.2 as cluster A is defined to represent the *relax* state. It is chosen to perform a LOO validation, thus leaving one TP out, as the SOM and subsequent clusters for testing purposes are also formed by leaving one participant out and not calculated based on all TP. After ten runs of LOO validation, the clustering performance has an average of 74.91% and a standard deviation of 6.36%.

5.3.2 Testing Performance

The calculation of testing performance is analogue to sec. 5.3.1, though using the labels of the testing TP. The right figure in Fig. 5.5 shows the clusters with overlayed labels of the test person. After ten runs of LOO validation, the testing performance has an average of 77.63% and a standard deviation of 7.82%. The testing performance is the average of sensitivity and specificity. The sensitivity, indicating the ability to recognize stress phases, has an average of 85.64% and specificity, the ability to recognize relax phases, has an average of 69.62% after 10 LOO runs. It is clear that the overall testing performance is decreased by the lower rates of specificity, the ability to detect *relax* phases. Making a clear distinction between *stress* and *relax* is generally a very difficult task. Although sensitivity can be seen as the more important factor, as the goal is to detect stress.

A performance of 77.63% by unsupervised learning is considered satisfying compared to 83.04% attained by supervised learning with GSR and ECG features. Therefore, the current approach of unsupervised learning with *Self-Organizing Maps* is explored further.

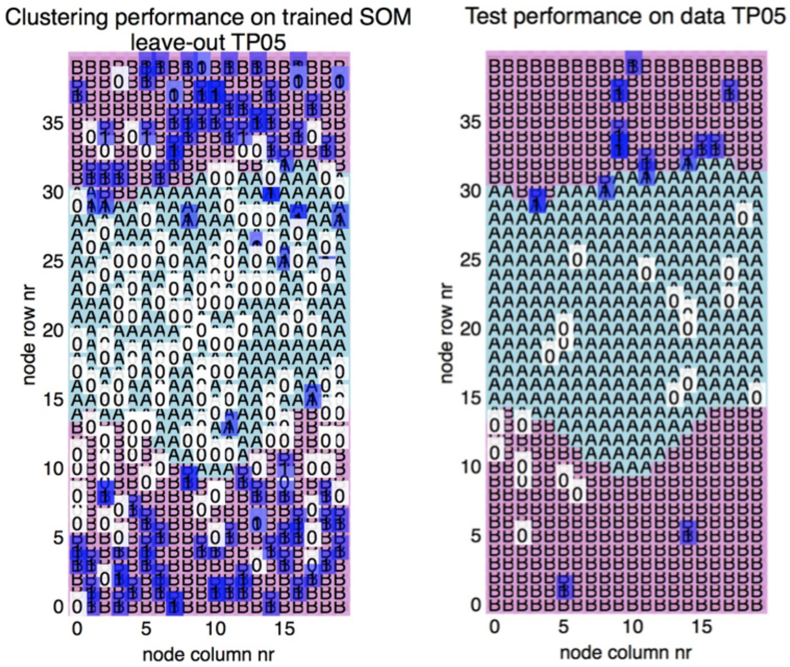


Figure 5.5: Left: Clustering performance. Right: Testing performance.
Labels A and B represent the assigned cluster of a node. Cluster A is associated with the *relax* state and cluster B is associated with the *stress* state. Labels 0 and 1 represent the objective labelling of section 4.4.1. They respectively correspond to *relax* and *stress* phases.

5.3.3 Optimal grid configuration

Using the subset *mean HR, RMSSD* on the laboratory data, a LOO validation has been run three times over a set of different grid parameters. All combinations of number of SOM rows and numbers of SOM columns in [10,20,30,40,50] have been executed. Averaging the outcome of three LOO runs, gave an optimal testing performance of 74.0% with standard deviation of 11.2% for a grid configuration of 40×20.

5.4 Stress detection in ambulant data

This section focusses on stress detection with the algorithmic pipeline based on *Self-Organizing Maps* in ambulant data. First, the ambulant input data is analysed by preprocessing it (sec. 5.4.1) and investigating the correlation between signals (sec. 5.4.2). Next, the input data is applied onto a SOM for training. Visually, it can be seen that samples from night-time are mapped onto the same area of the SOM and likewise for samples from day-time (sec. 5.4.3). Night-time samples are neglected and the nodes of the SOM are clustered.

The validation of the algorithmic pipeline consists of four phases. The quality of clustering is investigated by their cohesion and separation or *average silhouette coefficient* (sec.5.4.4). An optimal feature subset is defined both in terms of silhouette coefficients of ambulant data as in terms of classification performance on the laboratory data. Next, it is demonstrated that the outlined clusters are not random. Instead, similar patterns are repeated in clusters from different days and different participants (sec. 5.4.5). Moreover, these clusters represent *stress* and *relax* phases (sec. 5.4.6). Test samples are mapped onto the clustered SOM to detect *stress* or *relax*. The detected labels of these test samples are averaged over one hour intervals to derive average stress levels (sec. 5.4.7). The final step is to compare these detected predicted stress levels to the stress levels defined by participants (sec. 5.4.8).

Table 5.1: Percentage Health Patch data retained of high confidence and low activity. The threshold for low activity was set at a maximum of 0.04 for all three cases.

	% data with high confidence	% data with high confidence and low activity
confidence >0.6	71.27	49.99
confidence >0.7	63.88	47.21
confidence >0.8	54.45	42.86

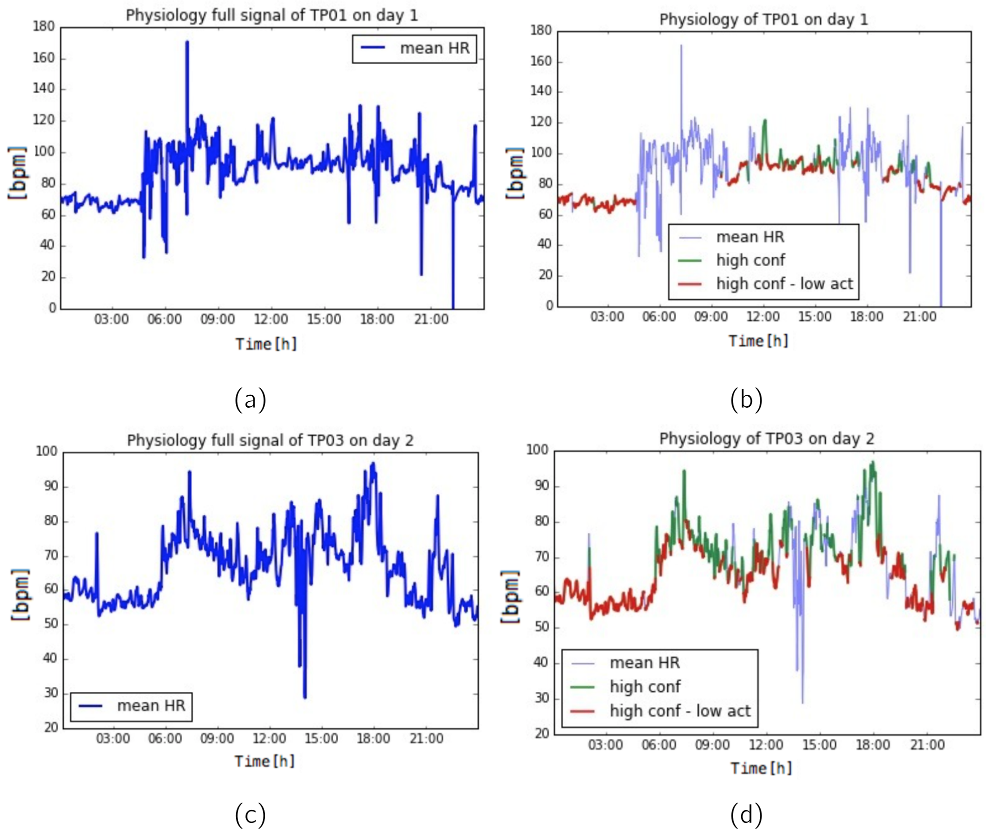
5.4.1 Preprocessing of ambulant data

As wearables are associated with several drawbacks regarding signal quality, it is highly necessary to perform a preprocessing of the data. Here, only an ECG wearable monitoring device is used, the Health Patch, which has a built in three-axial accelerometer. First, ECG data is retained which is regarded as data of high confidence, meaning data following the expected pattern of an ECG signal. The script for this procedure was provided by imec.

Next, accelerometer data was investigated. Data for which the standard deviation of the magnitude of acceleration was larger than a defined threshold was excluded. To calculate the standard deviation, a specific window size is used. Enlarging the window size will also enlarge the standard deviation. To obtain reasonable results, the standard deviation was allowed to variate between 0.02 and 0.04 for the given window size. The script and thresholds for this procedure were provided by imec.

The percentage of data retained after both steps averaged over all test persons can be seen in table 5.1. The first column is the threshold for data of high confidence. The threshold applied for low activity was 0.04. The column *% data with high confidence and low activity* represents the percentage of data retained from the full data set, not the data with high confidence. As can be seen in the table, less than half of the data is retained in all cases. Therefore, the threshold for low activity was set at a maximum of 0.04 to not loose more data. As the difference of total data loss between *confidence >0.6* and *confidence >0.7* is only about 2.5%, a confidence threshold of 0.7 was chosen, to retain as much and as confident data as possible.

In Fig. 5.6a and Fig. 5.6c the mean heart rate of the original ECG signal is displayed. In Fig. 5.6b and Fig. 5.6d the preprocessed signal is shown. The blue transparant curve represents the mean heart rate of the original ECG signal, the green curve is the signal after removing data with low confidence and the red curve is the remaining signal after removing high activity data from the confident signal. In Fig. 5.6b the ratio of remaining data is 58% for high confidence and 44% for high confidence low activity. In Fig. 5.6d the ratio of remaining data is 81% for high confidence and 55% for high confidence low activity. A large portion of the data is excluded, which impedes the stress detection.



