Institutionen för systemteknik
Department of Electrical Engineering

Examensarbete

Wearable Sensor Data Fusion for Human Stress Estimation

Examensarbete utfört i elektroteknik
vid Tekniska högskolan vid Linköpings universitet
av

Simon Ollander

LiTH-ISY-EX–15/4904–SE

Linköping 2015

Linköpings universitet
TEKNISKA HÖGSKOLAN

Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

Linköpings tekniska högskola
Linköpings universitet
581 83 Linköping

Wearable Sensor Data Fusion for Human Stress Estimation

Examensarbete utfört i elektroteknik
vid Tekniska högskolan vid Linköpings universitet
av

Simon Ollander

LiTH-ISY-EX–15/4904–SE

Handledare: **Martin Lindfors**
ISY, Linköpings universitet
Christelle Godin
Commissariat à l'Énergie Atomique et aux Énergies Alternatives
Aurélie Campagne
Université Pierre-Mendès-France

Examinator: **Gustaf Hendeby**
ISY, Linköpings universitet

Linköping, 28 oktober 2015

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering
SE-581 83 Linköping

Datum
Date

2015-10-28

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English
☐ _____

Rapporttyp
Report category

☐ Licentiatavhandling
☒ Examensarbete
☐ C-uppsats
☐ D-uppsats
☐ Övrig rapport
☐ _____

ISBN
—

ISRN
LiTH-ISY-EX–15/4904–SE

Serietitel och serienummer
Title of series, numbering

ISSN
—

URL för elektronisk version
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-XXXXX

Titel
Title

Fusion av data från bärbara sensorer för estimering av mänsklig stress
Wearable Sensor Data Fusion for Human Stress Estimation

Författare
Author

Simon Ollander

Sammanfattning
Abstract

I syfte att klassificera och modellera stress har olika sensorer, signalegenskaper, maskininlärningsmetoder och stressexperiment jämförts. Två databaser har studerats: MIT:s förarstressdatabas och en ny databas baserad på egna experiment, där stressuppgifter har genomförts av nio försökspersoner: Trier Social Stress Test, Socially Evaluated Cold Pressor Test och d2-testet, av vilka det sistnämnda inte normalt används för att generera stress. Support vector machine-, naive Bayes-, k-nearest neighbour- och probabilistic neural network-algoritmer har jämförts, av vilka support vector machine har uppnått den högsta prestandan i allmänhet ($99.5 \pm 0.6\%$ på förardatabasen, $91.4 \pm 2.4\%$ på experimenten). För båda databaserna har signalegenskaper såsom medelvärdet av hjärtrytmen och hudens ledningsförmåga, tillsammans med medelvärdet av beloppet av hudens ledningsförmågas derivata identifierats som relevanta. En ny signalegenskap har också introducerats, med hög prestanda i stressklassificering på förarstressdatabasen. En kontinuerlig modell har också utvecklats, baserad på den upplevda stressnivån angiven av försökspersonerna under experimenten, där support vector regression har uppnått bättre resultat än linjär regression och variational Bayesian regression.

Nyckelord
Keywords

stress, data fusion, classification, modelling, wearables, physiological sensors

Abstract

With the purpose of classifying and modelling stress, different sensors, signal features, machine learning methods, and stress experiments have been compared. Two databases have been studied: the MIT driver stress database and a new experimental database, where three stress tasks have been performed for 9 subjects: the Trier Social Stress Test, the Socially Evaluated Cold Pressor Test and the d2 test, of which the latter is not classically used for generating stress. Support vector machine, naive Bayes, k-nearest neighbor and probabilistic neural network classification techniques were compared, with support vector machines achieving the highest performance in general ($99.5 \pm 0.6\%$ on the driver database and $91.4 \pm 2.4\%$ on the experimental database). For both databases, relevant features include the mean of the heart rate and the mean of the galvanic skin response, together with the mean of the absolute derivative of the galvanic skin response signal. A new feature is also introduced with great performance in stress classification for the driver database. Continuous models for estimating stress levels have also been developed, based upon the perceived stress levels given by the subjects during the experiments, where support vector regression is more accurate than linear and variational Bayesian regression.

Sammanfattning

I syfte att klassificera och modellera stress har olika sensorer, signalegenskaper, maskininlärningsmetoder och stressexperiment jämförts. Två databaser har studerats: MIT:s förarstressdatabas och en ny databas baserad på egna experiment, där stressuppgifter har genomförts av nio försökspersoner: Trier Social Stress Test, Socially Evaluated Cold Pressor Test och d2-testet, av vilka det sistnämnda inte normalt används för att generera stress. Support vector machine-, naive Bayes-, k-nearest neighbour- och probabilistic neural network-algoritmer har jämförts, av vilka support vector machine har uppnått den högsta prestandan i allmänhet ($99.5 \pm 0.6\%$ på förardatabasen, $91.4 \pm 2.4\%$ på experimenten). För båda databaserna har signalegenskaper såsom medelvärdet av hjärtrytmen och hudens ledningsförmåga, tillsammans med medelvärdet av beloppet av hudens ledningsförmågas derivata identifierats som relevanta. En ny signalegenskap har också introducerats, med hög prestanda i stressklassificering på förarstressdatabasen. En kontinuerlig modell har också utvecklats, baserad på den upplevda stressnivån angiven av försökspersonerna under experimenten, där support vector regression har uppnått bättre resultat än linjär regression och variational Bayesian regression.

Acknowledgments

Many thanks to Christelle Godin at the Atomic Energy and Alternative Energies Commission for being a great supervisor and for sharing her expertise. I have very much appreciated her help, knowledge and experience.

I would also like to thank Aurélie Campagne at Pierre-Mendès-France University for a great experimental campaign and for all the help and advice on psychological and physiological measurements.

Furthermore I express my gratitude towards Gustaf Hendeby and Martin Lindfors at the Department of Electrical Engineering, Linköping University, for guiding me through this work.

Finally I thank Audrey Vidal and Sylvie Charbonnier for the hints and discussions.

Grenoble, August 2015
Simon Ollander

Contents

Notation

Abbreviations

Abbreviation	Meaning
ACC	Accelerometer
BVP	Blood volume pulse
CPT	Cold pressor test
CWT	(Stroop) Color word task
ECG	Electrocardiogram
EEG	Electroencephalography
EDA	Electrodermal activity
EMG	Electromyogram
GSR	Galvanic skin response
HF	High frequency
HPA	Hypothalamic-pituitary-adrenal
HR	Heart rate
HRV	Heart rate variability
IBI	Inter-beat interval
KNN	k-Nearest neighbour
LDA	Linear discriminant analysis
LF	Low frequency
MF	Medium frequency
MST	Mental arithmetic stress test
NB	Naive Bayes classifier
PCA	Principal component analysis
PNN	Probabilistic neural network
PPG	Photoplethysmogram
PSS	Perceived stress scale
RBF	Radial basis function
RMS	Root mean square
RSA	Respiratory sinus arrhythmia
RESP	Respiration

Abbreviations

Abbreviation	Meaning
SCL	Skin conductance level
SCR	Skin conductance response
SD	Standard deviation
SECPT	Socially evaluated cold pressor test
SMOTE	Synthetic minority over-sampling technique
ST	Skin temperature
SVM	Support vector machine
SVR	Support vector regression
TSST	Trier social stress test
VAS	Visual analogue scale
VBML	Variational Bayesian multiple linear regression
VLF	Very low frequency

1 Introduction

This chapter gives an introduction to stress detection and the background of this Master's thesis.

1.1 Motivation

In the daily life, stress is a normal reaction of the human body to external events of different kinds. However, if this reaction is too great or if it lasts too long, there is a risk of it resulting in physical or mental disorders. To prevent this, estimation of the stress level of a person using wearable sensors could give an early warning if the person is experiencing too high or too long-lasting stress. Recent works [43], [35], [20] have studied and found connections between non-invasive physiological measures and stress levels induced in laboratory conditions.

To contribute to the European Union project Bonvoyage 2020 [31], the physiological reactions of drivers in the context of stress has been studied, using the MIT Driver Stress Database. Furthermore, an experimental database based upon laboratory stress tasks has been recorded and analyzed for comparing different types of stress. All this data has been analyzed and modelled using signal processing and machine learning methods.

1.2 Purpose

The objective of this Master's thesis is to interpret physiological signals and to study the relation between the sensor measurements and stress levels induced by different tasks and conditions. The purpose is to explain and compare different

stress tasks, sensors, signal features, and modelling methods to give an understanding of their importance and applicability in the domain of stress detection.

1.3 Research Questions

The research questions that are to be answered by this work can be summarized by:

1. Which sensors are most relevant for detecting stress?
2. Which signal properties are most relevant for detecting stress?
3. Which signal properties and features are common for different types of stress?
4. Which machine learning techniques are most relevant for modelling stress?

1.4 Limitations

A limitation of this study is that only machine learning and black box methods are used for modelling.

1.5 The Company, CEA

The CEA (French Atomic and Alternative Energy Commission) is a public organization performing scientific research in the areas of energy, defense and security, and information and health technologies. It exists at 10 sites in France and consists of several divisions and laboratories, including the LETI (Laboratory for Electronics and Information Technology). The LETI mainly focuses on microelectronics and nanotechnologies, and is currently employing around 1,800 people. To a great extent, it focuses on helping companies increasing their competitiveness by innovation and by transferring its technical knowledge to the industry. Overall, research contracts with the industry are worth 75 % of CEA-LETI annual income. In particular, products integrating these technologies are industrialized at the DSIS (Department of Systems and Solution Integration), which is why a considerable part of its financing origins from the industry. DSIS is the department where this Master's thesis internship has been carried out, at the Laboratory for Multimodal Systems and Sensors (LSCM).

1.6 Thesis Outline

This Master's thesis is structured as explained below.

Chapter 2 explains the existing research regarding stress experiments and modelling, along with the necessary theory for understanding this work.

Chapter 3 presents the methods and results on the MIT driver database.

Chapter 4 introduces a new experimental database, along with its results.

Chapter 5 presents the conclusions that are made in this work.

2

State of the Art

The purpose of this chapter is to explain previous studies and results regarding estimation of human stress using sensors. It will discuss and compare different choices of sensors, signals, features and classifier methods along with their performance in existing research. This chapter will also introduce important terminology and notations that will be used throughout this Master's thesis.

2.1 Physiology and Stress

To deal with situations that humans normally do not have the resources to deal with, we have developed a biological and psychological reaction called stress. It can increase the performance of a human being for a short period of time, but longer exposure can lead to health problems.

A stressor is a stimulus that causes stress reactions. Examples include an exam, the death of a family member, moving house, loss of one's job, or a threat [30]. Stressful situations normally contain at least one of the following elements [12]:

- reduced or no control of the situation
- unpredictability, something unexpected is happening, or it is hard to predict what will happen
- novelty, something new that the person has never experienced is happening
- threat of ego, one's skill is put to test and one has doubt about one's capacities
- a threat in general

- time pressure

Physiologically, the body must first decide whether or not the situation is stressful. This is based upon sensory input in combination with stored memories. If the situation is indeed judged as stressful, the hypothalamus, located in the base of the brain, is activated [42, p. 34-48], [34] to start a stress reaction. The two main physiological components connected to the stress reaction are the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system. The parasympathetic nervous system is also involved, and together these two form the autonomic nervous system. Simplified, one can say that the sympathetic nervous system is responsible for "fight or flight" responses, while the parasympathetic nervous system deals with "rest and digest" mechanisms. [25, p. 411]. The short term effects are produced by the fight or flight response and consist of helping the body to deal with the stressor, e.g. giving the body more energy. The long term effects can however be negative, if the organism does not have enough time to recover from the stress.

2.2 Methods for Generating Stress

First of all, in order to model stress, a common way is to acquire data by generating stress in a person while recording physiological signals. Subsequently this data can be analyzed to find links and relations between these signals and the stress perceived by the person.

To generate stress, two main methods are used in the literature: stationary laboratory experiments and real-life data collections where the stress is more closely connected to daily life situations. [15] is an example of the latter, where participants are monitored during 5 weeks to compare the effects of different stress treatments.

One can distinguish between psychological stressors (e.g. mentally challenging tasks under time pressure or social stressors where other people are judging you) and physical stressors. These can also be combined in different ways to generate polyvalent and possibly higher stress levels.

[45, p. 227] compares the stress generated by several laboratory stress protocols. 20 healthy young men were subject to four of the most common tests in this category; the Trier Social Stress Test (TSST), Section A.1, a bicycle ergometer test, the Stroop Color Word Task (CWT), Section A.5, and the Cold-pressor Test (CPT), Section A.2. All four protocols increased the perceived stress levels of the participants, with TSST causing the highest level, followed by the ergometer, the CWT and lastly the CPT. The HPA axis response was the highest from the TSST, then the ergometer, the CWT and finally the CPT. These methods are further explained in Appendix A.

Due to the subjectiveness of stress there is no standardized method of evaluating the level of stress perceived by a participant. There are several questionnaires who try to solve this issue, e.g. the visual analogue scale (VAS) [45], [36], [32].

It simply lets the user place his or her perceived stress level on a line with two end points, one corresponding to no stress and the other to extreme stress. Other examples include Likert scales with various number of points and items. The "Perceived Stress Scale" (PSS) is has also been widely used since its publication in 1983, e.g. in [56].

Table 2.1 presents various studies and experiments where participants have been stressed in different ways. It varies from classical laboratory stressors (such as the TSST) to real-life situations (such as the driver task presented in [20]). If used, the questionnaire evaluating the perceived stress level of the participants is also presented along with the scale.

A series of factors can influence the results of these kinds of stress experiments, e.g. the time of day (which affects the cortisol level) and gender. Furthermore, the bare knowledge and anticipation of being stressed might increase the perceived stress of the subject, but it might as well have a soothing effect (since unpredictability is connected to stress). In the literature, the subject is sometimes informed about the purpose of the task, while in other studies the subject is not informed or even misinformed about why they are participating. An example is to incorrectly tell the subject that the MST is an easy test of intelligence and that most participants do not have any difficulties with it, as in [59].