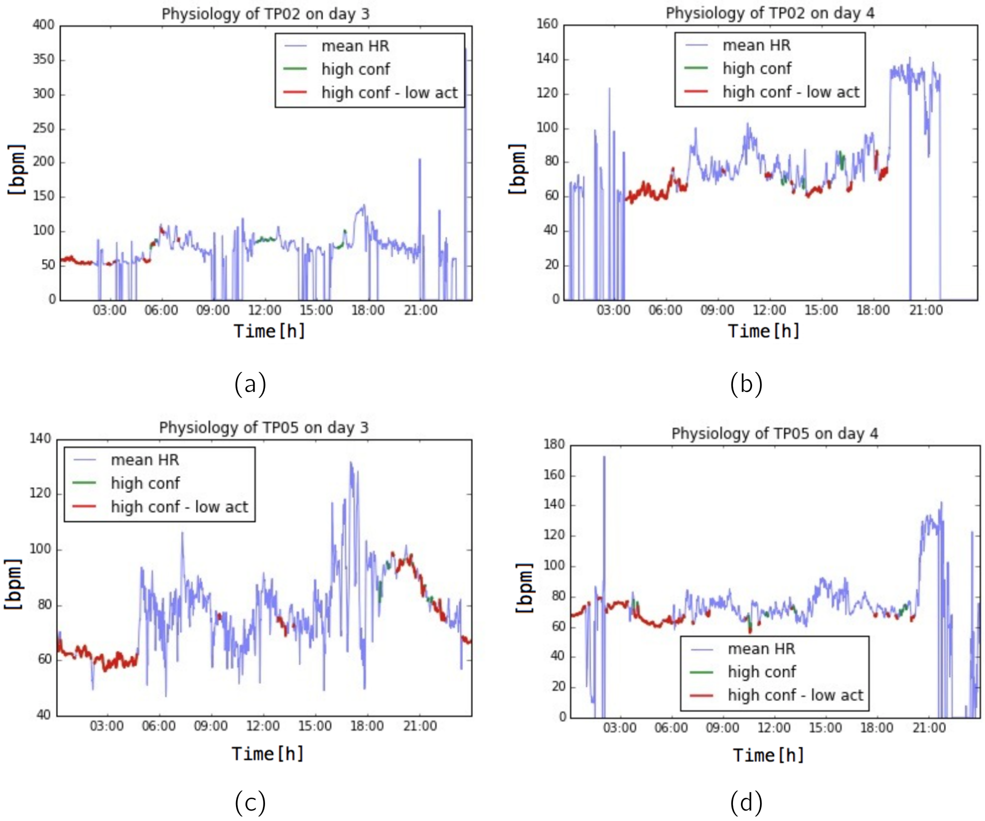
(a)

(b)

(c)

(d)

Figure 5.6: Fig. 5.6a and 5.6c represent the mean heart rate of the original ECG signal. Fig. 5.6b and 5.6d represent the preprocessed signals: The blue transparant curve represents the mean heart rate of the original ECG signal, the green curve is the signal after removing data with low confidence and the red curve is the remaining signal after removing high activity data from the confident signal.



(a)

(b)

(c)

(d)

Figure 5.7: Example cases in which the preprocessing of the Health Patch signal only retained a small portion of the original data. Fig. 5.7a and 5.7b represent cases in which the Health Patch was not used correctly by the participant, e.g. not changing the Patch in time or removing the Patch during night time. Fig. 5.7c represents a noisy signal, explained by sweat and activity. In Fig. 5.7d the signal itself is overall of good quality, though the script for selection of high confidence data has underperformed here.

Some input signals only retain a very small percentage of data after preprocessing. Fig. 5.7 illustrates these cases. A possible explanation for Fig. 5.7a is that sweat between the skin and the patch reduced the signal quality. A case in which the Patch should have been replaced. In Fig. 5.7b it is clear that the Health Patch has been removed during the night. Fig. 5.7c represents a noisy signal. Between 15h and 18h, the participant has been gardening which might explain the noise, because of sweat and activity. In Fig. 5.7d the signal itself is overall of good quality, though the script for selection of high confidence data has underperformed here.

The confident signals of features of the ambulant Health Patch data are depicted in Fig. 5.8. These figures show the range and evolution of different features over time. For the unsupervised learning of the ambulant data, normalised features are used, as depicted in Fig. 5.9. For the normalised features, it becomes visible that *mean HR* and *RMSSD* are negatively correlated signals. They both exhibit a clear pattern over time. For *LF* and *HF* most feature values are situated around zero. Features *SDNN* and *LF/HF* are mostly oscillating around zero.

5.4.2 Correlation of signals

The stress detection of the ambulant phase is focussed on ECG features. As these features are derived from the same signal, the correlation between features is investigated first. Fig. 5.10 displays correlation heat maps of two different test persons on arbitrarily chosen days. It is clear from both heat maps that a very large negative correlation exists between the features *mean heart rate* and *RMSSD*. For the depicted example cases, the correlation is -0.9959 for TP06 and -0.9921 for TP11. Both features are calculated similarly:

$$\begin{aligned}
\text{mean HR} &= \frac{60 \times f_{sampling}}{\text{average RR interval}} \quad \text{[bpm]}, \\
\text{RMSSD} &= \sqrt{\text{average}\left(\left(\frac{1000 \times \text{RR interval}}{f_{sampling}}\right)^2\right)} \quad \text{[ms]}.
\end{aligned} \tag{5.3}$$

5.4.3 Influence of day and night physiology

The Health Patch recorded the ECG for 24 hours continuously. This means that when training the Self-Organizing Map and subsequent clustering, it is influenced by data recorded both during night time as during daytime. In general, physiology is more at rest during night time and feature values are expected to be lower.

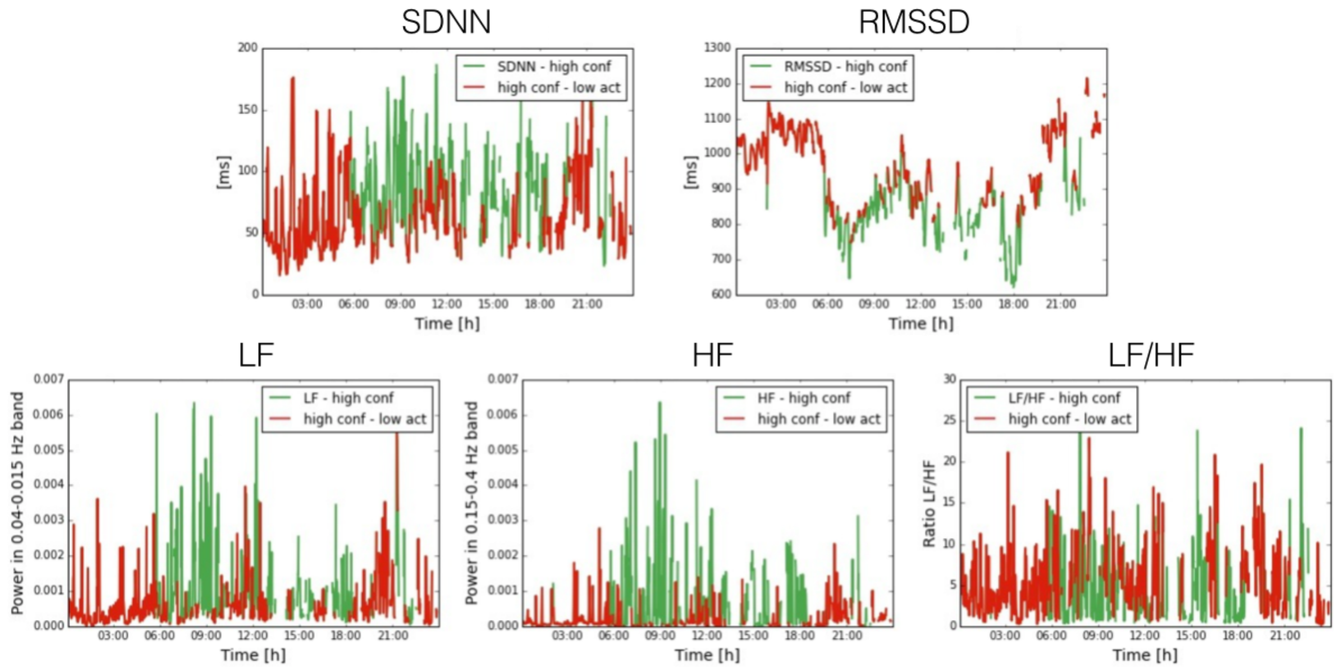


Figure 5.8: Graphical representation of ECG features of Health Patch data.Only the confident part of the data is retained (green) on which the low activity signal is indicated (red).

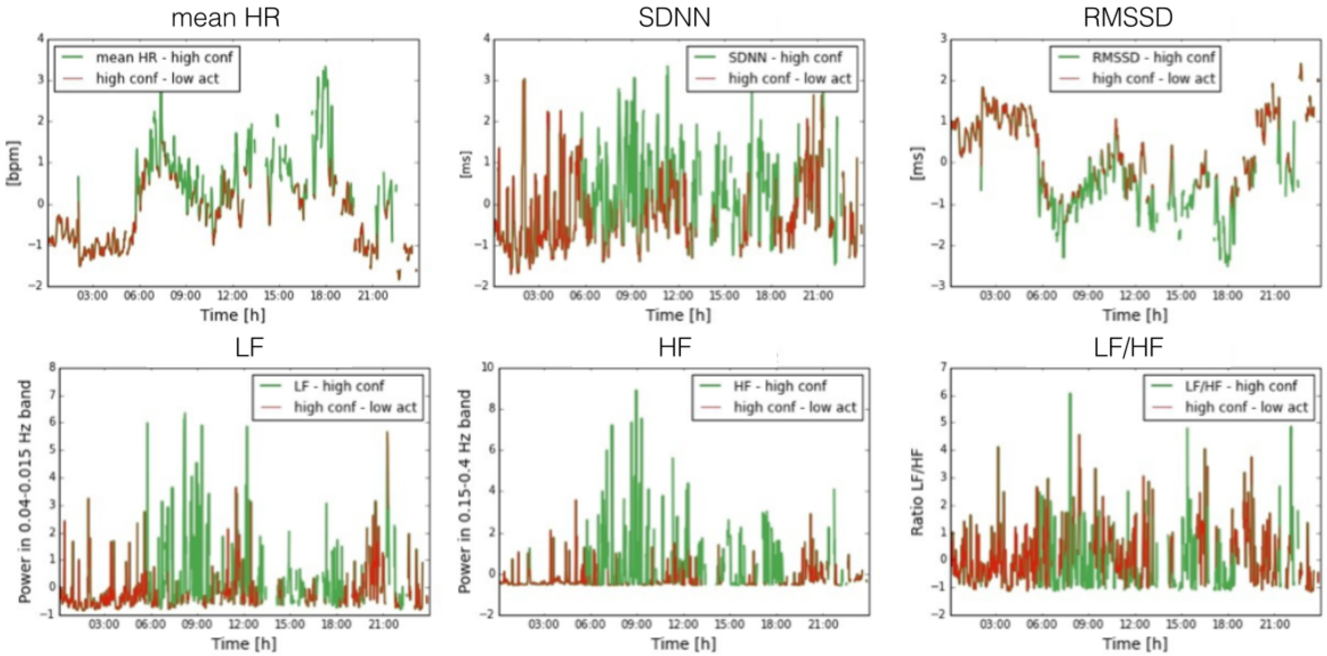


Figure 5.9: Graphical representation of normalised ECG features of Health Patch data. Only the confident part of the data is retained (green) on which the low activity signal is indicated (red). These normalised features, as used for the algorithmic pipeline.

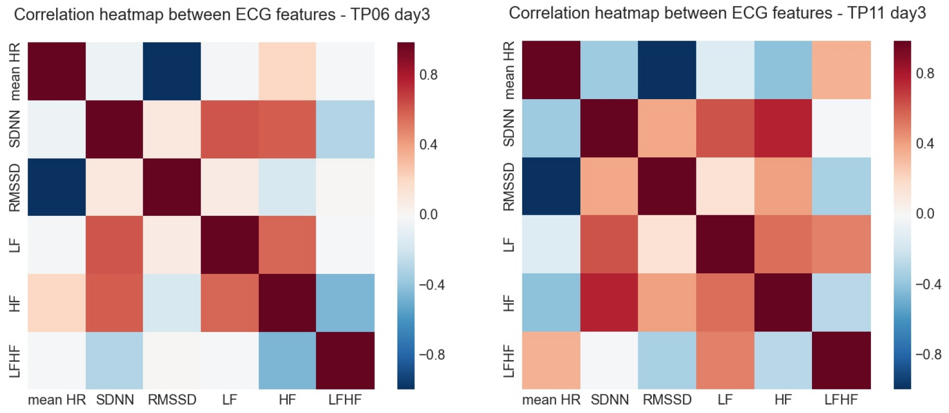


Figure 5.10: Correlation heat maps of ECG features on ambulant data. A large negative correlation can be seen between mean HR and RMSSD.

To check this assumption, every feature sample is labelled as day or night. *Day* is defined between 7am and 10pm and labelled as '1'. *Night* is defined between 23pm and 6pm and labelled as '0'. These labels are mapped onto the trained SOM. Every node of the SOM has a weight. Feature vectors are compared to these weights. The best matching node receives the label of the input feature vector.

The SOM is trained using six features, thus formed in a six-dimensional feature space. The labels and the SOM are visualised on intersections of the feature space. The node positions of the SOM along the feature space of *mean heart rate* are depicted in Fig. 5.11. Blue colours indicate nodes situated at low values of the feature, red values indicate nodes situated at high values of the feature. In the subsequent figures, the SOMs are overlayed once by *day* labels and once by *night* labels to visually check the correspondence with the colours of the SOM, i.e. node positions of the SOM.

As can be seen from Fig. 5.11 representing the *mean HR* space, *day* labels correspond very well with high *mean HR* and *night* labels with low *mean HR*. Searching for clusters representing *day* and *night* would be fairly straight forward when only using *mean HR* as a feature. This is however not the point of interest. The aim is to find two delineated clusters, representing *stress* and *relax* phases. Therefore, to exclude this effect of achieving a day/night clustering, samples labelled as *night* (i.e. between 11pm and 6am) will be removed.

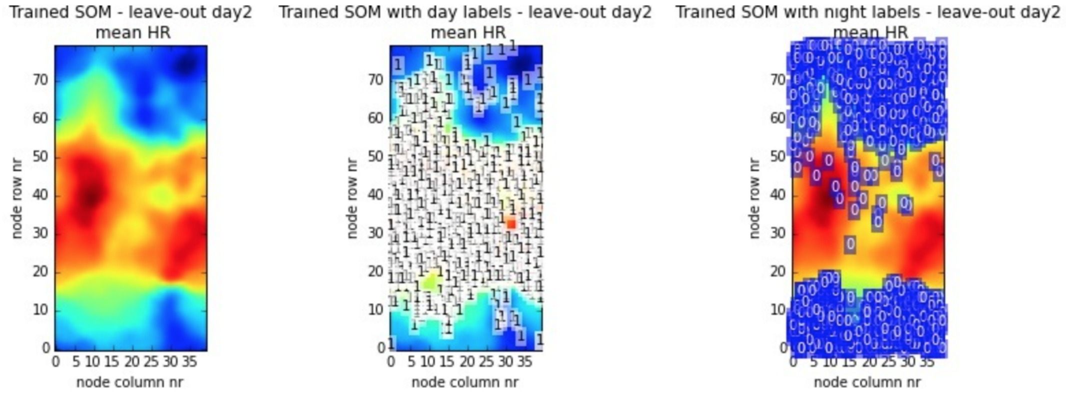


Figure 5.11: Correspondence of day and night physiology with trained SOM. Visually, it can be steen that a clear correspondence exists between high feature values of *mean HR* and mapping of day time samples, and low feature values with night time samples.

5.4.4 Cohesion and separation of clusters

After the Self-Organizing Map is trained and every node received a position in feature space, these nodes are clustered. This section investigates the quality of the clusters. As the clusters have no related labels, the quality is expressed in terms of cohesion and separation of the clusters. These factors are merged in the silhouette coefficient [46, 57].

The silhouette coefficient s of sample i is defined as:

$$s(i) = \frac{b(i) - a(i)}{max(a, b)}, \tag{5.4}$$

with a the mean intra-cluster distance and b the mean distance to all samples of the nearest cluster. As only two clusters are considered, b is simply the other cluster than to which i is assigned to.

The range of the silhouette coefficient is between -1 and 1. If $s(i)$ approaches 1, it implies that $a(i) \ll b(i)$ and the mean intra-cluster distance is much smaller than the mean distance to samples of the other cluster. Therefore, sample i is well-clustered and assigned to the right cluster. When $s(i)$ is about zero, the sample i lies equally far from both clusters and it is not clearly defined which is the right cluster. If $s(i)$ is close to -1, $a(i) \gg b(i)$, meaning that the sample is misclassified.

Fig. 5.12 displays on the left side the silhouette coefficients of samples in cluster 0 and cluster 1. The coefficients are plotted in ranked order. The red dashed line represent average silhouette score over all samples (here 0.28). The right side depicts the feature values of the test data, plotted in two-dimensional feature space of *mean HR* and *SDNN*, as a section of the original five-dimensional feature space. Colours represent their assigned cluster, i.e. blue for cluster 0 and red for cluster 1.

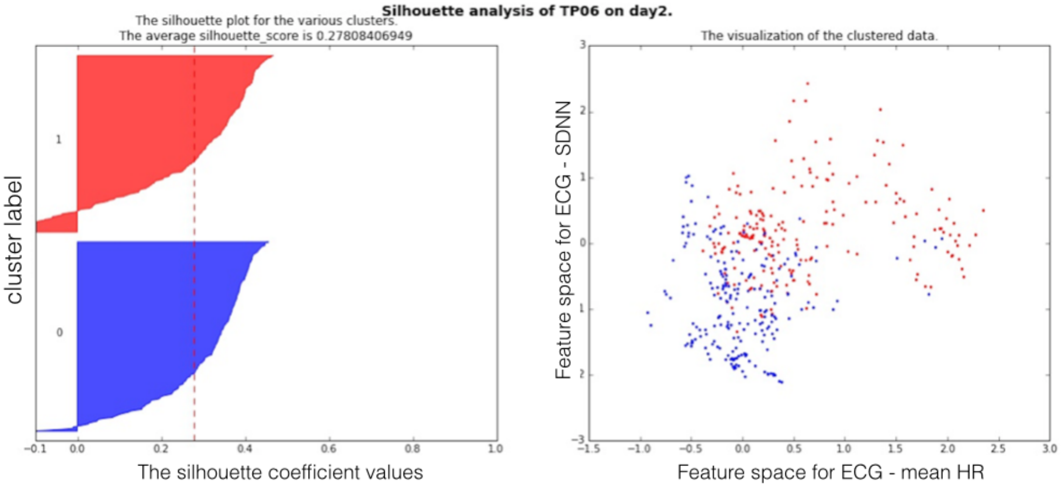


Figure 5.12: Left: Silhouette coefficients of samples in cluster 0 and cluster 1, in ranked order. Red dashed line represent average silhouette score over all samples (here 0.28). Right: Feature values of the test data plotted in two-dimensional feature space of *mean HR* and *SDNN*. Colours represent their assigned cluster, i.e. blue for cluster 0 and red for cluster 1.

Optimal feature subset

The silhouette coefficient is applied to derive an optimal subset of features from the total set of ECG features. The optimal subset achieves clustering with a maximal silhouette coefficient.