2.3 Physiological Signals

There are several methods of measuring properties of the human body. Figure 2.1 shows an overview of signals and features commonly used in affective signal processing (which also includes the analysis of other feelings than just stress). Different statistical methods are used to compute the features from the raw signals, which are further detailed in Section 2.5.

[60] tests four signals for continuously measuring physiological signals non-invasively: skin resistance, heart activity, the pupil diameter and the skin temperature. [35] records electroencephalography (electrical activity of the brain, EEG) and facial (corrugator and zygomatic) electromyography. [50, p. 97] combines speech signal and electrocardiogram to efficiently estimate the emotional states of persons. [52, p. 256] evaluates accelerometer, arterial blood pressure measurement, capnogram, electrocardiogram, electrodermal activity, impedance cardiography and temperature measurements. Thus there is a large choice of possible signals that one can acquire from the human body, measuring properties of the eye, the face, the brain, the muscles, the skin, the heart and even the movement of the body as a whole.

2.3.1 Electrocardiogram

An electrocardiogram (ECG) records the electrical activity of the heart using electrodes placed upon the body. It can be measured using electrodes, Figure 2.2,

source	setting	stressor	# subjects	questionnaire	stress scale
[22]	real-life	calls	9	Likert	7-point
[38]	real-life	daily life	18	PSS	14-item
[20]	real-life	driving	24	free, forced	1-5, 1-7
[2]	real-life	meetings	5	feeling	
[35]	lab	blood sample, CWT, MST, TSST	12		2 point
[45]	lab	CPT, CWT, physical, TSST	20	VAS	VAS, 100 mm
[37]	lab	CPT, MST	54		
[44]	lab	CPT, MST, TSST	22	EMA	0-1
[48]	lab	CWT	9	POMS, ZBW	
[60]	lab	CWT	32		4 point
[14]	lab	CWT, physical	15	self-assessed	0-5
[53]	lab	d2	456		
[25]	lab	MIST	33		0-1
[18]	lab	MST + noise, CPT	81		
[59]	lab	MST, social	44	Brief COPE	1-4
[10]	lab	MIST	42		
[29]	lab	MST	10		
[56]	lab	MST	30	PSS	
[32]	lab	SECPT	61	VAS	
[41]	lab	SECPT	70	11 point	0-100
[26]	lab	SECPT	72	[41]	0-100
[51]	lab	social	60	STAI-T	0-10
[36]	lab	TSST	26	VAS	
[9]	lab	TSST	39		1-10
[39]	lab	TSST	136, 44, 41	Dislocations Scale	1-7, 0-6, 0-15
[27]	lab	TSST	155		
[58]	lab	verbal	80	self-report, Likert	15-item, 5-point
[23]	lab	video	50		3 point

Table 2.1: *Comparison of stress generating methods and experiments, along with their methods of stress evaluation. The laboratory stressors are further detailed in Appendix A. For further details regarding the experiments and their stress scales and questionnaires, see the cited source.*

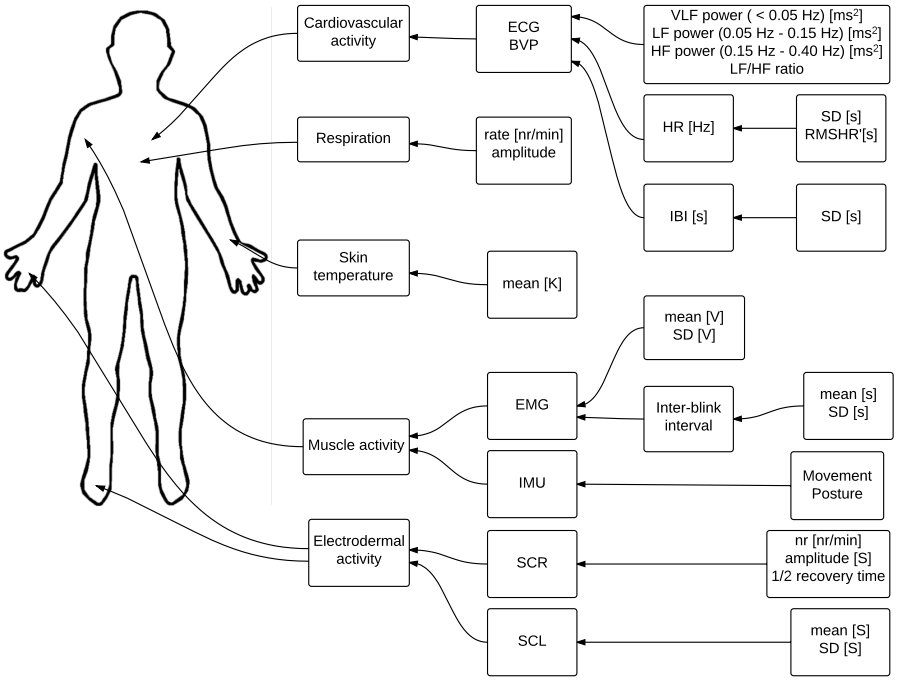

Figure 2.1: *Common physiological signals and features that might be used for stress detection, and how they are extracted [50, p. 6]. ECG = electrocardiogram, BVP = blood volume pulse, EMG = electromyogram, IMU = inertial measurement unit, SCR = skin conductance response, SCL = skin conductance level, VLF = very low frequency, LF = low frequency, HF = high frequency, HR = heart rate, IBI = inter-beat interval, SD = standard deviation, RMS HR' = root mean square of successive differences in heart rate.*

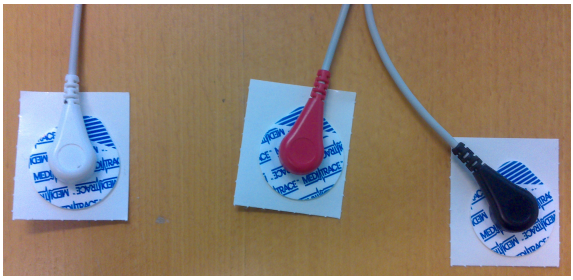

Figure 2.2: *ECG electrodes.*

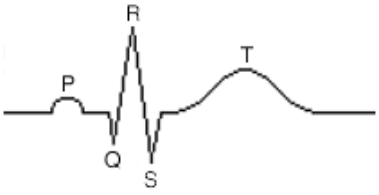

Figure 2.3: *A typical ECG signal representing a heartbeat, with the usual elements: P wave, QRS complex and T wave [54].*

which is further detailed in Section 4.1.2. The ECG signal is usually periodic, consisting of three parts: the P wave, the QRS complex and the T wave. These are given a graphical representation in Figure 2.3. The ECG signal is affected by the breathing cycle through a phenomenon called respiratory sinus arrhythmia (RSA). Expiration slows down the heart rate while the opposite is true for inspiration [33]. A main interest of the ECG is to calculate the heart rate (HR), normally done through the inter-beat intervals (IBI) of the R waves. The heart rate variability (HRV) is a denotation that combines all measures related to how the heart rate varies, e.g. its standard deviation or the difference between successive HR values. An alternative to ECG is measuring the blood volume pulse (BVP), from which the HR also can be derived. This method is called photoplethysmogram (PPG), and measures the differences in light caused by the blood volume pulsations.

2.3.2 Electromyogram

An electromyogram (EMG) records the electrical potential generated by skeletal muscle cells. Needle electrodes are used in this purpose, usually placed on an arm, a leg or a shoulder. Facial electromyography is also possible, in this case the electrodes are placed upon various facial muscles.

[57, p. 43] describes an experiment where the several features of the EMG signal were shown to change significantly between stressful and not stressful conditions, including amplitude, gaps, increased significantly during mental stress tasks. It concludes that EMG is a useful parameter to detect stress and that EMG sensors, together with other physiological sensors, can be included in a wireless system

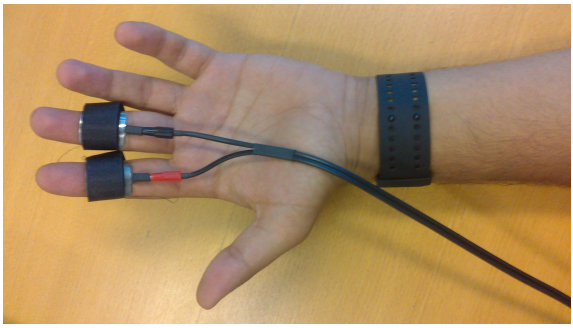

Figure 2.4: *GSR electrodes and the Empatica E4 wristband.*

for ambulatory monitoring of stress levels. However, since normal muscle movements have a large impact on the EMG, one must be careful to distinguish the source affecting an EMG signal.

2.3.3 Electrodermal Activity

The sweat glands and the skin blood vessel are only connected to the sympathetic nervous system, not the parasympathetic one. The heart rate, on the other hand, is influenced both by the sympathetic and the parasympathetic nervous systems. Sweat secretion increases the conductance of the skin proportionally, thus the electrodermal activity (EDA) is measured by the conductivity of the skin. The density of sweat glands is highest around the palms of the hands or the feet, so this is usually where it is measured. Another common name for EDA is the galvanic skin response (GSR). Two systems for measuring the GSR are presented in Figure 2.4: finger electrodes and the Empatica E4 wristband, with wrist electrodes. These are further detailed in Section 4.1.2. The skin conductance level (SCL) is the part of the EDA signal that changes slowly. It can indicate psychophysiological activation, but is subject to great individual variation.

The skin conductance response (SCR) is a peak in the EDA signal caused by a single stimulus, normally delayed by around 1.5 – 6.5 seconds (the latency). Common features of the SCR are its amplitude, the latency and its recovery time, Figure 2.5. The recovery time is the time required for the skin to regain its original conductance level.

Spontaneous fluctuations (non-specific SCR) can also occur, and their frequency and mean are of interest for psycho-physiological measures. They also vary between individuals, and their shapes are similar to the one of a specific SCR [25, p. 411].

[2] specifically uses solely a GSR sensor to detect stress by analyzing different peaks in the GSR signal.

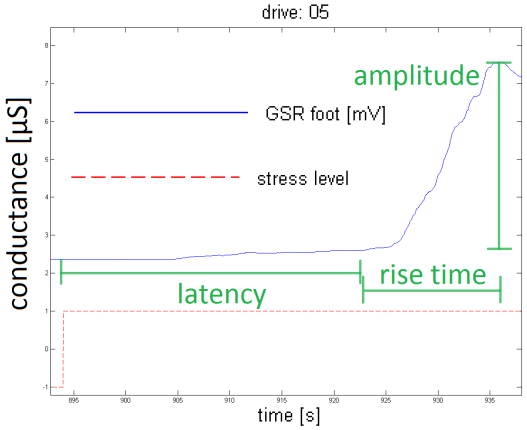

Figure 2.5: *A typical response in skin conductivity to a stressful stimulus (in this case switching from rest to intense driving).*

2.3.4 Skin Temperature

A person's skin temperature (ST) can be influenced by narrowing of blood vessels (vasoconstriction) that can be caused by a sympathetic response followed by pain or mental stress. A temperature sensor can be placed e.g. upon the distal phalanx of the thumb [14]. The ST is however not known as a signal being strongly influenced by stress.

2.3.5 Respiration

It is possible to measure the respiration (RESP) of a person by recording chest expansion. This can be done using a resistor, by measuring its impedance. The respiration of a person might influence the ECG signal, by causing peaks in the low frequencies (< 0.3 Hz) of the ECG spectrogram [44]. See also RSA, explained in Section 2.3.1. Respiration is usually not known as a signal being highly correlated to stress.

2.3.6 Acceleration

An accelerometer (ACC) can be used mainly in combination with other sensors to record when an individual has been moving or not. In this way it is possible to distinguish between physiological reactions caused by movement, and those caused by other means (e.g. psychological stress).

2.4 Sensor Systems

To measure physiological signals, a sensor system is required. Table 2.2 compares commercial sensor systems for measuring stress-related bio-signals. It also indicates whether the systems are wearable or not, and their sampling rates. Some

system	type	wearable	signals	F_s	source
Affectiva Q Sensor	wristband	yes	SCL, ST, ACC	2 – 32 Hz	[22]
Autosense	armband, chestband	yes	ECG, GSR, RESP, ST	10 – 60 Hz	[44]
BodyBugg	armband	yes	GSR	every min	[43]
BodyMedia Sensewear	armband	yes	GSR	every min	[43]
Emotion Board	wristband	yes	EDA, SCL	16 Hz	[25]
Emotiv Research Edition SDK	headset	yes	EEG	128 Hz	[43]
Empatica E4	wristband	yes	PPG, EDA, ACC, ST	4 – 64 Hz	[13]
Wild Divine IOM Device	electrodes	yes	HR, EDA		[23]
Zephyr BioHarness BT	chestband	yes	ECG, HR, RESP, ST	1 – 250 Hz	[14]
Biopac GSR100C	electrodes	no	GSR, HR	1 kHz	[43]
Thought Technology FlexComp		no	GSR, HR, RESP	0 – 40 kHz	[43]

Table 2.2: *Comparison of systems for measuring bio-signals. HR means heart rate, measured by ECG or BVP.*

systems output only raw signals, while others also extract features (such as heart rate from ECG). The sampling rate depends on the signal, generally one wants it to be higher than 128 Hz for ECG, and at least around 16 Hz for EDA. Another aspect to consider is that combining several sensor systems might be complicated in terms of synchronization and adapting the sampling rates. In general, the ECG and the EDA sensors seem to be popular choices for stress estimation.