The total ECG set of features is *mean HR, RMSSD, SDNN, LF, HF* and *LFHF*. Every combination out of this set, including single features is tested. Single features are represented as a two-dimensional feature space of the single feature on both axes. The feature subset is applied to calculate features of the training data. As every participant has ambulant data over several days, a leave-one-day-out cross validation is performed. The feature vectors of the training data are applied to construct self-organizing maps. The nodes of the SOM received a location in feature space and are subsequently clustered. The quality of these clusters is evaluated in terms of cohesion and separation with the average silhouette coefficient. The average silhouette coefficients over all days is averaged out. This is the final average silhouette coefficient for a specific feature subset for a specific participant. The procedure is repeated for all participants. The average silhouette coefficients of all participants for a certain subset is averaged. Next, the most optimal feature set is selected to continue further calculations.

For all participants the same trends were visible, being single feature subset reaches the largest silhouette coefficients. The maximal silhouette coefficient was attained for a single feature subset *LFHF* with 0.62, with single feature subsets *LF* and *HF* being in the same range. Single feature subset *mean HR* has an average silhouette coefficient of 0.52. The combination of *mean HR* and *RMSSD* reached a value of 0.49. The silhouette coefficient of the complete feature set was low with a score of 0.16.

The silhouette coefficients plot for a single feature subset *LFHF* is shown in Fig. 5.13. A very high average value of 0.62 is reached (red dashed line). The plot on the right hand side with the two-dimensional feature space of *LFHF* reveals that feature *LFHF* mainly clusters values close to zero. Fig. 5.9 shows that *LFHF* is indeed oscillating around zero.

The silhouette coefficients for a feature subset *mean HR, RMSSD* is shown in 5.14. The two-dimensional feature space of *mean HR* and *RMSSD* depicts a clear negative correlation between both features. The negative correlation might help clustering in feature space as the nodes of the SOM are attracted to two distant parts of the feature space.

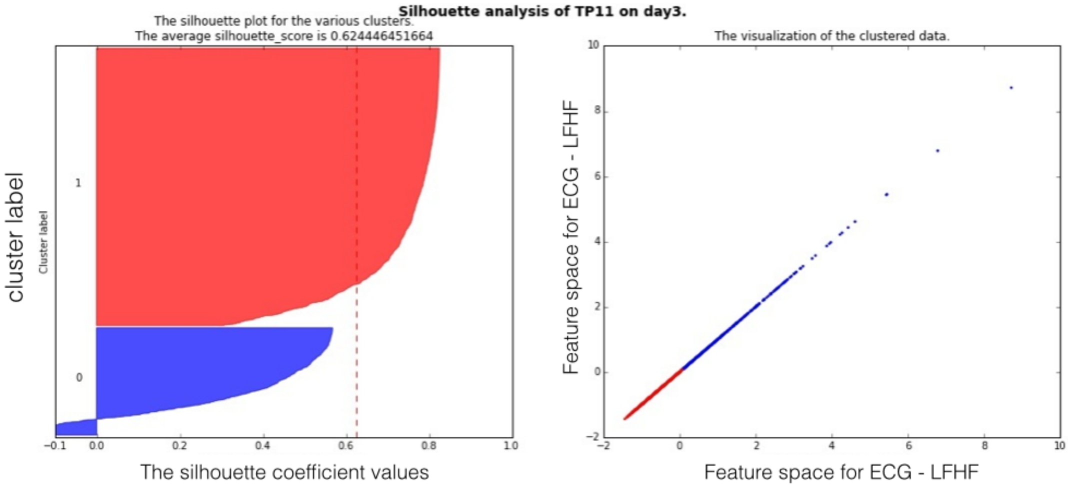


Figure 5.13: The symmetric two-dimensional feature space of *LFHF* reveals that feature *LFHF* mainly clusters values around zero.

Comparison to laboratory data

To evaluate the performance of clustering, it is compared to the average silhouette coefficient derived from the laboratory data. As the laboratory data consists of seven test persons for which approximately 15 minutes of data exists, a leave-one-participant-out cross validation is applied. The optimal subset of features of previous section is derived from the training data. The silhouette coefficient is calculated on the clustering of the training data. Silhouette coefficients are averaged after cross validation, which reaches a final averaged value of 0.63 for *LFHF*, 0.51 for *mean HR* and 0.56 for *mean HR, RMSSD*. The silhouette coefficients of the lab data are almost equal to these from the ambulant data for *LFHF* and *mean HR*, though slightly higher for *mean HR, RMSSD*.

Additionally, the optimal feature subset is calculated on the wearable data measured in laboratory conditions. The lab data contains labels, as such a testing performance

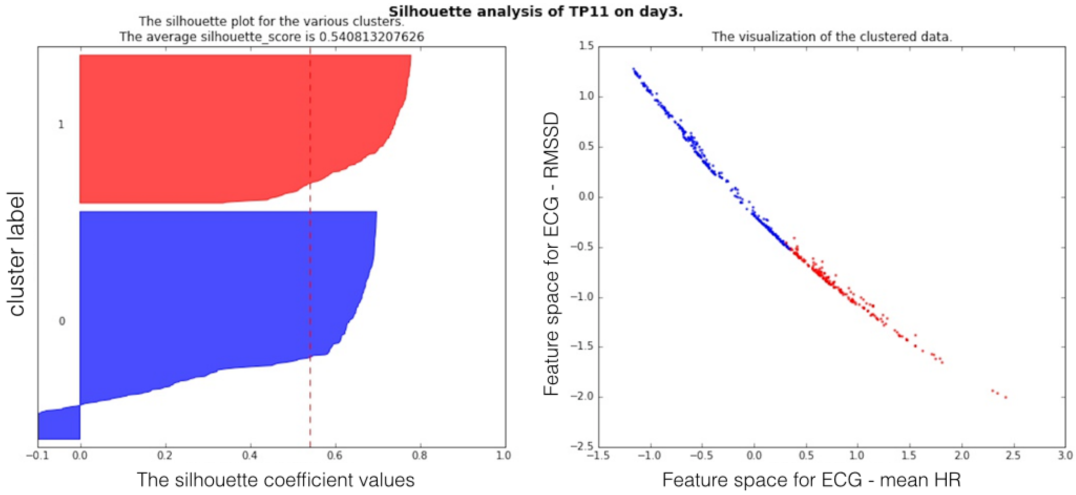


Figure 5.14: The two-dimensional feature space of *mean HR* and *RMSSD* depicts a clear negative correlation between both features. This might help clustering in feature space.

can be calculated. A complete LOO validation is run over every combination of the feature set. The most performant feature sets is *mean HR* in combination with *RMSSD*. Repeating the LOO validation ten times, results in an average testing performance of 76.42% (stand. dev. of 10.10%). This is 6% more compared to the average testing performance of *mean HR* (70.20%). The testing performance of *LFHF* is only 56.86% (stand. dev. of 10.90%). Using the whole feature set the testing performance is even less with 52.50% (stand. dev. of 13.50%)

Silhouette coefficients are similar over ambulant and laboratory data of the Health Patch, when using a particular feature set. However, testing performances are dependent on the applied features subset. An optimal feature set based on the silhouette coefficient does not result in the same optimal feature based on testing performance of laboratory data. As laboratory data is considered as the gold standard, further calculation are made with the optimal feature set based on testing performance, i.e. *mean HR, RMSSD*.

A typical trained SOM with corresponding clustering is shown in Fig. 5.15. The left figures represent the trained SOM in *mean HR* and *RMSSD* feature space. The right figure represents the clustering on this SOM.

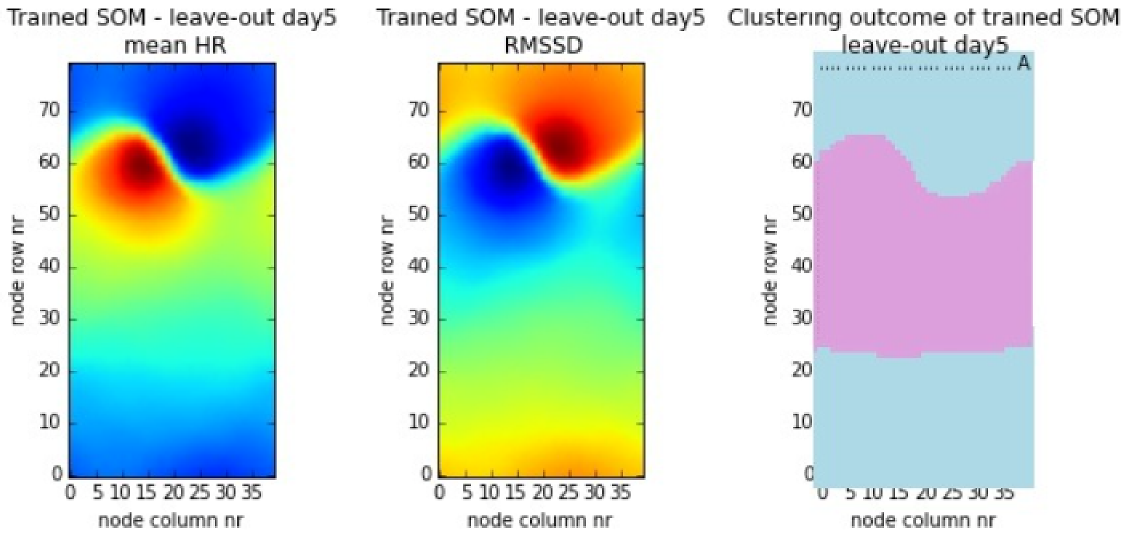


Figure 5.15: A trained SOM is depicted with an intersection of *mean HR* and *RMSSD* feature space. The right figure represent the clustering on this SOM.