2.5 Features

When working with large data quantities (such as bio-signals over a large time duration) it is normally a good idea to work with some kind of feature extraction and selection. Feature extraction means reducing the raw data into more comprehensive measures. One example of feature extraction is computing features of signals, e.g. by statistical methods. Some are signal-specific, e.g. rise time of the GSR after a stressful event, while others are more general, e.g. the mean value of a signal during a time window. To decide which features to compute, one can search the raw data for patterns in the signals, especially between different classes. Another example is using a more generic method, such as principal component analysis (PCA). PCA is useful for reducing the dimensionality of a feature

source	signals	# feat.	important feature
[23]	BVP, EDA	12	GSR: mean, sum
[14]	ACC, ECG, HR, RESP, ST	16	
[44]	ECG, GSR, RESP, ST	26	
[43]	BP, BVP, EMG, GSR, HR, ST, RESP	13	EEG features, GSR features, HRV
[25]	ECG, EDA, RESP	16	EDA: mean peak height, slope
[35]	EEG, EMG, face	5	EEG: alpha asymmetry, alpha/beta ratio
[38]	ACC, mobile phone usage, SCL	140	ACC: small median during the 2nd quarter of sleep, ACC: small SD 6-9pm, SMS: few or short sent, screen ons: small # or % of screen on 6-9 p.m. or 9 p.m.-12 a.m.
[29]	BVP, EDA, PPG, RESP	5	HR: mean, mean RESP rate, HRV: LF power, HF power, LF/HF power ratio
[29]	ECG, EMG, RESP, GSR	19	ECG: HRV, LF, HF, EMG: RMS, static load, median load, peak load

Table 2.3: *Features of bio-signals commonly used for stress detection. BP = Blood pressure, face: face measurements. SDNN = standard deviation of all normal RR intervals. HRV = heart rate variability.*

space. It projects the data points onto the axis where the most variance is found, i.e. where there is most information. This gives a transformation from the original feature space to a reduced one, where the data can be studied in three or two dimensions (depending on the number of principal components one chooses). It can give a good overview of the separability of the data. The principal component analysis is independent of the class of the features, it simply transforms the feature space to axes with decreasing importance (they contain less and less information). These features can be calculated over a time window T_f, which is chosen depending on the time constants found in the data.

Furthermore, one must analyze which of these features that are most correlated to the output signal, i.e. the stress level. Table 2.3 compares the signals, the number of features and, if possible, the most important feature indicating stress in different previous studies. The purpose is to reduce the dimensionality of the data and to facilitate the work of classification methods (Section 2.6). Working with recognizable properties in the signals rather than raw data makes the models easier to understand, while they are also more likely to be generalizable (e.g. between different persons). This is further explained in Section 2.5.1.

[60, p. 4] calculates the following features from the following physiological signals:

- BVP: amplitude
- HR: power spectrum, ratio between power low and high frequency, mean, standard deviation
- GSR: number, mean, amplitude, rise time, energy
- ST: mean after finite impulse response (FIR) filtering
- Pupil diameter: mean

The data was normalized as well, using baseline measures from a resting period. The purpose of normalization is to remove subjective differences between individuals (e.g. different resting heart rates). It also forces all features to the same order of magnitude (which also makes them lose their physiological meaning). Having similar data on the same order of magnitude facilitates the work of machine learning algorithms.

[44, p. 2] calculates the following features from the following physiological signals:

- HR: mean, deviation, squared deviation
- ECG: RSA, integration over the HF band (0.15 – 0.5 Hz), integration over power in MF (0.09 – 0.5 Hz) and LF (0.00 – 0.09 Hz) bands, sum of energy in bands 0.0 – 0.1 Hz, 0.1 – 0.2 Hz, 0.2 – 0.3 Hz and 0.3 – 0.4 Hz, LF and MF ratio
- Respiration: mean period, deviation of period, amplitude
- Skin conductance: mean level, deviation, squared deviation, mean absolute deviation
- GSR: number of responses, amplitude of responses in a window, sum of the duration of GSR responses in a window, sum of the area of GSR responses in a window
- ST: mean, deviation, squared deviation

[50, p. 86] calculated heart rate variability (HRV) from the ECG. By distinguishing the P, Q, R and S waves of a regular heart beat, the heart rate and its variance and mean absolute deviation are calculated.

[43, p. 1295] suggests Fourier transformations and wavelet transformations for transforming EEG, GSR and HRV signals to the frequency domain. Wavelet transformation is more suitable for data with sharp spikes and discontinuities. It also suggests principal component analysis (PCA) and independent component analysis for extraction of features from EEG.

2.5.1 Feature Selection

When the features are extracted, one needs to examine which ones contain the most useful information, and remove those who are not contributing to improving the model. Feature selection means choosing a subset among the extracted features that gives a good prediction performance and a small generalization error. The generalization of a machine learning measures its capacity to predict unseen data. A high generalization error means that the model does not perform well on new data. Too many features might lead to overfitting (overly complex models) while including too few features means a risk of losing useful information. One must also keep in mind that some features can perform poorly alone, but can prove very useful in combination. Thus one must be careful while analyzing features one by one.

In this Master's thesis, we define the two classes: (see Section 2.6)

- class 1: "not stress", NS
- class 2: "stress", S

The correlation coefficient, R, is a simple tool for studying the relevance of features and ultimately selecting them [16, p. 4], [11, p. 614]. It assigns a number between -1 and 1 to each feature, indicating their linear correlation with the output signal. $R = -1$ indicates a perfect negative linear correlation and $R = 1$ is a perfect positive one. The linear correlation coefficient is calculated between the feature f and its stress level s by

$$R = \frac{\text{cov}(f, s)}{\sigma_f \sigma_s}, \tag{2.1}$$

where cov is the cross-covariance between two variables and σ is the standard deviation. This can give a preliminary indication of the importance of a feature, but one must keep in mind that it only analyzes linear correlations.

The heteroscedastic t-Test score T for feature f is calculated by [61]

$$T = \frac{\mu_S - \mu_{NS}}{\sqrt{\frac{\sigma_S^2}{N_S} + \frac{\sigma_{NS}^2}{N_{NS}}}}, \tag{2.2}$$

where μ_S is the mean of f over class "stressed", N_S is the total number of samples in the class "stressed". It examines how different the feature values of each class is. For example, if the means are identical, the t-Test score will be equal to 0, indicating a low separability between the classes for the feature [62, p. 243]. The denominator takes into account the standard deviation of the feature over each class, along with the number of samples in each class. This assumes a Gaussian distribution of the feature.

The Fisher score, FS, indicates a ratio between how much a feature separates

itself between two classes [16, p. 4]. It is given by

$$FS = \frac{(\mu_S - \mu_{NS})^2}{\sigma_S^2 + \sigma_{NS}^2}, \tag{2.3}$$

where μ_S is the average of a feature over class "stressed" and σ_S is the standard deviation over "stressed". The same applies for NS, which indicates the class "not stressed". A higher Fisher score means that the feature contains a lot of information for separating two classes. In the case of perfect class balance ($N_S = N_{NS}$), $FS = N_S T^2$.

[16] compares different methods of feature selection for the classification of emotions, mainly Fisher score and correlation coefficient. It also tests Chi-square Score, Gini Index, Information Gain, Correlation Based Filter and Fast Correlation Based Filter. It concludes that they lead to a similar analysis. A summary of these methods can be found in [49].

[56] first uses correlation analysis to reduce 19 extracted features to 9, followed by principal component analysis to further reduce the number to 7. ECG, EMG, SCL, and respiration are recorded for stress detection during driving tasks in [20, p. 1]. It concludes that skin conductivity and heart rate metrics are most closely correlated with the stress levels of the participants. In general, useful features are related to HRV, along with different characteristics of the GSR (e.g. rise time, and slope).

Other, more automated methods include features selection by classification (Section 2.6) accuracy, by adding or removing features based upon if they are making prediction easier or harder. This is called forward and backward feature selection, and many versions and combinations of them exist. [28] compares these kinds of wrapper methods to the filter approach (which is independent of any classification algorithms).

Forward and backward feature selection algorithms were implemented according to Algorithms 1 and 2, either single-user cross validation or multi-user cross validation. The single-user mode means cross validating within the data of a single user, then computing the performance as a mean of all the users. The multi-user mode means leaving one user out as validation data, and using the other users for training the classifier. The performance is then computed as a mean of all the users. The forward algorithms start with an empty feature space, successively adding the feature that increases the classifier performance the most. This is done until all the features have been added and afterwards one can observe the performances to decide what combination of features that was most efficient. The backwards algorithms work identically, except that they start with all the features, successively deleting the feature that decreases the classifier performance the least.

```
featureSpace = [];
while featureSpace is not full do
    for feature f do
        features = featureSpace;
        remove f from features;
        for dataset d do
            validationData = d;
            trainingData = all datasets except d;
            validationData = duplicate(validationData);
            trainingData = duplicate(trainingData);
            model = trainModel(trainingdata);
            [TPrate, TNrate] = predict(validationData.X, model);
            [TPrateOnTraining, TNrateOnTraining] =
            predict(trainingData.X,model);
            performance[d] = mean(TPrate, TNrate);
            performanceOnTraining[d] = mean(TPrateOnTraining,
            TNrateOnTraining);
        end
        performanceOverDatasets[f] = mean(performance);
        performanceOverDatasetsOnTraining[f] =
        mean(performanceOnTraining);
    end
    featureToAdd = max(performanceOverDatasetsOnTraining);
    add featureToAdd to featureSpace;
end
```

Algorithm 1: Multi-user forward feature selection. The validation is done by leaving out one dataset while letting the remaining data sets predict it (cross validating between persons).

```
featureSpace = allFeatures;
while featureSpace is not empty do
    for feature f do
        features = featureSpace;
        remove f from features;
        for dataset d do
            validationData = d;
            trainingData = all datasets except d;
            validationData = duplicate(validationData);
            trainingData = duplicate(trainingData);
            model = trainModel(trainingdata);
            [TPrate, TNrate] = predict(validationData.X, model);
            [TPrateOnTraining, TNrateOnTraining] =
            predict(trainingData.X,model);
            performance[d] = mean(TPrate, TNrate);
            performanceOnTraining[d] = mean(TPrateOnTraining,
            TNrateOnTraining);
        end
        performanceOverDatasets[f] = mean(performance);
        performanceOverDatasetsOnTraining[f] =
        mean(performanceOnTraining);
    end
    featureToRemove = max(performanceOverDatasetsOnTraining);
    delete featureToRemove from featureSpace;
end
```

Algorithm 2: Multi-user backward feature selection. The validation is done by leaving out one data set while letting the remaining data sets predict it (cross validating between persons).

2.6 Classification

To predict the output class of an input where data is already existing, statistical classification methods can be used [19]. These can be used on already labeled data (supervised learning), or with the purpose of discovering new patterns in the data (unsupervised learning) [11, p. 17].

In supervised learning, the purpose is to find a function f that maps the input data x as accurately as possible to the output labels Y. There is an unknown function g (called the ground truth), and the purpose of f is to approximate it. Mathematically this becomes: given training input data X and output data Y, with m training data points: $(X_1, Y_1) \dots (X_m, Y_m)$, to find a classifier $\hat{y} = f(x, \theta)$, where θ are parameters related to the classification function (e.g. tuning). The output $\hat{y}$ of f is the predicted class membership of the input x. This function f can be chosen in different ways. A risk when using classification methods is overfitting the model to the data, i.e. finding patterns that are not generalizable

to new data sets. This puts a restriction on how f can be chosen. One wants it to have as good as possible performance on new data, often called test data or validation data (Section 2.6.7).

In the case of stress detection X is often represented by various features calculated over a time window of one or several sensor inputs. Y is then the stress levels associated with each time window, ranging from two-class problems ("not stressed" and "stressed") to five or more different stress levels [14]. Classification resulting in two classes is usually called detection, which becomes "stress detection" when applying this technique on human stress. In the literature, Y is often given by experiment protocols or questionnaires, while in other cases unsupervised learning is applied, where the algorithm has to find appropriate patterns which distinguish the stress levels [2].

The VC dimension for a set of functions, VC, is the maximum number of points that can be separated in all possible ways by that set of functions. For a classifier $\hat{y} = f(x, \theta)$, this means that there exists a θ such that f can shatter every number of points below VC. Shatter means separating the data points without making any errors (perfect separation). In the case of a two-dimensional classification problem, where the n data points (not placed on the same line) are to be separated by a classifier using a straight line as model, $VC = 3$. This means that the model with the correct parameters θ can shatter any combination of three points [1], however if four data points are present there exists configurations where a line cannot shatter them.

A summary for the notation used for explaining the classifiers:

- f, the classifier function
- θ, parameters related to a classifier
- c, the number of classes
- $y_j \in (y_1, y_2, \ldots, y_c)$, the classes
- X, the training input data
- Y, the training output data (labels)
- m, the number of training data points
- x, new input data
- $\hat{y}$, the predicted class membership of x

[43, p. 1296] compares the following pattern recognition techniques for stress classification:

1. Bayesian classification
2. Decision trees
3. Artificial neural networks

source	signals	classes	classifiers	precision	margin
[38]	ACC, mobile phone usage, SCL	2	SVM, KNN, PCA	> 75 %	
[29]	BVP, EDA, PPG, RESP	2	SVM	85 %	
[60]	BVP, GSR, PD, ST	2	NB, DT, SVM	78.65 %, 88.02 %, 90.10 %	
[25]	ECG, EDA, RESP	2	LDA, SVM, NCC	82.8 %, 79.8 %, 78.0 %	
[56]	ECG, EMG, GSR, RESP,	2	LB, QB, KNN, FSL	78.14 %, 77.78 %, 76.30 %, 79.26 %	2.50 %, 2.07 %, 1.68 %, 1.40 %
[44]	ECG, GSR, RESP, ST	2	SVM	60 %	6.5 %
[14]	ECG, HR, RESP, ST	2, 6	BN	90.35 %, 39.7 %	
[35]	EEG, EMG face	2	FDA	79 %	
[22]	SCL	2	SVM	73.41 %	
[50]	speech	2	KNN, SVM, ANN	89.74 %, 89.74 %, 82.37 %	

Table 2.4: *Comparison of classification methods used in previous studies. BN = Bayesian network, DT = decision tree, FDA = Fisher's discriminant analysis, ANN = artificial neural network, LB = linear Bayes, QB = quadratic Bayes, NCC = nearest centroid classifier, PD = pupil diameter, SVM = support vector machine, KNN = k-nearest neighbor, DT = decision tree, NB = naive Bayes. Note that these studies are based upon different data.*

4. Support vector machines (SVM)
5. Markov chains and hidden Markov models
6. Fuzzy techniques

The conclusion from [43, p. 1297] is that SVM is superior for learning stress models. [44, p. 4] and [60, p. 6] also enjoy success with the same method. Using SVM, up to 90.1% accuracy in differentiation between "relaxed" and "stressed" emotional states could be reached in [60].

Table 2.4 compares the pattern classification methods used in previous studies, along with their performance. The SVM classifier is a popular choice, and seems to be performing well in general. The method is however very black box, and it is hard to analyze the model it outputs. Other suggestions are tree classifiers or Bayesian ones.

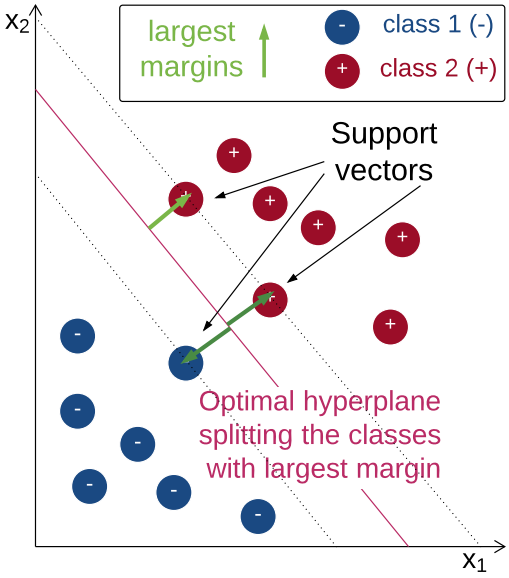

Figure 2.6: *A two-dimensional classification problem, where the two classes "-" and "+" have been split by an optimal line.*

2.6.1 Support Vector Machines

Support vector machines (SVM) work by finding the optimal hyperplane capable of splitting two classes [19, p. 417-419]. This plane is defined as the one with the largest margin towards the closest data point of each class. The idea is to choose the classifier decision function f from hyperplanes $w \cdot x = 0$ (x coordinate vector and w hyperplane cofficients). A two-dimensional example of this is visualized in Figure 2.6, where the optimal line (one-dimensional hyperplane) splitting the two classes has been found. The line is chosen as the one with the largest margins with respect to a misclassification error. Note that this is normally done after a kernel transformation, which is explained later in this section. For hyperplanes of dimension n, $VC = n + 1$, which means SVM classifier can shatter one more point than its hyperplane dimension. This makes it possible for the SVM to deal with data of high dimensions.

Consider the case where w is normalized with respect to a set of training points X^* such that $\min_i \|w \cdot x_i\| = 1$. Vapnik & Chervonenkis showed that the VC dimension for the set of decision functions $f_w(x) = \text{sign}(w \cdot x)$ on $x \in X$, where $\|w\| \leq A$ has the restriction

$$VC \leq (\max_{x \in X} |x|)^2 A^2. \tag{2.4}$$

The problem of obtaining a classifier of greatest margins thus becomes equivalent to minimizing $\|w\|^2$. By choosing $f(x, \theta) = (w \cdot x) + b$ (b corresponding to where the hyperplane intersects the origin) we obtain

$$\frac{1}{m} \sum_{i=1}^{m} l(w \cdot x_i + b, y_i) + \|w\|^2, \tag{2.5}$$

subject to $\min_i |w \cdot x_i| = 1$, where l is the zero-one loss function, $l = 1$ if both arguments equal, otherwise $l = 0$.

If the data are completely separable by a hyperplane the problem is reduced to: Minimize $\|w\|^2$ under

$$y_i(w \cdot x_i + b) \geq 1, \tag{2.6}$$

which is a quadratic program. This optimization problem means finding the hyperplane with the biggest margins between the classes.

In the case of non-separability, the problem is similar. The margin is still maximized, but some points are allowed to be on the wrong side of the hyperplane. For this, the slack variables $\xi = (\xi_1, \xi_2, \ldots, \xi_m)$ are introduced one has to minimize

$$\|w\|^2 + C \sum_{i=1}^{2m} \xi_i, \tag{2.7}$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0,$$

which is also a quadratic program. C is the cost parameter, which in the separable case is equal to ∞ [19, p. 417-422].

To extend the SVM from only dealing with linear separation, the data can be transformed if it is not separable by a plane. This transformation is called $\Phi(x)$ and the decision function becomes

$$f(x) = w \cdot \Phi(x) + b. \tag{2.8}$$

It can increase the dimensionality e.g. by mapping the data from a one-dimensional space to a two-dimensional one. What previously required a quadratic function to separate can then become separable by a line in the new space.

To facilitate the work of the quadratic solver, the kernel trick is used, where the kernel function is defined as

$$K(x_i, x) = \Phi(x_i) \cdot \Phi(x). \tag{2.9}$$

One of the most popular kernels is the radial basis function (RBF),

$$K(x, x') = \exp(-\gamma \|x - x'\|^2), \tag{2.10}$$

which works with the distance of the point to the origin [11, p. 259]. Other possible choices include linear, polynomial and sigmoid kernels [19, p. 423-426].

As seen in Table 2.4 and in [43, p. 1297] SVM classifiers perform well in the domain of stress detection. Due to the kernel transformation and the properties of the VC dimension of SVM:s, data of very high dimensions is not necessarily a big issue. The various kernel choices and tuning parameters makes the choice a sensitive part of the modelling process using SVM:s. The quadratic program has a risk of encountering problems in numerical stability, which in combination with the kernel transformation can be computationally demanding [19, p. 423].

Pros and cons of SVM:s include:

+ Accurate in stress detection
+ Can deal with data of very high dimensions
- Memory-demanding
- Risk of numerical stability problems

2.6.2 Decision Tree Learning

Decision tree learning is the construction of a decision tree from class-labeled training data [19, p. 308-317]. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node. This makes the whole model presentable by a single two-dimensional diagram, which helps its see-through possibility and makes in more white box than e.g. the SVM, which works with transformations of the data. The branch splitting in itself performs a a kind of feature selection by choosing the most discriminative parts of data. The nodes can be both numerical and categorical, which allows a combination of these in the same model. Decision trees can be used for classification, where the nodes represent an attribute test (except for the terminal nodes, which contain a class label). By combining several decision trees, they can be expanded into "random forests", with a potential for higher classification rates. Pruning might be used to reduce the complexity of a decision tree, it means removing the size and number of decisions. This is also useful for avoiding overfitting, which is easily achieved if one keeps expanding the tree with more nodes, i.e. creating a too large tree. A potential problem with decision trees is that the calculation time grows exponentially when the problem size increases, adding more nodes.

Pros and cons of decision trees include:

+ White box, see-through possibility
+ Can select the most important features
+ Can combine numerical and categorical data
- Calculation grows exponentially when the problem size increases

2.6.3 Naive Bayes

The naive Bayes classifier is a simple probabilistic classifier [19, p. 210-211]. Given Y, it assumes conditional independence of all feature variables X in class j, i.e. that no correlations exists between them,

$$P(X_1, \ldots, X_m|Y) = \prod_i P(X_i|Y). \tag{2.11}$$

This independence is an optimistic assumption, however even if the individual estimations of the class density are biased, the posterior probabilities might not be hurt (in particular near the decision regions).

Assuming a normal (Gaussian) distribution of the data, the mean and the standard deviation of each feature are calculated. However there are also other alternatives for this distribution. Given a new data point, its value is compared to the mean and standard deviation of all other points, using the theorem of Bayes. This outputs the probability of the new data point belonging to each class. If new training points are introduced, the only adaptation needed is to recompute the mean and the standard deviations for each class, which facilitates the relearning process of a model.

The naive Bayes classifier prediction $\hat{y}$ as a function of possible classes y_j and the trained conditional probabilities X is

$$\hat{y} = \arg\max_{y_j} P(Y|X_1, \ldots, X_m). \tag{2.12}$$

The product sum is taken over all predictors f, and j represents each class. Pros and cons of NB classifiers include:

+ Simple
+ Possible to adapt to new incoming data easily
- Does not deal with dependent variables

2.6.4 k-Nearest Neighbour

The k-nearest neighbour (KNN) algorithm uses the k nearest samples [19, p. 463-475] to "vote" for the class membership of a new sample, Figure 2.7. k is usually chosen as a small number, and different weighting of each neighbour is sometimes used. A small k is more sensible to noise, but a large k makes the algorithm computationally expensive. For binary classification, odd k is a good idea since this prevents ties in the voting process. Note that the KNN usually works in the feature space. If new data points are introduced, relearning the KNN is simple, since it simply means rechecking the neighbours.

The most commonly used distance to decide which class is nearest for a query point x is the Euclidean distance, d,

$$d_{(i)} = \|x_{(i)} - x\|. \tag{2.13}$$

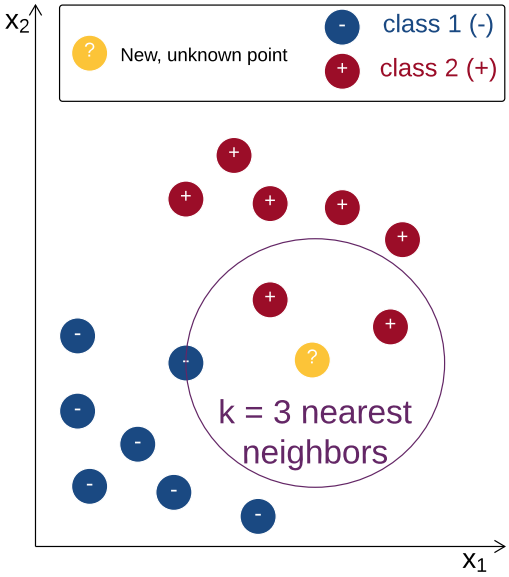

Figure 2.7: *A two-dimensional classification problem, with the two classes: "-" and "+". The class of the new point x (marked by "?") is decided by voting using its $k = 3$ nearest neighbours (1 "-" and 2 "+"), thus the new point is assigned to the "+" class. The Euclidean distance d has been used to define which are the 3 nearest neighbours to x.*

Other options include the Manhattan Distance, the Hamming Distance and the Minkowski Distance [40].

The decision rule for input data x works by creating V_x, which is the vector with the k nearest neighbours in the training data X to x according to the distance measure. Then x is assigned to the most frequent class appearing in V_x.

[11] shows that that for $k = 1$ the error bound for the KNN classifier P has the limitation

$$P^* \leq P \leq P^*(2 - \frac{c}{c-1}P^*), \tag{2.14}$$

where P^* is the Bayes error rate (the lowest possible error rate for a given class of classifier) and c is the number of classes. This means that the KNN error rate never higher than twice the the Bayes error rate, which is promising for an algorithm of this complexity (compare with the training process for e.g. the SVM).

It is a simple classification algorithm, however it has enjoyed success in problems such as handwritten digits, satellite image processing and ECG patterns. Pros and cons of KNN classifiers include:

+ One of the simplest algorithms
+ Updates the model quickly with new data
- Sensible to local structure of data

2.6.5 Probabilistic Neural Network

A probabilistic neural network (PNN) [11, p. 173] is a classifier based upon the statistical algorithm called "kernel discriminant analysis" and consists of a feedforward network containing four layers: input, pattern, summation and output. The first layer represents the set of measurements, while the second one computes Euclidean distances using an RBF kernel. The third layer calculates an average over each class of the outputs of the second layer, while the fourth one decides the associated class by voting [46]. To train a PNN, firstly all training input patterns $\mathbf{x}$ are normalized such that $\sum_i x_i^2 = 1$. Then the first weight w_1 is set equal to the x_1, the first pattern unit. This is repeated for all patterns, such that $w_k = x_k$. The trained net activation net for an input $\mathbf{x}$ and the weights $\mathbf{w}$ is given by their inner product,

$$net = \mathbf{w}^t\mathbf{x}. \tag{2.15}$$

The activation function is then given by

$$e^{(net-1)/\sigma_{PNN}^2}, \tag{2.16}$$

where σ_{PNN} (also known as the smoothing factor) is the width of the Gaussian window, the only tuning parameter needed for the PNN. If the smoothing factor is too small it approximates poorly and if it is too large it has a risk of smoothing out important details.

The PNN decision function (its fourth layer) for c classes with m_j data points in class j becomes

$$\hat{y}_j(x) = \frac{1}{m_j} \sum_{i=1}^{m_j} \exp\Big(- \frac{(\|x_{j,i} - x\|)^2}{2\sigma_{PNN}^2} \Big) \qquad (2.17)$$
$$j = 1, \ldots, c.$$

The class membership of $\hat{y}$ is chosen as class s if $y_s > y_j$, $j \in [1, \ldots, c]$. This reminds of the decision function used by the naive Bayes classifier (Section 2.6.3), which also calculates a value associated with each class, and decides class membership using this value.

The layer-based learning process of the PNN is quick, however it has extensive memory requirements for larger data sets. If the training samples are changed, the model is easily adapted by retraining the relevant network nodes. Pros and cons of PNN classifiers include:

+ Fast learning process
+ Flexible, training samples can be added or removed without extensive retraining
- Large memory requirements for large data sets

2.6.6 Class Imbalance

A common problem when classifying real-world data sets is class imbalance. This means, for example in binary classification, that one of the classes has a lot more data points of one class than the other one. This in turn will bias the classification algorithms to always predicting the majority class, which can give high accuracy but low generalization.