Optimal grid configuration for ambulant data

Using the optimal subset *mean HR, RMSSD* the optimal grid configuration was found to be 40 × 20 (sec. 5.3.3). As the ambulant data contains more samples, the relative dimensions of the grid are kept, though double in size 80 × 40. With enlarging the dimensions, the resolution of the Self-Organizing Map is enlarged. This enables more precise predictions.

5.4.5 Similarity of clusters

Previous section demonstrates that the algorithmic pipeline using feature set *mean HR, RMSSD* outlines two clusters with large intra-cluster cohesion and inter-cluster separation. This section demonstrates that these clusters are not two randomly outlined clusters from the data. Instead, it is shown that among clusters from different days and different participants, similar patterns are repeated. Therefore, the feature values in clusters are compared. It is observed that clusters contain similar feature values. The feature values within a cluster are represented as boxplots (Fig 5.16). The boxplots depict cluster A (blue) and cluster B (purple) of features *mean HR* (left) and *RMSSD* (right) of different participants. The median is printed in bold font above the boxplot. Visually, it can be seen that the same pattern of boxplots repeats over different clustered self-organizing maps.

If clusters are not presenting the same pattern, the clustering had an undesirable outcome. This will be clearly seen in corresponding boxplots and silhouette plot (Fig. 5.17). The resulting silhouette coefficient is lower as the average calculated from ambulant and lab data (0.52 and 0.51 respectively, see previous section). Therefore a minimum threshold is set at an average silhouette coefficient of 0.45. Clusterings with a lower coefficient will be excluded for further analysis, which was one out of 40 days. An example of an incorrect clustering can be observed visually from the corresponding boxplots (Fig. 5.17a). The pattern does not match to other boxplots. Comparison by median of the boxplots would indicate a correct clustering, though the interquartile distance shows irregularity. The silhouette plot (Fig. 5.17b) clearly indicates a false clustering. The average silhouette coefficient is about 0.36 and below the threshold of 0.5. The clustering outcome is one cluster with all the extreme values of *mean HR* (and *RMSSD*) and one cluster with all the intermediate values, as illustrated in Fig. 5.18. As a conclusion, the similarity of clusters has to be determined by comparison of boxplot median and value of average silhouette coefficient.

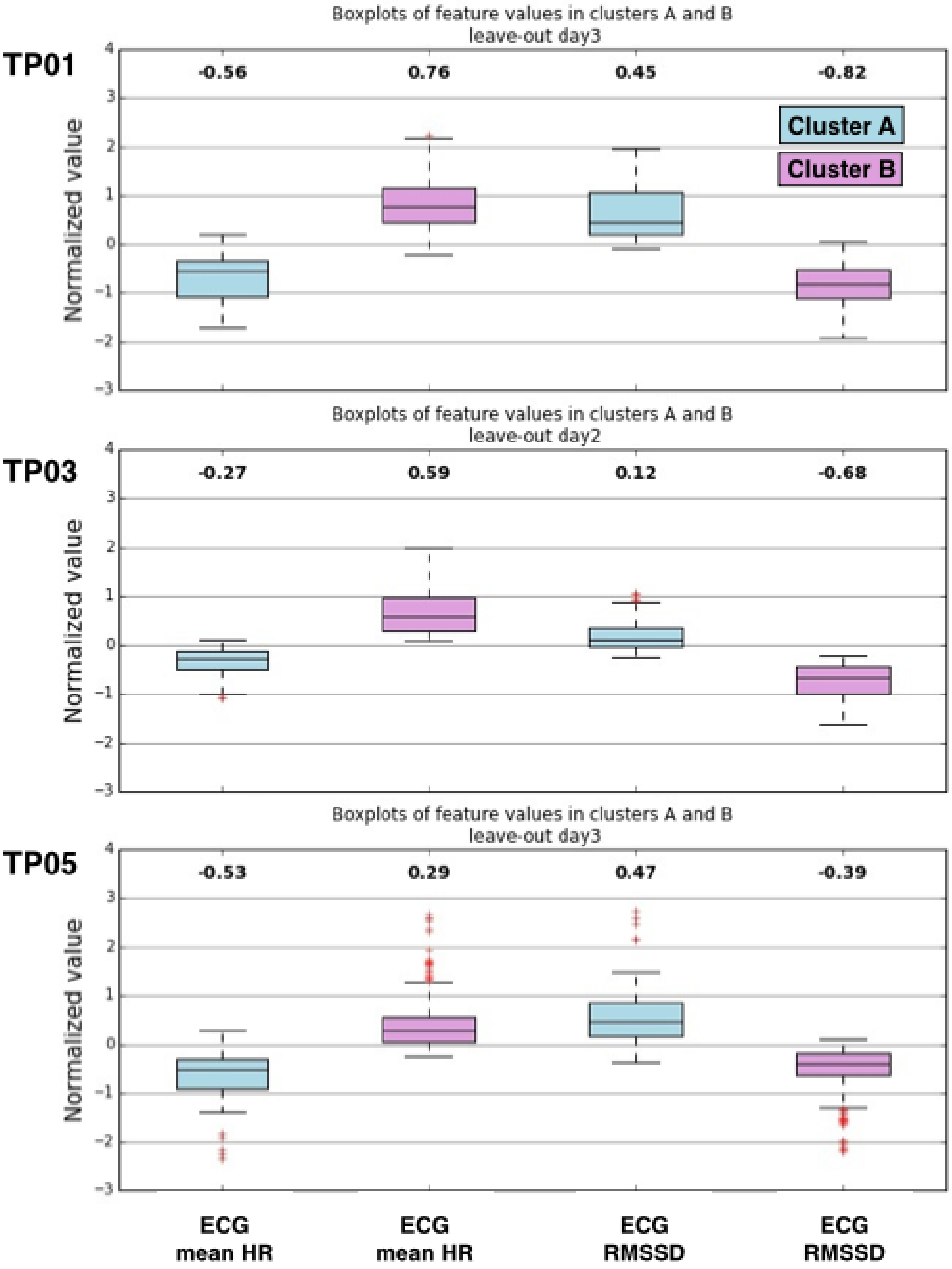
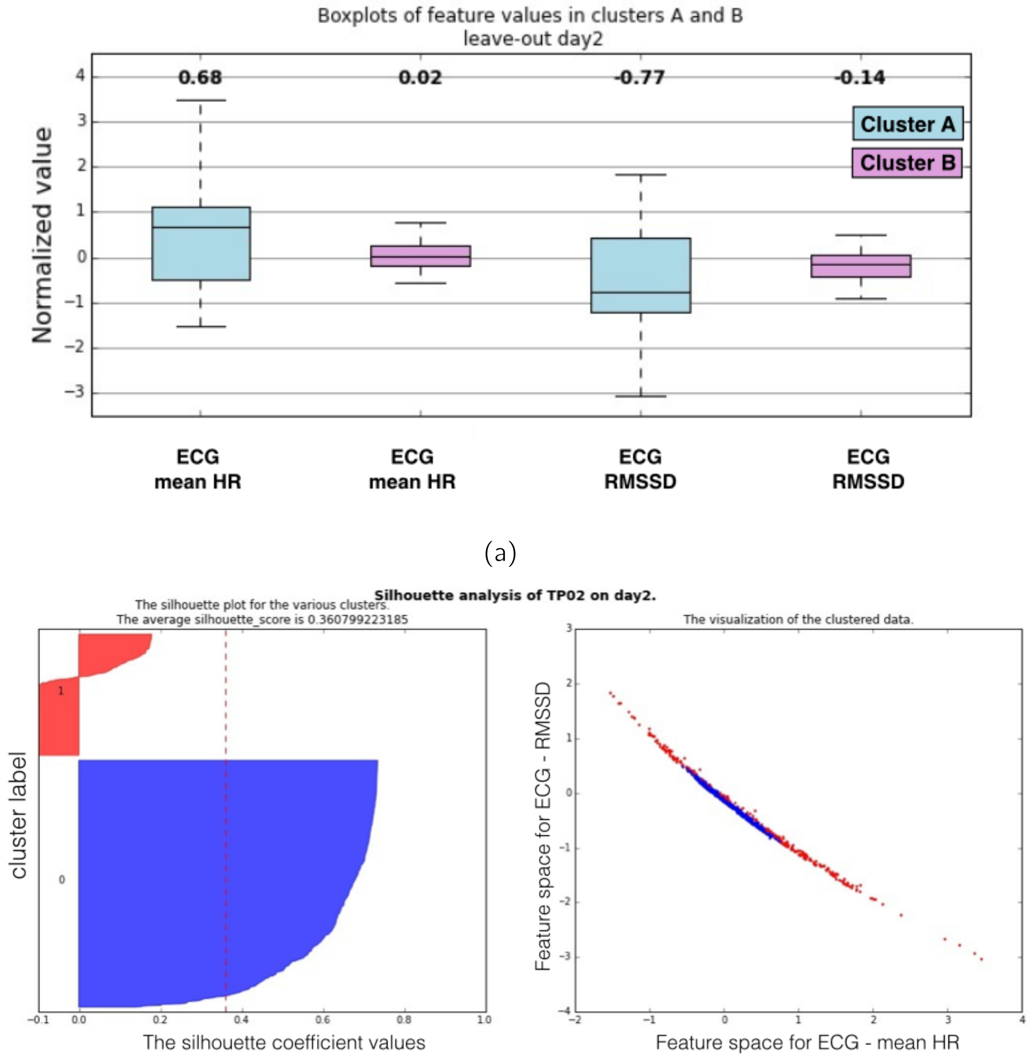


Figure 5.16: Boxplots representing the features values contained in cluster A (blue) and cluster B (purple) of feature *mean HR* (left) and *RMSSD* (right) of different participants. The median is printed in bold font above the boxplot.



(a)

(b)

Figure 5.17: Undesirable clustering can be derived visually from the corresponding boxplots (Fig. 5.17a). Their pattern will not match to other boxplots. The interquartile distance shows irregularity. The silhouette plot (Fig. 5.17b) as well clearly indicates an undesirable clustering. The average silhouette coefficient is about 0.36 and below the threshold of 0.5. Extreme values are clustered together.