To solve this problem, several methods that are more or less complex exist. Two examples are:

1. duplication
2. SMOTE

The duplication technique means simply copying the minority class until both classes have the same number of samples. Each sample of the minority class is copied the same number of times. The duplication method was implemented by copying the minority class cyclically until class balance is achieved. [8, p. 324] explains similar methods, including random oversampling and oversampling of data points near the class boundaries.

The SMOTE (Synthetic minority over-sampling technique) [8, p. 329] uses a number of neighbours of each sample in the minority class in order to deal with class imbalance. It replicates the minority samples by taking a random step in the direction to the neighbour sample. In this way one can achieve the same number of

samples in each class by introducing artificial data points.

2.6.7 Validation and Performance Measures

To validate the performance of a model, one must introduce new data and observe if the model predicts the correct labels or not [19, p. 219].

If the data is plentiful, one can split it into a training set and a validation set. In this case, the validation data is not used for training, and is only presented when the performance of the final classification algorithm is to be tested. Once this is done one must be careful when trying to improve the model for further accuracy, since one risks to overfit it to the data.

For smaller data sets, it is preferable to keep as many samples as possible for learning. In this case, cross validation is widely used [19, p. 241]. A common type is leave-one-out cross validation, where one observation of the data is removed, and the rest are used for training the model. The model is then tested by letting it predict the observation that was removed. This is then repeated for each observation, to obtain a robust measure of performance. Finally the mean of these results is calculated as the performance. One must consider if it is suitable to split the data in time segments (following each other temporally), or in random segments. This depends on how the data is generated, and how correlated nearby samples are to each other.

The performance of a binary classifier on a given data set is related to four factors derived from a prediction and the corresponding true value [8, p. 323]:

- number of true positives (TP)
- number of false positives (FP)
- number of true negatives (TN)
- number of false negatives (FN).

TP is the number of accurately predicted positives. In the domain of stress detection, it corresponds to the number of samples predicted as "stress" when the person actually is stressed. Similarly, FP is the number of falsely predicted positives, TN is the number of correctly predicted negatives and FN is the number of falsely predicted negatives.

Using these four measures, one can define the confusion matrix $M_{confusion}$:

$$M_{confusion} = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}.$$

The confusion matrix summarizes the performance of a classifier. The values at the top left and bottom right (TP and TN) need to be as high as possible for a good classifier, while the ones at bottom left and top right needs to be as low as possible (optimally 0).

Using the values from $M_{confusion}$ one can then define [8, p. 322-326]:

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{TN + FP}$$

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the percentage of correctly predicted positives among all existing positives. It is also known as positive predictive value. It is the proportion of samples correctly predicted as "stress" among all samples predicted as "stressed".

Sensitivity is also known as true positive rate, hit rate, or recall. It is the proportion of samples correctly predicted as "stress" among all "stressed" samples.

Specificity is also known as the true negative rate and measures the proportion of correctly classified negatives (correctly predicted as "not stress" among all "not stressed" samples).

Accuracy is the proportion of correctly predicted samples among all samples, accuracy = 100% means that all samples are correctly classified.

In the case of class imbalance (where one of the classes is heavily underrepresented), one can introduce the g_{mean} [47, p. 3362],

$$g_{mean} = \sqrt{\text{sensitivity} \cdot \text{specificity}}, \tag{2.18}$$

which is a non-linear measure of a binary classifier's performance, punishing majority class misclassification more than minority class misclassification.

In this work we then define the performance p of a classifier as a choice between one of these measures or a combination of them. Examples include:

- p = accuracy
- $p = \frac{\text{sensitivity+specificity}}{2}$
- $p = g_{mean}$

For an accuracy the associated margin of error (E) can be calculated,

$$E_{95} = 1.96\sqrt{\frac{\text{accuracy}(1 - \text{accuracy})}{n}}, \tag{2.19}$$

which describes the margin of error at 95 % confidence interval (i.e. 1.96 standard deviations for a Gaussian distribution), denoted E_{95}. n is the number of predicted examples.

2.7 Continuous Stress Models

[20, p. 10] creates a continuous stress metric by testing different correlations between physiological signal features and a stress level based upon video recording.

2.7.1 Linear Regressive Model

One of the simplest regressive models is a linear regressive model [19, p. 44], a linear model with parameters adapted by least squares between features and stress levels. The model consists of a simple vector of dimension $1xA$, where A is the number of features. It is trained by

$$model_{lin} = Y_{train}X^{+}_{train}, \tag{2.20}$$

where X^{+}_{train} is the Moore-Penrose pseudoinverse [17, p. 257-258] of the training data (implemented by the Matlab function `pinv`).

A prediction $\hat{y}_{lin}$ of new data X_{val} is then performed by

$$\hat{y}_{lin} = model_{lin}X_{val}. \tag{2.21}$$

2.7.2 Support Vector Regression

Support vector regression (SVR) is based upon the same mathematical foundations as the SVM (Section 2.6.1), but instead of predicting a class membership, $\hat{y}$ is a regressive prediction in the case of SVR, taking continuous values.

The problem formulation of ν-SVR can be described as [6]

$$\begin{aligned} \min \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\Big(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\Big) \\ (\boldsymbol{w}^T\phi(x_i) + b) - y_i \leq \epsilon + \xi_i \\ y_i - (\boldsymbol{w}^T\phi(x_i) + b) \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1, \ldots, l, \epsilon \geq 0, \end{aligned} \tag{2.22}$$

where C is once again the cost parameter, ξ_i the slack variables and w the hyperplane coefficients, and Φ the transformation. ϵ decides what errors to include, ignoring errors of size less than ϵ (which in the case of SVM:s corresponds to the non-support vector points that are on the correct side of the decision boundary). The ν parameter controls the number of support vectors and training errors, by being an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors. A more elaborate explanation, including kernel transformation and prediction can be found in [19, p. 434-437].

2.7.3 Variational Multiple Bayesian Linear Regression

The variational multiple Bayesian linear regression model (VBML) [3, p. 486-490], combines linear regression with variational inference. The model parameters θ for training data pairs $(Y_i, X_i), 1 \leq i \leq m$, consisting of f regressors, has the

likelihood function

$$p(Y|\theta, \alpha, \beta) = \prod_{i=1}^{m} \mathcal{N}(Y_i|\theta^T \Phi(X_i), \beta^{-1}), \tag{2.23}$$

where $\mathcal{N}$ corresponds to a Gaussian distribution, β is the noise precision parameter and Φ is the basis function [3, p. 139]. The prior over θ is given by

$$p(\theta|\alpha) = \mathcal{N}(\theta|0, \alpha^{-1} I), \tag{2.24}$$

where α is a precision parameter. For the precision of a Gaussian, the conjugate prior is a gamma distribution [3, p. 688],

$$p(\alpha) = \Gamma(\alpha|a_0, b_0). \tag{2.25}$$

Similarly, for β the the conjugate prior is also a gamma distribution,

$$p(\beta) = \Gamma(\beta|c_0, d_0). \tag{2.26}$$

The joint distribution for all variables is then given by

$$p(Y, \theta, \alpha, \beta) = p(Y|\theta, \alpha, \beta) p(\theta|\alpha) p(\alpha) p(\beta). \tag{2.27}$$

To approximate the posterior distribution $p(\theta, \alpha, \beta|Y)$, the variational posterior distribution

$$q(\theta, \alpha, \beta) = q(\theta) q(\alpha) q(\beta) \tag{2.28}$$

is used, since no analytical solution exists.

The variational re-estimation equation for the posterior distribution over θ can then be found by using [3, p. 466] and identifying coefficients,

$$\begin{aligned} \ln q^\star(\theta) = \mathbb{E}_{q(\alpha)q(\beta)}[\ln p(Y, \theta, \alpha, \beta)] + \text{const} \implies \\ q^\star(\theta) = \mathcal{N}(\theta|\mu_n, \Lambda_n), \end{aligned} \tag{2.29}$$

where $\Lambda_n = (\mathbb{E}[\beta]\Phi(X)^T \Phi(X) + \mathbb{E}[\alpha] I)^{-1}$ and $\mu_n = \mathbb{E}[\beta] \Lambda_n \Phi(X)^T Y$.

Similarly, for α

$$\begin{aligned} \ln q^\star(\alpha) = \mathbb{E}_{q(\beta)q(\theta)}[\ln p(Y, \theta, \alpha, \beta)] + \text{const} \implies \\ q^\star(\alpha) = \Gamma(\alpha|a_n, b_n), \end{aligned} \tag{2.30}$$

where $a_n = a_0 + \frac{f}{2}$ and $b_n = b_0 + \frac{\mathbb{E}[\theta^T \theta]}{2}$.

For the noise precision parameter β this becomes

$$\begin{aligned} \ln q^\star(\beta) = \mathbb{E}_{q(\theta)q(\alpha)}[\ln p(Y, \theta, \alpha, \beta)] + \text{const} \implies \\ q^\star(\beta) = \Gamma(\beta|c_n, d_n), \end{aligned} \tag{2.31}$$

where $c_n = c_0 + \frac{m}{2}$ and $d_n = d_0 + \frac{Y^T Y - 2\mathbb{E}[\theta]^T \Phi(X)^T Y + \mathbb{E}[\theta]^T \Phi(X)^T \Phi(X) \mathbb{E}[\theta]}{2}$.

Using the properties of the Γ distribution [3, p. 688], $\mathbb{E}[\theta]$, $\mathbb{E}[\theta^T \theta]$, $\mathbb{E}[\alpha]$ and $\mathbb{E}[\beta]$

can be obtained as

$$\begin{aligned} \mathbb{E}[\theta] &= \mu_n \\ \mathbb{E}[\theta^T \theta] &= \mu_n^T \mu_n + \Lambda_n \\ \mathbb{E}[\alpha] &= \frac{a_n}{b_n} \\ \mathbb{E}[\beta] &= \frac{c_n}{d_n}. \end{aligned} \tag{2.32}$$

To evaluate the variational posterior distribution, the parameters of either $q(\theta)$, $q(\alpha)$ or $q(\beta)$ are initialized, then they are alternately re-estimated until the difference in free energy [5, p. 349] of the model converges.

Given a new input x, the predictive distribution over Y_i is approximated as

$$p(Y_i|x, Y) \approx \mathcal{N}\Big(Y|\mu_N^T \Phi(x), \mathbb{E}[\beta]^{-1} + \Phi(X)^T \Lambda_n \Phi(X)\Big). \tag{2.33}$$

For more details, see [3, p. 486-490].

2.7.4 Performance Measures

The performance of these regressive models can be computed as the root mean square (RMS) error:

$$RMSE = \sqrt{\frac{\sum_s^n (\hat{y} - Y_{val_s})^2}{n}}. \tag{2.34}$$

This value calculates the difference between the true value Y_{val} and the predicted value $\hat{y}$ in each sample. It squares this difference and adds together all errors, then dividing by the number of data points n.

3

MIT Driver Database

As a first study of stress feature selection and classification, The MIT Driver Stress Database [21] was chosen to be analyzed, while waiting for experimental data to be made available (Chapter 4). The MIT database origins from experiments with subjects in rest and driving in cities and on highways, done at the Massachusetts Institute of Technology. In Table 3.1 the data sets and the signals of the database are presented. The recorded signals are ECG, HR (derived from ECG), EMG, GSR of the foot and the hand and finally the respiration. The ECG was recorded at 496 Hz, the GSR and the respiration signals at 31 Hz and the EMG at 15.5 Hz. The signals finally available in the database are down-sampled versions of these, normally to 15.5 Hz. The "marker" signal indicates a change of situation, i.e. the start of a period of rest, city driving or highway driving.

3.1 Method

In Figure 3.1 an overview of the data processing from the database to a model with associated accuracy is presented. The block "read database" is adapted to a given database, it restructures the data and outputs raw signals, along with their sampling frequency. The pre-processing steps consists of refining the data, such as removing artifacts.

The pre-processing step is also responsible for creating a binary stress level signal from the marker signal. The features of each signal are then extracted, which is followed by a split of the data. It is possible to modify how many parts one wants to split it in, and if the parts should be picked ordered by time or randomly. The normalization parameters are then estimated using the input features, along with

drive	$F_s[Hz]$	ECG	HR	EMG	foot GSR	hand GSR	RESP	marker
01	15.5	x	x	x	x	x	x	
02	15.5	x	x		x		x	x
03	31.0	x			x	x	x	
04	15.5	x	x		x	x	x	x
05	15.5	x	x	x	x	x	x	x
06	15.5	x	x	x	x	x	x	x
07	15.5	x	x	x	x	x	x	x
08	15.5	x	x	x	x	x	x	x
09	15.5	x	x	x	x	x	x	x
10	15.5	x	x	x	x	x	x	x
11	15.5	x	x	x	x	x	x	x
12	15.5	x	x	x	x	x	x	x
13	15.5	x	x	x	x		x	x
14	15.5	x		x	x	x	x	x
15	15.5	x	x	x	x		x	x
16	15.5	x	x	x	x	x	x	x
17a	15.0	x	x	x	x	x	x	x
17b	15.5	x	x	x	x	x	x	x

Table 3.1: *The MIT driver database. The sampling frequency and the signals recorded in each data set (drive) are indicated.*

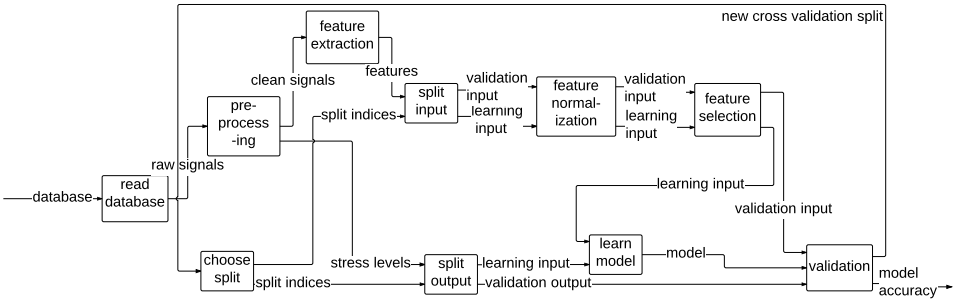

Figure 3.1: *Data flow chart of how the signals were processed.*

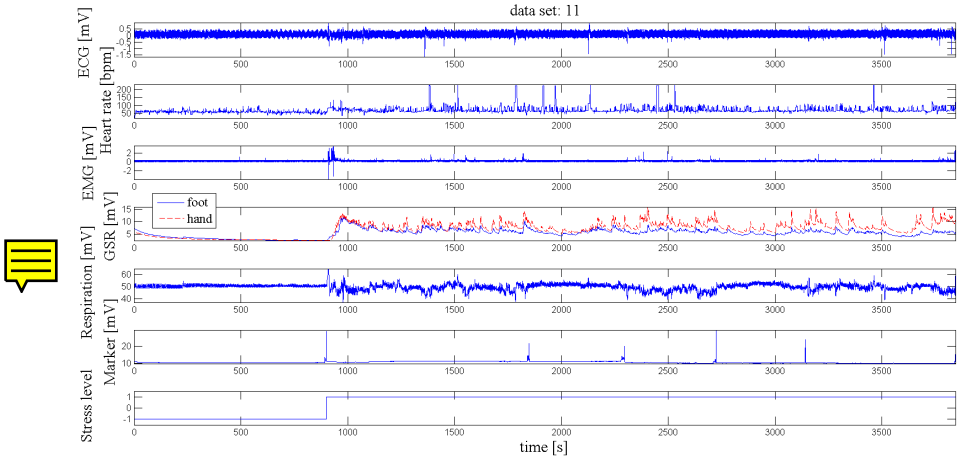

Figure 3.2: *The raw signals recorded in drive 11, and the stress levels derived from the marker signal.*

their output labels. The feature selection block simply reduces the features to a pre-defined set, but this is possible to modify for comparison of different feature sets. Finally a model is trained (several different classification methods can be used), and in the end the accuracy of the model is calculated by validation. By adding a loop testing different signal partitions, a cross validation is obtained.

3.1.1 Preprocessing

The "marker" signal was converted into binary stress levels (−1 and +1), depending on if the subject is resting or driving during the current sample. Furthermore, data points before the first marker and after the last marker were removed, since these correspond to situations where the subject is being equipped and unequipped with sensors, respectively.

How binary s

In Figure 3.2 an example of a data set containing raw signals can be obs The stress levels signal is however derived from the marker signal by lo for its peaks (indicating that a button was pressed to record that a new phase of the experiment is commencing). Data corresponding to noise while equipping sensors has also been removed at the beginning and the end of the set.

Since drives 01 and 03 lack the marker signal, they were not used for learning any models. Drives 02, 13 and 15 do not contain a GSR signal from the hand, thus they were excluded. This decision is motivated by the fact that it is interesting to compare both GSR signals, especially for feature selection. Drive 05 was excluded for missing ECG and HR data during 300 seconds (the sensor probably lost its connection at this point).

Since the signals available are all down-sampled to 15.5 Hz, this makes especially

the ECG very challenging to analyze since the peaks are hard to distinguish (both for the human eye and by peak detecting software). Consequently, the HR-signal was used for calculating the desired features related to heart activity. However, drive 14 lacks this HR-signal so it was eliminated.

The drives 17a and 17b are in fact recorded during the same session, but split into different data sets with different sampling rates. These were interpolated to 15.5 Hz and merged into one single data set, where 17b follows 17a. This interpolation of drive 17a introduced unwanted noise in its GSR signals, thus they were median filtered to smoothen them and avoid false detection of local maxima.

The decision was taken to exclude the EMG signal, since it opens up the possibility to use an additional data set (drive 04), but mainly because of its placement upon the shoulder might record muscle movement (using the steering wheel) rather than a psychological stress response from the traffic (e.g. muscle tension). Consequently one ECG signal with its derived HR values, two GSR signals and the respiration signal were used as inputs for the stress modelling process. Finally, the following 10 data sets were complete and useful for training stress-predicting models: 04, 06, 07, 08, 09, 10, 11, 12, 16 and 17.

Artifacts are present, mainly in the HR signal, which sometimes drops to unreasonably low 30 bpm or unreasonably high 200 bpm. Thus it was filtered by removing points outside a tuning constant times the standard deviation of a signal window. At a few time windows of some seconds, the GSR signal turned negative, and visual inspection confirmed that it is very probably a simple error of sign behind these artifacts. These values were therefore multiplied by -1. This is motivated by the fact that when the negative values occurred, the signal looked exactly as if coherent with nearby values, but with the wrong sign. GSR values suddenly falling to exactly zero correspond to a sensor problem, and were replaced by the precedent value. The signals were all visually analyzed to confirm that no unreasonable physiological data was to be used in steps further ahead. After visual analysis of the "post-stress" period (i.e. the resting period after driving) it was decided to remove this part since the signals look like a mix between the pre-driving rest period and the driving session. This kind of contradictory data would give the classifier a hard time.

Each data series was split into time segments, to make it work with features over a time window rather than samples. This choice is motivated by the fact that it is desirable to work with time windows that are sufficiently large to contain enough samples in the very low frequency (VLF) spectrum ($0.01 - 0.04$ Hz) while also keeping the possibility of detecting the stress within a reasonable time. Too big time segments will also reduce the number of points available for learning. Since each drive has a duration of around $1 - 1.5$ hours, time segments of $T_f = 60$ s were deemed suitable, giving around 75 different segments for each drive. The chosen duration of each time segment can however be modified in the program. It is also important to keep a balance between the number of features and the number of samples to avoid overfitting.

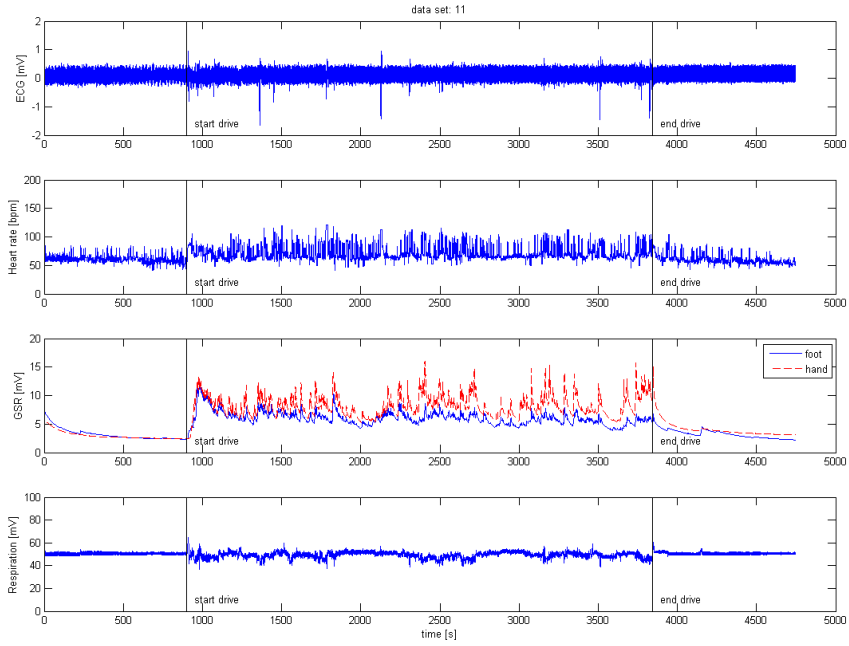

Figure 3.3: *The pre-processed data of drive 11. Artifacts have been removed.*

The segments corresponding to the change between rest and driving were removed, in order to avoid feeding mixed data into the learning algorithm (since it is not clear whether the subject is stressed or not during this segment). Furthermore, the last segment is not of the same duration as the rest (except if the total duration of the drive is evenly divisible by the segment duration). Because of this, the last segment was also removed from the data.

The final product of the pre-processing is 10 complete data sets with 5 signals and one label vector ("not stressed" or "stressed"), all split into time segments of one minute.

In Figure 3.3 the data of drive 11 have been pre-processed (compare with Figure 3.2). The artifacts have been removed (mainly present in the HR signal in the form of outlier values) and the GSR signals have been turned positive where they were negative before.

3.1.2 Feature Computation

Based upon visual data inspection and the previous studies analyzed in Section 2.5, the decision was taken to calculate the following features:

From the ECG and HR signals:

1. mean of HR
2. standard deviation of HR
3. root mean square of successive differences
4. standard deviation of successive differences
5. HRV power in VLF (0.01 – 0.04 Hz)
6. HRV power in LF (0.04 – 0.15 Hz)
7. HRV power in HF (0.15 – 0.50 Hz)
8. ratio between HRV power in LF and HF

The mean is defined as the arithmetic mean of n samples of a variable x,

$$\text{mean}(x) = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{3.1}$$

The standard deviation (SD) is a measure of the difference between the samples of the same dimension as the samples

$$SD(x) = \sqrt{\text{mean}(x_i - \text{mean}(\text{x}))^2}. \tag{3.2}$$

The successive differences of the heart rate (HR') are firstly calculated by Algorithm 3, which simply outputs the difference between the current heart rate and the previous heart rate for each value, as a measure of heart rate variability. This is the numerical derivative computed by the Matlab function `diff`.

for *sample s* **do**
  HR'(s) = HR (s) - HR (s-1);
end
Algorithm 3: The algorithm for calculating the successive differences of the heart rate (HR').

Feature 3 is computed as

$$f_{RMSHR'} = \sqrt{\text{mean}((\text{HR}')^2)}, \tag{3.3}$$

and Feature 4 is computed as

$$f_{SDHR'} = SD(\text{HR}'). \tag{3.4}$$

Features 5, 6 and 7 are calculated by Welch's method [55], estimating the power spectra in the respective frequency bands. This method converts the signal to the frequency domain and is based upon periodogram spectrum estimates.

For a signal x, the mth windowed, zero-padded frame is

$$\begin{aligned} x_m(n) &= w(n)x(n+mR) \\ n &= 0, 1, \ldots, M-1 \\ m &= 0, 1, \ldots, K-1, \end{aligned} \tag{3.5}$$

with the parameters R as the window hop size and K the number of available frames. The mth block has the periodogram

$$P_{x_m,M}(w_k) = \frac{1}{M}|FFT_{N,k}(x_m)^2|, \tag{3.6}$$

where FFT is the fast Fourier transform, defined as

$$X(k) = \sum_{n=1}^{N} x(n) \exp\Big(- i2\pi(k-1)\frac{n-1}{N}\Big), 1 \leq k \leq N. \tag{3.7}$$

for an input signal x of length N.

Welch's method for estimating the power spectral density is then calculated as

$$\hat{S}_x^W(w_k) = \frac{1}{K}\sum_{m=0}^{K-1} P_{x_{m,M}}(w_k). \tag{3.8}$$

The mean value of the HR is subtracted before the estimation is done. The signal is calculated upon windows of length equal to the sampling frequency, in this case $F_s = 15.5$ Hz.

From the GSR signals measured at foot and hand:

1. mean
2. mean of derivative
3. mean of negative derivative
4. mean of absolute derivative
5. proportion of negative samples in the derivative vs all samples
6. number of local maxima

Features 2, 3, 4, 5 and 6 are all computed upon the derivative of the GSR signal. This derivative (GSR') is calculated by Algorithm 4 (similarly to Algorithm 3, the numerical derivative computed by the Matlab function `diff`). The five derivation features are then chosen as the mean, the negative part, the absolute value, the proportion of negative values of GSR', and finally all values where GSR' is positive while GSR'' (its second derivative) is negative.

for *sample s* **do**
  GSR'(s) = GSR (s) - GSR (s-1);
end
Algorithm 4: The algorithm for calculating the derivative o the GSR signal (GSR').

From the respiration signal:

1. SD of breathing signal
2. range (or greatest breath)
3. mean of breathing rate

Feature 3 is computed first by choosing a symmetric Hanning window,

$$w(n) = 0.5(1 - \cos\left(\frac{2\pi n}{N-1}\right)), \tag{3.9}$$

where N is the total number of samples. The FFT (3.7) is then calculated upon each Hanning window.

To find the breathing rate, the maximal value of the absolute value of the fast Fourier transform is found. This maximal value corresponds to the most common frequency in $X(k)$.

Another feature, "feature x" was calculated from one of the signals. Due to a possible future patent, the nature of this feature will not be explained here, but it is compared in the following analysis.

These features add up to a total of 24. A 25th feature, "random values", consisting of random values between 0 and 1, was also added for comparison purposes. It was calculated using the Matlab function `rand`. The idea is to introduce a benchmark feature that approximates a feature completely unrelated to stress detection.

The calculated features of drive 11 are presented in Figure 3.4.

After computation, each feature f was normalized using

$$f_{norm} = \frac{f - mean(f_{rest})}{SD(f_{rest})}, \tag{3.10}$$

where f_{rest} are samples f taken from a resting period, which is this case consists of samples from the "non-stress" class. The window was chosen at the end of the "non-stress" rest period, taking the last 6 time segments. These values were subsequently removed from the data to avoid it influencing any classification accuracy (since every other feature value in the series is influenced by the normalization data). This results in each feature having values in the same order of magnitude (around −10 to 10), and compensates for individuals having e.g. different resting heart rates.