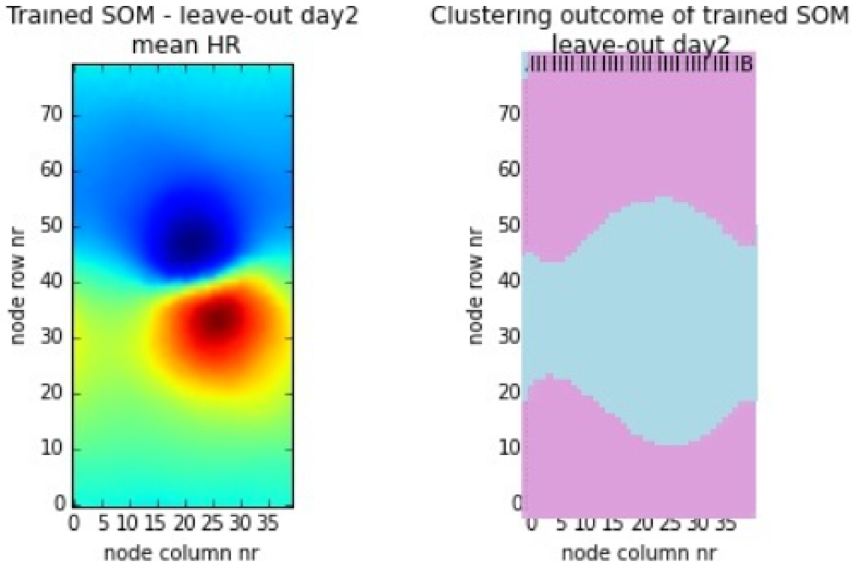


Figure 5.18: Cluster (blue) with extreme low and high values of *mean HR* and *RMSSD* and a cluster (purple) with the intermediate values.

5.4.6 Defining stress and relax clusters

This section evaluates if the found clusters are not only similar over different datasets, but also represent a *stress* and a *relax* cluster. To determine which clusters represents *stress* or *relax*, the boxplots of feature values are examined. During the supervised learning phase, it was observed that a high *mean HR* indicates a period of stress (Fig. 4.14). This was confirmed during the cluster identification of unsupervised learning in lab phase (sec. 5.2) and by Taelman et al. [13] and by Vrijkotte et al. [33]. Visually, there is a clear indication that this alternation of high and low feature values is present in the found clusters of Fig 5.16.

5.4.7 Detection of stress and relax intervals

The SOMs are trained and every node is assigned to a *relax* or *stress* cluster, the algorithmic pipeline can now predict stress level of new, unseen samples. Samples are mapped onto the SOM by comparison of its feature vector and the weights of the SOM nodes. The sample receives the label of the best-matching node, *stress* (label **1**) or *relax* (label **0**).

Every minute a sample or feature vector is calculated, with a window size of five minutes. Stress predictions are averaged over one hour intervals to obtain a stress level instead of a series of a series of **0** and **1**. As many data points have been excluded because of low confidence and high activity, some time intervals might contain a low number of samples and bias the averaging procedure. Therefore a threshold is set to maintain a level of confidence about the prediction. A minimum of 15 samples, thus 15 minutes, per hourly interval is required in order to make a prediction on this interval. The threshold is calculated based on TP who retained around 50% of their data. A lower threshold would not provide enough confidence. A higher threshold would exclude to many predictions. The retained data after preprocessing (sec. 5.4.1) of three days of different TPs was below 20%, thus no confident prediction could be made as not enough samples were available. The data of this day was however included for training of the SOM.

5.4.8 Validation

Previous sections demonstrated the robustness and quality of the unsupervised algorithmic pipeline based on SOMs and clustering. According to literature (sec. 2.6), the last phase of validation is comparison against the participants' questionnaire. The scores of stress level by the TPs in the questionnaire are normalised for every TP as TPs rate minimum and maximal levels of stress differently. The RMSE between predictions and TP stress scores is calculated. The outcomes per TP are summarised in table 5.2. The averaged RMSE over all TP is 0.4953 with a standard deviation of 0.0915. As Self-Organizing Maps have a certain amount of randomness in the initialization of their weights, results might differ slightly over different runs. Two examples of a prediction with the questionnaire stress level is depicted in Fig. 5.19. In the left figure the same trend is followed, though in the right figure this cannot be seen.

Table 5.2: Root Mean Squared Error of prediction and questionnaire for all test persons.

	TP01	**TP02**	**TP03**	**TP04**	**TP05**
RMSE	0.5748	0.4624	0.4046	0.5384	0.6291
	TP06	**TP07**	**TP11**	**TP12**	**TP13**
RMSE	0.4151	0.3643	0.4294	0.5390	0.5960

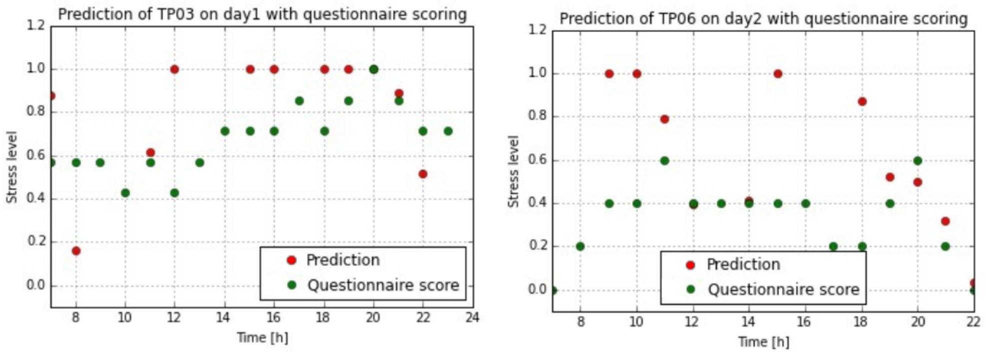


Figure 5.19: Stress level predictions (red) with questionnaire stress level (green).

5.5 Conclusion

First, the feasibility of the *Self-Organizing Maps* was tested on the laboratory data and compared a posteriori with the objective labels. Using a subset of ECG features, the classification performance was 76.42%. This is a comparable result to supervised learning with ECG features, indicating the principal functioning of the SOMs for stress detection.

In a second phase, the SOM was applied on the ambulant data. Training the SOM with ECG features *mean HR* and *RMSSD* from the ambulant data, enabled clustering from the feature space. The clusters were well separated with large cohesion, with an average silhouette coefficient of 0.49. Moreover, the clusters were similar over different test persons and days. According to literature the center values of the features in each cluster can indicate stress and relax phases. By mapping test samples on the trained and clustered SOM, stress predictions were made. Comparison against the subjective stress levels from the questionnaire was however poor with an RMSE of 0.50.

It is suggested to further explore the use of *Self-Organizing Maps* as it solely relies on the physiological data, excluding subjective labelling. Important improvements can be made by applying multimodal feature sets, including for example galvanic skin response.

6 Discussion

6.1 Features and classification performance in supervised versus unsupervised learning

6.1.1 Laboratory data

During supervised learning with Random Decision Forests on the lab data (sec. 4.6.3), the model using GSR features on the Nexus data reaches an average classification performance of 82.66%. The combined model based on both GSR features (Nexus) and ECG features (Health patch) reaches an even higher classification performance of 83.04%. Smets et al. [45] had a similar experimental set-up and reported a maximum performance rate (for non-personalized models) of 82.7% using SVM. Similar features for ECG and GSR were applied, with additional Temperature and Respiration features. Therefore, it can be concluded that supervised classification of lab data was successful and applicable as a gold standard for the subsequent unsupervised phases.

It was found that GSR *SCph* and ECG *mean heart rate* were the most important features in this model. Therefore these features were also applied for unsupervised learning with Self-Organizing Maps of the lab data. Here, a classification performance of 77.63% was reached (sec. 5.3.2). This can be considered as a very good result, comparing with supervised learning, as unsupervised learning is a more challenging task. Bornoiu and Grigore [42] applied a Self-Organizing Map as well for stress detection, using similar GSR features and reported an average recognition rate of 86.25%. They implemented a non-wearable GSR in laboratory setting with stress-inducing experiments as well. Different was the their labelling system for validation of their outcomes. An expert observer evaluated the GSR signal in combination with participant questionnaires to manually label the input signal. Obviously, recognition rates will be higher as labelling of the data is based on a priori evaluation of the physiological signals. Furthermore, it is not clear how their *average recognition rate* is computed. A classification rate based on the average of sensitivity and specificity will generally be lower as purely reporting the sensitivity 5.3.2. Moreover, their number of participants is not reported for comparison.

6.1.2 Ambulant data

Optimal ECG feature subset

Ambulant data was measured by the Empatica and by the Health Patch. As the data of Empatica was found to be of low quality during lab measurements, it was decided to solely investigate Health Patch signals during the ambulant phase. Therefore examination of ECG features is important as these are the basis for unsupervised learning of ambulant data (overview of ECG features table 4.5). Both the analysis with supervised as unsupervised learning reveal that *mean HR* and *RMSSD* are the most important feature of the ECG feature set. A similar study by Hovsepian et al. [4] confirms *mean HR* being an informative feature.

Supervised learning based on all the ECG features of the Health patch lab data reaches an average classification performance of 76.89%. When applying all ECG features during unsupervised learning, a very poor classification result is obtained of 52.04%.

The most optimal feature subset among all ECG features for unsupervised learning is the feature subset *mean HR, RMSSD*, which reaches a classification result of 76.42%, a result similar to supervised learning with all ECG features. It is clear that including other features actually hinder the classification with SOM. Therefore, it is important to carefully select the features in case of rraining a SOM in a two-dimensional feature space, as every feature has a great influence on the training procedure.