3.1.3 Class Imbalance Problem

After deletion of the "post stress" period of the data sets, each data set consists of around 80% "stress" class and 20% "not stress" class. The two methods presented in Section 2.6.6 for dealing with this class imbalance were examined (along with the option of doing nothing):

1. no extension

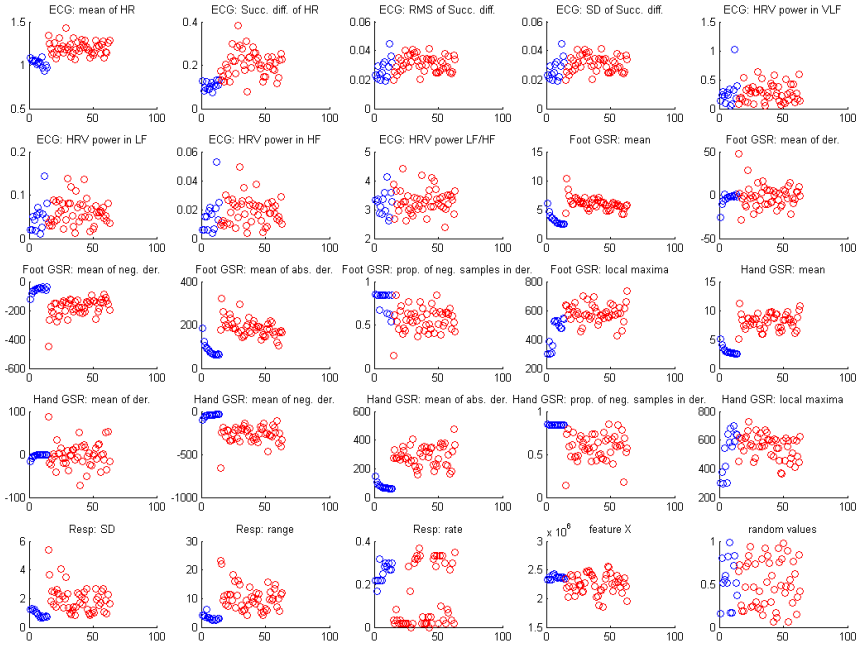

Figure 3.4: *The calculated features of drive 11. Each feature is calculated over T_f = 60 seconds, and the value is represented by a dot. Blue circles are from the rest period, while the red circles are from the driving session. The x-axis represents the feature value at the corresponding time segment of the y-axis. y-axis units (from the upper left corner): Hz, Hz, Hz, Hz, s, s, s, dimensionless, μS, $\frac{\mu S}{s}$, $\frac{\mu S}{s}$, $\frac{\mu S}{s}$, dimensionless, dimensionless, μS, $\frac{\mu S}{s}$, $\frac{\mu S}{s}$, $\frac{\mu S}{s}$, dimensionless, dimensionless, V, V, Hz, unit X, dimensionless.*

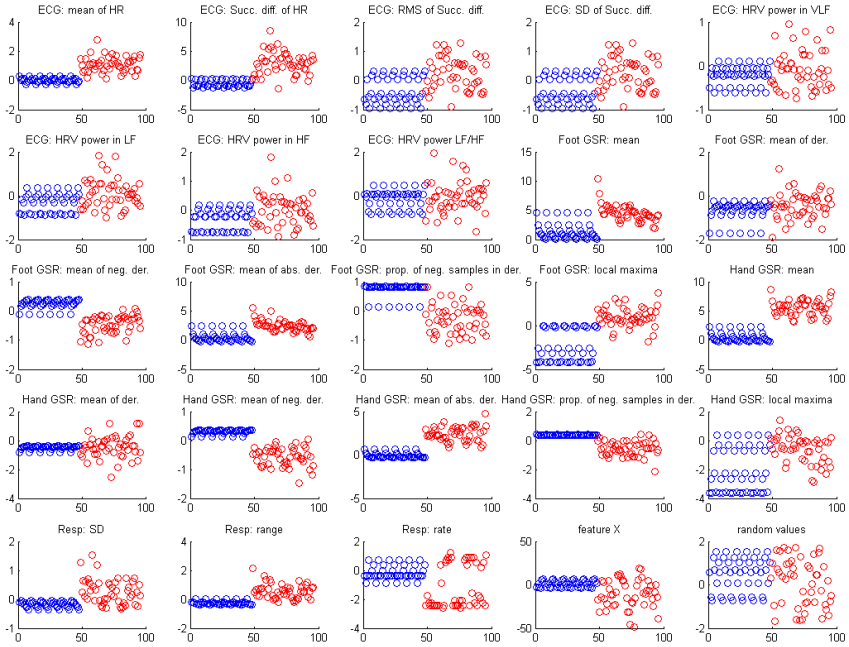

Figure 3.5: *The extended features of drive 11. The features of the "not stress" class have been extended using duplication to achieve class balance. Blue circles are from the rest period, while the red circles are from the driving session. The x-axis represents the feature value at the corresponding time segment of the y-axis. These features are normalized, making all y-axis values dimensionless.*

2. duplication
3. SMOTE

The result of the duplication method of feature extension on drive 11 can be studied in Figure 3.5. The SMOTE method was implemented and tested upon the data (Figure 3.6), however as the duplication does not produce any artificial data points, it was decided to be used in the subsequent analysis. The duplicated features in Figure 3.5 are more regular than in Figure 3.6. The reason behind this is that the data in the duplication method consists of pure copying while the SMOTE method creates new data points based upon the existing data.

3.1.4 Feature Selection

The following analyses were made:

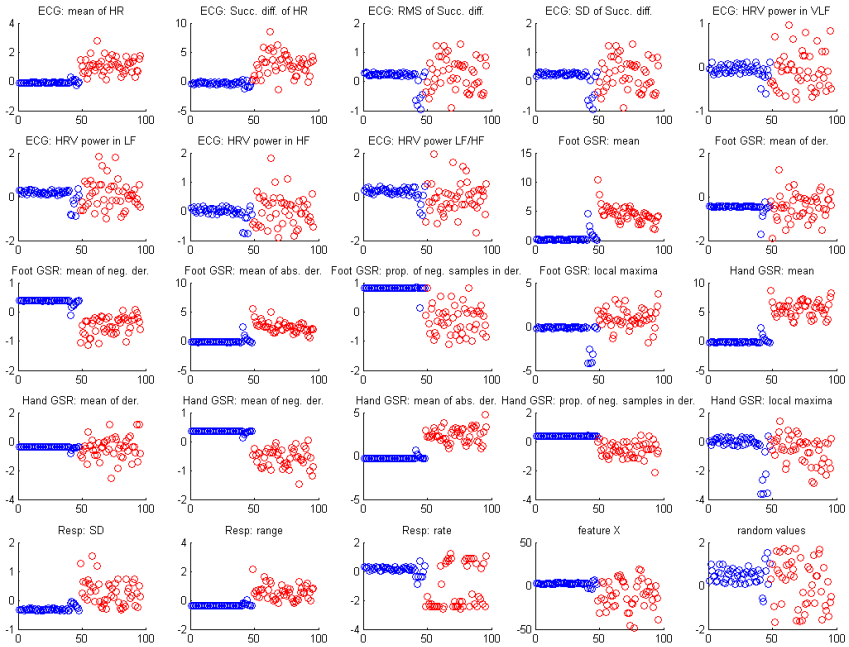

Figure 3.6: *The extended features of drive 11. The features of the "not stress" class have been extended using SMOTE to achieve class balance. The added data points are placed at the end of the rest period. Blue circles are from the rest period, while the red circles are from the driving session. The x-axis represents the feature value at the corresponding time segment of the y-axis. These features are normalized, making all y-axis values dimensionless.*

1. linear correlation coefficient, (2.1)
2. t-Test, (2.2)
3. Fisher score, (2.3)
4. forward feature selection, Algorithm 1
5. backward feature selection, Algorithm 2
6. PCA, 1, 2 and 3 components

The Gini index, and the Gram-Schmidt orthogonalization process were also examined, but led to similar results which are not presented in this work. Histogram plots were also done for each feature and driver, to observe if the feature values were indeed changing between the classes or not.

For deciding what features to keep and remove in the forward and backward selection algorithms, an SVM with a radial basis function was used.

3.1.5 Classification

A naive Bayes classifier using a Gaussian distribution and two classes was implemented in Matlab. It consists of two functions (Algorithms 5 and 6).

```
nbTrain X,Y: model:
class1data = X(Y==class1);
class2data = X(Y==class2);
for feature f do
    class1.means(f) = mean(class1data(f));
    class1.SDs(f) = SD(class1data(f));
    class2.means(f) = mean(class2data(f));
    class2.SDs(f) SD(class2data(f));
end
class1.Nosamples = length(class1.data);
class2.Nosamples = length(class2.data);
model.class1 = class1;
model.class2 = class2;
```

Algorithm 5: The implemented Naive Bayes classifier's training function.

```
nbClassify X,model: prediction:
class1mean = model.class1.means;
class1SD = model.class1.SDs;
class2mean = model.class2.means;
class2SD = model.class2.SDs;
totalSamples = model.class1.noSamples + model.class2.noSamples;
PClass1 = class1noSamples / totalSamples;
PClass2 = class2noSamples / totalSamples;
newData = X;
for newData d do
    newPoint = newData(:,d);
    pClass1 = normpdf(newPoint,class1mean,class1SD);
    pClass2 = normpdf(newPoint,class2mean,class2SD);
    evidence = PClass1 * prod(pClass1) + PClass2 * prod(pClass2);
    posteriorClass1 = PClass1 * prod(pClass1) / evidence;
    posteriorClass2 = PClass2 * prod(pClass2) / evidence;
    if posteriorClass1 > posteriorClass2 then
        prediction (d) = class1;
    else
        prediction (d) = class2;
    end
end
```

Algorithm 6: The implemented Naive Bayes classifier's classification function.

The SVM classifier from LIBSVM [7] 3.18 was used. The RBF kernel was used, with $\gamma = 1/\text{numberOfFeatures}$. The PNN and the KNN classifiers from the toolboxes included with [11] were used. $k = 1$ was chosen for the KNN classifier in single user mode, since it is not possible to use more neighbours if one wants to perform leave-one-out cross validation. For multi-user mode and every other case where leave-one-out cross validation was not done, $k = 3$ neighbours were chosen. For the PNN classifier, the Gaussian width was chosen as $\sigma_{PNN} = 0.5$. $\sigma_{PNN} = 1$ was also tried, but with worse results. Decision tree classifiers were left out due to a suitable implementation not being available, and lack of time of making an own implementation.

3.2 Results

The results of the analysis made on the MIT driver database are presented in this section.

3.2.1 Feature Selection

The results of the feature selection analysis explained in Section 3.1.4 are visualized in this section.

In Figure 3.7 the Fisher scores for each feature and data set can be compared. Prominent features include:

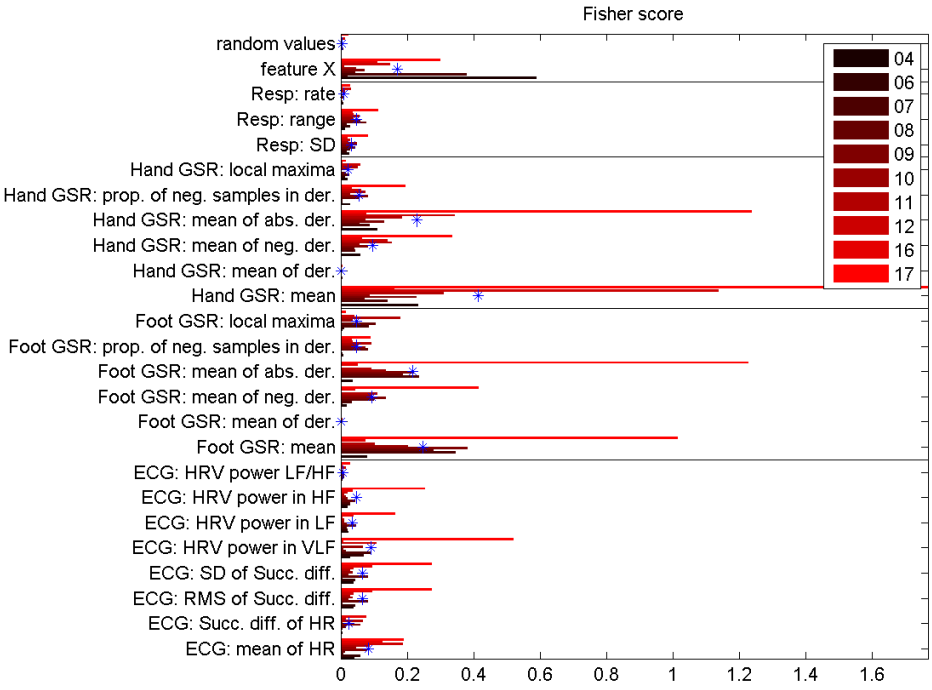

Figure 3.7: *Fisher score, calculated for each drive. Mean is indicated by "*".*

1. Hand GSR: mean
2. Foot GSR: mean
3. Hand GSR: mean of abs. der.
4. Foot GSR: mean of abs. der.
5. feature x

These are chosen because they have higher Fisher scores than the other features among a majority of the users.

In Figure 3.8 the linear correlations for each feature and data set can be compared. Features having the same effect on each user are all pointing the same way (left for negative correlation i.e. the feature value is lower in the driving state and right for positive correlation, i.e. the feature is higher in the driving state). This occurs for the mean of the negative derivative and the proportion of negative samples in the derivative in both GSR signals (negative correlations). This is also true for the mean of HR and the LF/HF power ratio (positive correlations). Prominent features, that have mean absolute linear correlations higher than 0.5, but behave exceptionally in some cases, include:

1. ECG: mean of HR
2. feature x
3. Hand GSR: mean
4. Hand GSR: mean of neg. der.
5. Hand GSR: mean of abs. der.
6. Foot GSR: mean
7. Foot GSR: mean of neg. der.
8. Foot GSR: mean of abs. der.