The feature subset *mean HR, RMSSD* was also the most important ECG feature subset during RDF training (Fig. 4.13). A Random Decision Forest is able to select informative features, though *mean HR* and *RMSSD* are highly negatively correlated (sec. 4.6.2 and 5.4.2). This explains why both *mean HR* and *RMSSD* were almost equally important during RDF training, as both features basically contain the same, though a high information gain.

Silhouette score

The reason for this discrepancy between classification performances of the complete feature set and the subset with *mean HR* can be found when observing the silhouette scores. The silhouette scores for single features *LF, HF* and *LFHF* are highest with approx. 60% while the silhouette score for *mean HR* is lower, about 50%. These scores were similar over lab and ambulant data. Observing both lab signals in Fig. 4.12 as ambulant signals Fig. 5.9, it can be seen that the normalised features *SDNN, LF, HF, LF/HF* have many values at zero or oscillate around zero. Therefore the Self-Organizing Map is attracted in feature space to these values around zero. The subsequent clustering will cluster the dense values around zero and cluster all remaining values. These clusters do not necessarily indicate stress and relax clusters, compared to the observation that *mean HR* follows a more varying pattern over time and achieves acceptable classification rates.

For the data being observed in this thesis, a larger silhouette score does not imply higher classification outcomes. The reason is that certain features (*SDNN, LF, HF, LF/HF*) lead to clusters with a good cohesion, although not separating *stress* and *relax* data . Obviously, there is a large grey zone between *stress* and *relax* which makes this task challenging.

6.2 Challenges of wearables for ambulant data collection

6.2.1 Challenge of multi-modal signals

Analysis of the laboratory was based on features from two different sensors measuring different signals. i.e. skin conductance with Nexus and ECG with the Health Patch. Both supervised as unsupervised learning methods achieved good classification outcomes. During supervised learning it was demonstrated that the features set combining GSR and ECG achieved the highest classification rates, while the feature set based exclusively on GSR performed almost equally well. As GSR enables the use of very performant features (Fig. 4.7, 4.15b), it is recommended to include this signal for stress detection. Unfortunately, the GSR signal of the wearable wristband Empatica was generally of inferior quality during lab experiments.

The ECG feature set was evaluated in more detail as only the ECG signal of the ambulant data could be analysed. Both with the supervised as the unsupervised method *mean HR* and *RMSSD* were the two most performing features. Other features had little (RDF, supervised) or negative (SOM, unsupervised) impact on the classification outcome. Furthermore, *mean HR* and *RMSSD* are highly negatively correlated. To avoid the correlation of features, the use of multi-modal signals, i.e. signals from different sources, is encouraged. Furthermore, the expansion of the ECG feature set could be beneficial. The use of time-series features was not incorporated as only chunks of high-confident and low-activity data could be retained (table 5.1). Time-series features require continuous data streams. Sarker et al. [38] added the missing data after exclusion of high-activity data in order to compute time-series based features. These features include information from prior time periods, such as *duration of previous stress episode* or *slope of best-fit line to past stress likelihood values*.

6.2.2 Challenge of qualitative data

The basis for a good evaluation outcome is a qualitative input. The Random Decision Forest and Self-Organizing Map achieved good results on data from the laboratory phase, being the GSR signal measured by Nexus and the ECG signal measured by the Health Patch. The wearable Empatica measuring GSR could however not be used for analysis as the signal was noisy or data was missing. The quality of the Health Patch and Empatica signals was verified against the fixed sensor Nexus, which served as the ground truth. The incorporation of wearable GSR data could have boosted classification outcomes as more data and data from a different source would be available.

During ambulant phase, large portions of data were excluded(table 5.1) because the signals did not match typical ECG patterns (low confidence) and the standard deviation of the acceleration magnitude exceeded the threshold (high-activity). Especially the requirement for low-activity reduced the amount of usable data.

For many time intervals, no prediction could be made as not enough samples were available. Ideally, this high-activity data is included and compensated for in the data set. How to account for high-activity data during stress detection is an ongoing topic of research [28, 17] (see sec. 2.4.2). Additionally, the script for detection of high-confidence data needs further improvement.

6.3 Outcome clustering

The proof of concept of clustering nodes of a SOM was carried out first on lab data with the use of the two most performant features during supervised training, i.e. *GSR-SCph* and *ECG-meanHR*. Only two features were chosen for better visualisation and lower complexity. The outcome of clustering has proven to be sufficient on lab data with an average clustering performance of 74.91% and a standard deviation of 6.36% (sec. 5.3.1) and a testing performance of 77.63% and a standard deviation of 7.82% (sec. 5.3.2). Calculation of the testing performance is completely based on the clustering, as such it is a good measure for clustering as well.

A large constraint for clustering on ambulant data was the fact that only ECG features could be applied, as GSR features had proven to be more performant in general. The clustering performance of the ambulant data is based on the average silhouette coefficient and the similarity of boxplots. It was observed that single features had the largest silhouette coefficients, though not all of these features were ideal for *stress - relax* detection. Features *SDNN, LF, HF* and *LF/HF* did not exhibit large varying patterns and mainly clustered values around zero. Only two correlated features remained, i.e. *mean HR* and *RMSSD*, constructing a two-dimensional feature space. Therefore the full potential of the algorithmic pipeline of training a Self-Organizing Map and clustering was not exploited. The SOM has the ability of compressing the information of a high-dimensional space and thus simplifying the subsequent clustering step. Clustering could be performed in a higher-dimensional feature space. With the use of features highly related to stress, better separated clusters would emerge, increasing the confidence of predictions.

Applying *mean HR* and *RMSSD*, 39 out of 40 days were clustered with a silhouette coefficient above the established thresholds and boxplots exhibiting repeating patterns. Although some of the clustering procedures failed with a typical clustering pattern as displayed in Fig. 5.17b. Here the clustering outcome was one cluster with all the extreme values of *mean HR* (and *RMSSD*) and one cluster with all the intermediate values, as illustrated in Fig. 5.18. Ruling out this cluster with intermediate values could increase the confidence of stress predictions with the cost of excluding data. Clustering with variational Bayesian Gaussian Mixture has the possibility for automatic selection of the best number of cluster, though this property hasn't been exploited. It is suggested to further explore this method with the potential of finding multiple clusters of which one cluster with these intermediate values.

6.4 Validation on ambulant data

The validation procedure of ambulant data consisted of multiple steps. First, it was examined if well separated clusters with large cohesion were obtained. Subset of feature were found which provided good average silhouette scores (0.50 and more). Next, the similarity of clusters over different test persons and days was observed to verify that the algorithm did not randomly outlined 2 clusters. Third, it was proven that these clusters represented *stress* and *relax* states. The final step was to compare the predictions made to the participants subjective feelings of stress. The stress levels documented in the diary were normalised for every participant. The RMSE between prediction and subjective stress level had an average value of 0.50 with a standard deviation of 0.09. Sometimes the stress detections follow certain trends of the questionnaire or have an elevated stress level when the participant is under a lot of stress. However no confident predictions can be made.

The first three phases proved a working concept. The final validation step consisted of comparison against subjective stress levels, although the purpose of unsupervised learning is to exclude these subjective labels. Stress levels derived from questionnaires are highly subjective, not accurate and often filled in with a delay [20, 4]. Sarker et al. [38] reviewed the consistency of self-reported responses. Often responses were inconsistent and participants were biased towards neutral self-assessment. Therefore, they supported the value of an objective sensor-based model of stress. It is a conflict in deciding what defines a correct stress criterion:

the unsupervised clustering based on physiological data or subjective stress levels declared by participants. The result of this thesis is a correct clustering of *stress* and *relax* phases based on consistent feature patterns indicating stress [13, 33].

6.5 Future work

Improvements can be made by applying multimodal feature sets. Signals from different sensors sources are intrinsically uncorrelated and therefore automatically add new information. Galvanic Skin Response features are recommended to further explore as they had a large and important contribution in RDF training. In SOM training as well, the phasic skin conductance (GSR - SCph) lead to good performance. Furthermore, non-linear features should be taken into account, including time-series features as in [35, 38]. These incorporate information from previous time intervals, which is interesting as *stress* is not an isolated event.

The SOM has the ability of compressing the information of a high-dimensional space and thus simplifying the subsequent clustering step. Clustering could be performed in a higher-dimensional feature space. Increasing the number of features would also diminish the influence of features that do not greatly contribute to stress detection, such as *SDNN* mainly clustering around zero. Generally, other unsupervised learning techniques are suggested to be explored as well, which are for example more robust for bad feature selection. Exploratory work of unsupervised learning methods for identification of stress states has been done by Medina [44], though only on laboratory data and using ECG signals. Donner et al. [58] describe different types of networks for analysis of recurrence-based time series, though outside the field of stress detection.

Advances can be made in more detailed evaluation of the clustering step. Other parameters or other cluster algorithms can be explored to better separate *stress* from *relax* clusters. An interesting aspect would be to add more clusters to capture stress levels directly from the SOM or determine the confidence of a predicted stress level. A first step would be to add a third cluster outlining intermediate values and focus on the extreme values of *stress* and *relax*.

A general improvement for stress detection is the incorporation of highly-active data. As activity influences physiology, many data points got excluded because high activity was detected in these intervals. Perhaps mapping of high activity data onto a trained SOM could reveal useful patterns about changing physiology. Furthermore, retaining as much data as possible leads to more continuous data streams and enables the use of time-series features [38].

It is suggested to validate the developed model of unsupervised learning with *Self-Organizing Maps* against the field-data and field self-reports of the *cStress* model [4]. This model aims to provide a gold standard for continuous stress assessment of ambulant data. The field self-report is an Ecological Momentary Assessment (EMA) in which participants are prompted 15 times a day for instantaneous self-report of *stress*. This is an advantage over the questionnaire used in this thesis as immediate reporting is required. Furthermore, they expand these self-reports to a self-reported stress at every minute, based on the reported stress of the previous minute, as well as the physiological response of the previous minute. As such, the model allows for arbitrary lags between physiological response and the lingering memory of a past stress event captured in self-report.

Future work on a long-term basis is to provide feedback to the users of wearables on their stress levels. Proper information about his or her stress level could unveil stressful habits or situations, allowing proper counteracting [4].

Bibliography

[1] "Calculating the costs of work-related stress and psychosocial risks – a literature review," 2014. [Online; Accessed 27.01.2017].

[2] S. C. Segerstrom and G. E. Miller, "Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry," *Psychological bulletin*, vol. 130, pp. 601–630, 07 2004.

[3] J. Siegrist, "Effort-reward imbalance at work and cardiovascular diseases.," *International Journal of Occupational Medicine and Environmental Health*, vol. 23, no. 3, pp. 279–285, 2010.

[4] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cstress: Towards a gold standard for continuous stress assessment in the mobile environment," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, (New York, NY, USA), pp. 493–504, ACM, 2015.

[5] M. Myrtek and G. Brügner, "Perception of emotions in everyday life - studies with patients and normals," *Biological Psychology*, vol. 42, no. 1–2, pp. 147 – 164, 1996.

[6] M. Myrtek, E. Aschenbrenner, and G. Brügner, "Emotions in everyday life: an ambulatory monitoring study with female students.," *Biological Psychology*, vol. 68, no. 3, pp. 237–255, 2005.

[7] R. D. Lane, D. M. Quinlan, G. E. Schwartz, and S. B. Zeitlin, "The levels of emotional awareness scale: a cognitive-developmental measure of emotion.," *Journal of Personality Assessment*, vol. 55, no. 1-2, pp. 124–134, 1990.

[8] B. Verkuil, J. F. Brosschot, M. S. Tollenaar, R. D. Lane, and J. F. Thayer, "Prolonged non-metabolic heart rate variability reduction as a physiological

marker of psychological stress in daily life," *Annals of Behavioral Medicine*, vol. 50, no. 5, pp. 704–714, 2016.

[9] R. Karasek. and T. Theorell, *Healthy Work: Stress, Productivity, and the Reconstruction of Working Life*. Basic Books, 1992.

[10] E. Pereira, F. Kothe, F. Bleyer, and C. Teixeira, "Work-related stress and musculoskeletal complaints of orchestra musicians," *Revista Dor*, vol. 15, pp. 112 – 116, 06 2014.

[11] J. Siegrist, *Effort-reward imbalance at work: theory, measurement and evidence*, 2012.

[12] L. K. McCorry, "Physiology of the autonomic nervous system," *American Journal of Pharmaceutical Education*, vol. 71, 2007.

[13] J. Taelman, S. Vandeput, A. Spaepen, and S. V. Huffel, *Influence of Mental Stress on Heart Rate and Heart Rate Variability*, pp. 1366–1369. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[14] T. Hoehn, S. Braune, G. Scheibe, and M. Albus, "Physiological, biochemical and subjective parameters in anxiety patients with panic disorder during stress exposure as compared with healthy controls," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 247, no. 5, pp. 264–274, 1997.

[15] S. Freeman, *Biological Science*. Pearson Prentice Hall, 2005.

[16] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens, and J. Penders, "Towards mental stress detection using wearable physiological sensors," in *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2011*, pp. 1798–1801, IEEE Engineering in Medicine & Biology Society, August 2011.

[17] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "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*, UbiComp '16, (New York, NY, USA), pp. 1185–1193, ACM, 2016.

[18] D. M. Almeida, "Resilience and vulnerability to daily stressors assessed via diary methods," *Current Directions in Psychological Science*, vol. 14, no. 2, pp. 64–68, 2005.

[19] R. Larson and M. Csikszentmihalyi, "The experience sampling method," *Naturalistic Approaches to Studying Social Interaction*, vol. 15, pp. 41–56, 1983.

[20] P. Adams, M. Rabbi, T. Rahman, M. Matthews, A. Voida, G. Gay, T. Choudhury, and S. Voida, "Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild," in *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '14, pp. 72–79, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

[21] S. S. Dickerson and M. E. Kemeny, "Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research," *Psychological Bulletin*, vol. 130, no. 3, p. 355–391, 2005.

[22] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, pp. 156 – 166, 2005.

[23] J. Wijsman, B. Grundlehner, H. Liu, J. Penders, and H. Hermens, "Wearable physiological sensors reflect mental stress state in office-like situations," in *Humaine Association Conference on Affective Computing and Intelligent Interaction. ACII 2013*, (Los Alamitos, CA, USA), pp. 600–605, IEEE Computer Society, September 2013.

[24] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 2, pp. 279–286, 2012.

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

[26] J. Zhai and A. Barreto, "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, 2006.

[27] M. Kusserow, O. Amft, and G. Tröster, "Psychophysiological body activation characteristics in daily routines," in *ISWC 2009: Proceedings of the 13th International Symposium on Wearable Computers*, pp. 155–156, IEEE, 2009.

[28] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, M. Griss, and G. Yang, *Activity-Aware Mental Stress Detection Using Physiological Sensors*, pp. 211–230. Springer Berlin Heidelberg, 2012.

[29] J. C. Sriram, M. Shin, T. Choudhury, and D. Kotz, "Activity-aware ecg-based patient authentication for remote health monitoring," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, (New York, NY, USA), pp. 297–304, ACM, 2009.

[30] W. Boucsein, *Electrodermal Activity*. Springer, 2012.

[31] C. Kappeler-Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410–417, 2010.

[32] D. Carroll, J. Rick Turner, and J. C. Hellawell, "Heart rate and oxygen consumption during active psychological challenge: The effects of level of difficulty," *Psychophysiology*, vol. 23, no. 2, pp. 174–181, 1986.

[33] T. Vrijkotte, L. V. Doornen, and E. D. Geus, "Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability," *Hypertension*, p. 886, 2000.

[34] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, "The effect of mental stress on heart rate variability and blood pressure during computer work," *European Journal of Applied Physiology*, vol. 92, no. 1, pp. 84–89, 2004.

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

[36] M. T. F. of the European Society of Cardiology the North American Society of Pacing Electrophysiology Marek Malik, "Heart rate variability standards of measurement, physiological interpretation, and clinical use," *Circulation*, no. 5, pp. 1043–1065, 1996.

[37] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cstress: Towards a gold standard for continuous stress assessment in the mobile environment," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, (New York, NY, USA), pp. 493–504, ACM, 2015.

[38] H. Sarker, M. Tyburski, M. M. Rahman, K. Hovsepian, M. Sharmin, D. H. Epstein, K. L. Preston, C. D. Furr-Holden, A. Milam, I. Nahum-Shani, M. al'Absi, and S. Kumar, "Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 4489–4501, 2016.

[39] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, ACII '13, (Washington, DC, USA), pp. 671–676, IEEE Computer Society, 2013.

[40] N. Sharma and T. Gedeon, "Modeling a stress signal," *Applied Soft Computing*, vol. 14, Part A, pp. 53 – 61, 2014. Special issue on hybrid intelligent methods for health technologies.

[41] J. Ramos, J.-H. Hong, and A. K. Dey, "Stress recognition - a step outside the lab.," in *PhyCS* (A. Holzinger, S. H. Fairclough, D. Majoe, and H. P. da Silva, eds.), pp. 107–118, SciTePress, 2014.

[42] O. Grigore and I.-V. Bornoiu, "Kohonen neural network stress detection using only electrodermal activity features," *Advances in Electrical and Computer Engineering*, vol. 14, no. 3, pp. 71–78, 2014.

[43] J.-S. Han and G.-J. Kim, "A method of unsupervised machine learning based on self-organizing map for bci," *Cluster Computing*, vol. 19, no. 2, pp. 979–985, 2016.

[44] L. Medina, "Identification of stress states from ecg signals using unsupervised learning methods," in *Portuguese Conf. on Pattern Recognition - RecPad*, vol. -, pp. —, October 2009.

[45] E. Smets, P. Casale, U. Großekathöfer, B. Lamichhane, W. De Raedt, K. Bogaerts, I. Van Diest, and C. Van Hoof, *Comparison of Machine Learning Techniques for Psychophysiological Stress Detection*, pp. 13–22. Springer International Publishing, 2016.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[47] M. Kusserow, O. Amft, and G. Tröster, "Monitoring stress arousal in the wild," *IEEE Pervasive Computing Magazine*, vol. 12, pp. 28–37, Apr. 2013.

[48] "imec." `http://http://www.imec-int.com/`. [Online; Accessed 07.02.2017].

[49] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[50] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2-3, pp. 81–227, 2011. DOI: 10.1561/0600000035.

[51] E. Peper, R. Harvey, I.-M. Lin, H. Tylova, and D. Moss, "Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony?," *Biofeedback*, vol. 35, no. 2, pp. 55–61, 2007.

[52] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Ambulatory stress monitoring with minimally-invasive wearable sensors," *Orient.J. Comp. Sci. and Technol*, 2010.

[53] S. Nikam, "A comparative study of classification techniques in data mining algorithms.," *Orient.J. Comp. Sci. and Technol*, vol. 8, no. 1, pp. 13–19, 2015.

[54] T. Kohonen, M. R. Schroeder, and T. S. Huang, eds., *Self-Organizing Maps*. Springer-Verlag New York, Inc., 3rd ed., 2001.

[55] B. Arnrich, C. Kappeler-Setz, R. La Marca, G. Tröster, and U. Ehlert, "Self organizing maps for affective state detection," in *Machine Learning for Assistive Technologies*, 2010.

[56] scikit-learn developers, "Variational Bayesian Gaussian Mixture." `http://scikit-learn.org/stable/modules/mixture.html#bgmm/`. [Online; Accessed 10.12.2016].

[57] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[58] R. Donner, M. Small, J. Donges, N. Marwan, Y. Zou, R. Xiang, and J. Kurths, "Recurrence-based time series analysis by means of complex network methods," *International Journal of Bifurcation and Chaos*, vol. 21, no. 04, pp. 1019–1046, 2011.