In Figure 3.9 the t-Test scores for each feature and data set can be compared. Features having the same effect on each user are all pointing the same way (left for if the feature value is lower in the driving state and right if the feature value is higher in the driving state). Prominent features include:

1. ECG: mean of HR
2. feature x
3. Hand GSR: mean
4. Hand GSR: mean of neg. der.
5. Hand GSR: mean of abs. der.
6. Foot GSR: mean
7. Foot GSR: mean of neg. der.

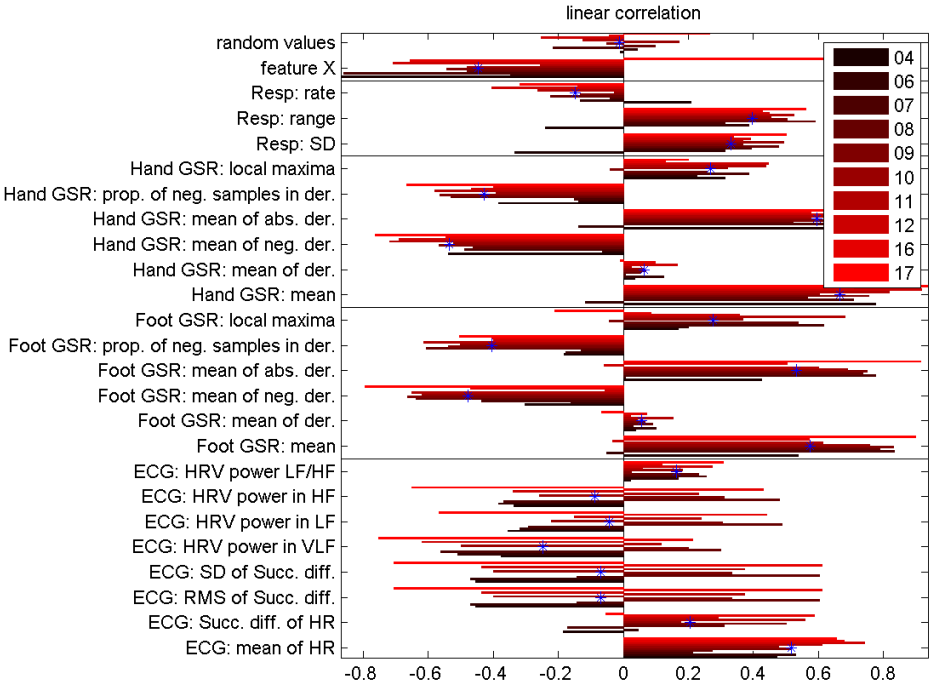

Figure 3.8: *Linear correlation coefficient, calculated for each drive. Mean is indicated by "*".*

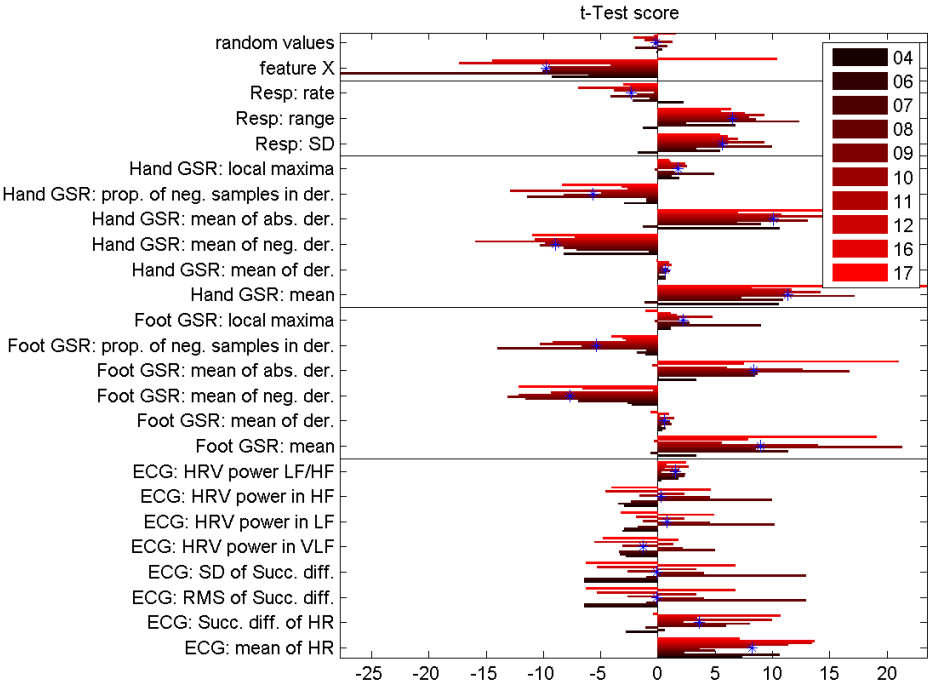

Figure 3.9: *t-Test score, calculated for each drive. Mean is indicated by "*".*

8. Foot GSR: mean of abs. der.

These were chosen because they have a higher t-Test score than the other features, and also because T in general have the same sign for all users (only negative or only positive). This indicates that the mean of the feature changes more and in the same way for all the subjects compared to other features.

The results of the forward feature selection algorithms can be studied in Figure 3.10 and Figure 3.11 (Algorithm 1, multi-user case). The feature space tested is accumulated upwards, meaning that firstly it only consists of the feature that is located furthest down, secondly it consists of the feature furthest down combined with the feature second furthest down etc. The feature highest up is the last one added, which means that all features are in feature space at this point. The feature choice is based upon only the training data, but the performance on the validation data is also presented in each graph. The features under the black line in each of these figures represent the respective feature space chosen the algorithms in each case. This will later be compared in Section 3.1.5. In every case, the feature x is chosen as the most important feature (firstly chosen for forward algorithms and lastly deleted for backward algorithms). The local maxima feature, especially from the hand GSR is also popular.

Figures 3.14 and 3.15 present the two principal components after PCA transformation. They give a visual representation of the separability of the classes "not stress" (diamonds) and "stress" (+). In general, the values associated with "not stress" are closer to the origin, while the "stress" values tend to be more dispersed toward the edges, independently of which data set they belong to. A few "stress" samples tend to mix with the "non-stress" towards the center for each data set.

3.2.2 Classification

Table 3.2 is a comparison between the classification performances of different choices of feature spaces. The single and the multi-user modes are both tested, as well four classifiers: SVM, NB, KNN and PNN. The "intuitive choice" (i.c.) is the author's own choice, based upon Figures 3.7, 3.8 and 3.9 and the selection algorithms. It consists of:

1. ECG: mean of HR
2. Hand GSR: mean
3. feature x

The KNN fails however completely when using only one single principal component, falling down to a performance of 50.0 %, just like random guessing. This is related to that the notion of neighbour disappears when there is an input of only one principal component (a single observation of each feature, transformed by importance in terms of variance). Given this, the SVM, NB and the PNN are more trustworthy, especially the SVM which never falls below 78.9 % in performance. Naturally, the single-user case is a simpler problem than the multi-user one, due to differences between persons, which can be observed in the higher classification

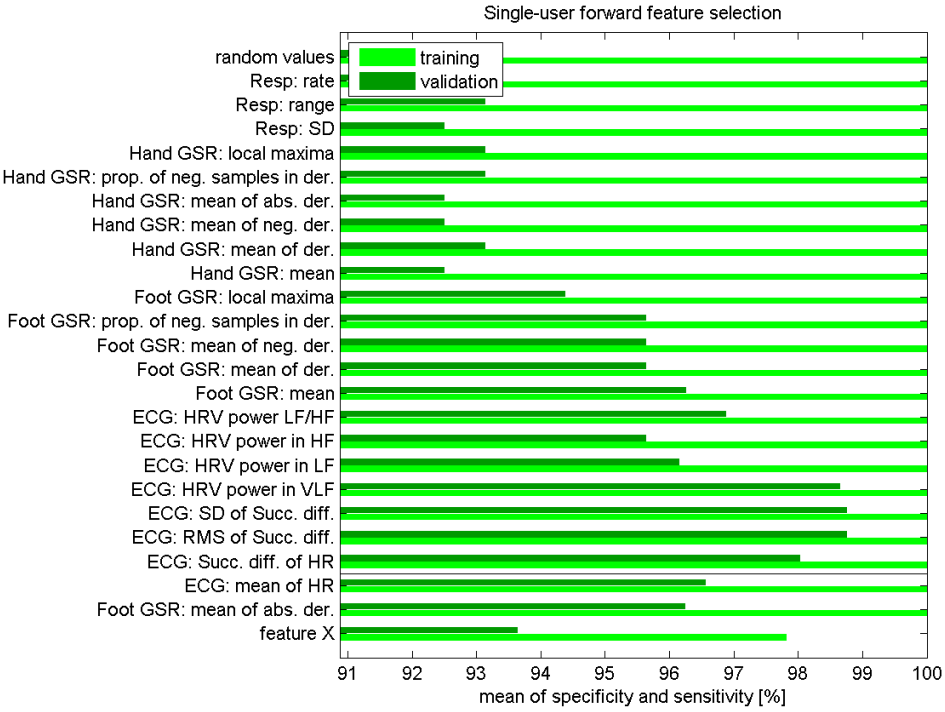

Figure 3.10: *Single-user forward feature selection. The feature space is growing upwards. Features are selected based upon the training data performance.*

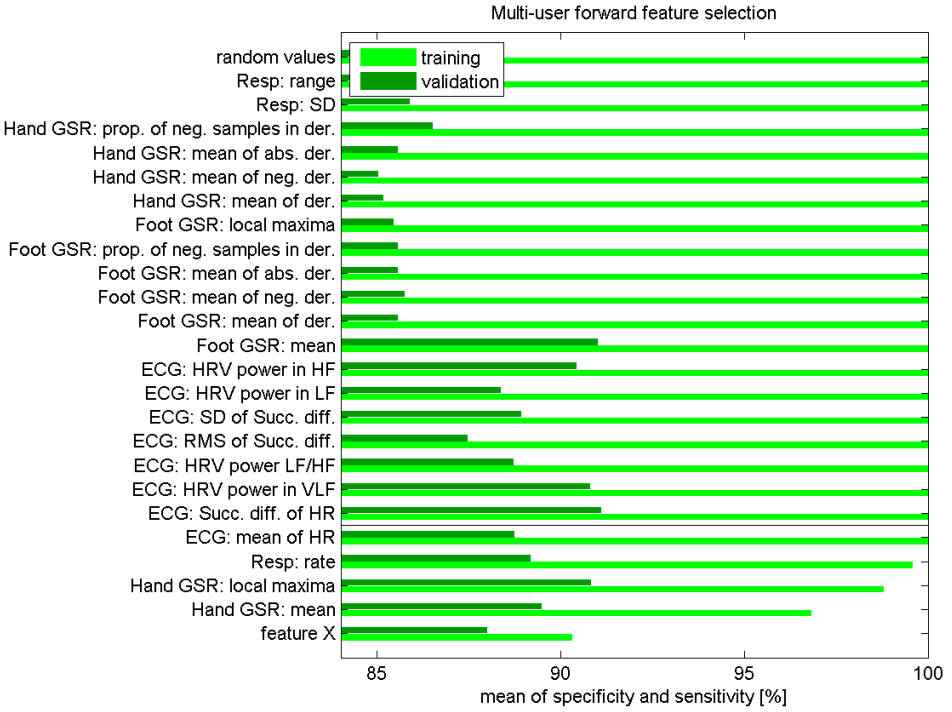

Figure 3.11: *Multi-user forward feature selection. The feature space is growing upwards. Features are selected based upon the training data performance.*

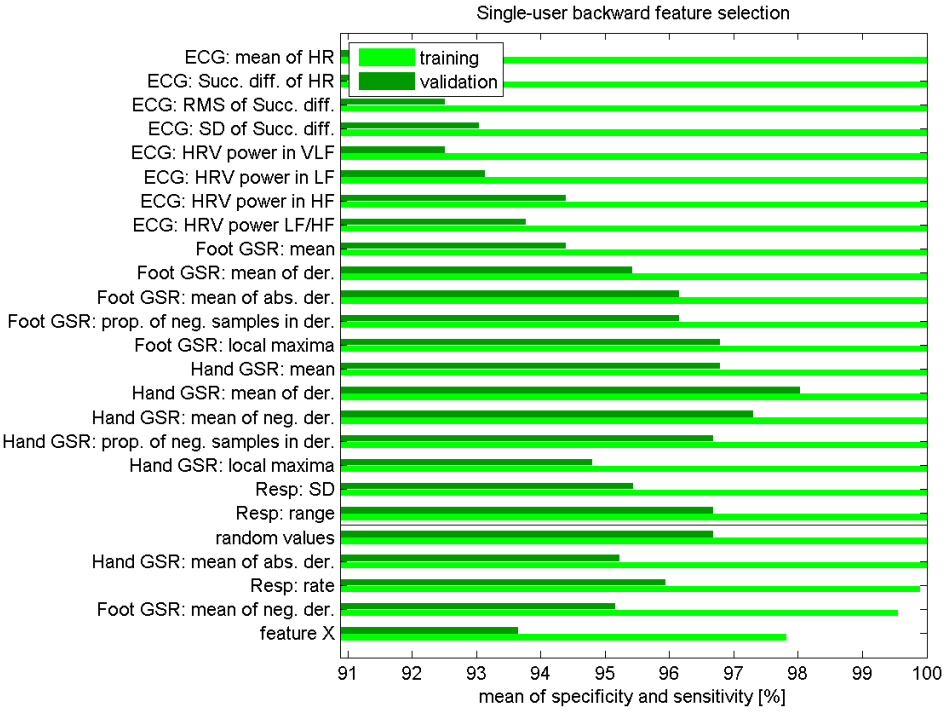

Figure 3.12: *Single-user backward feature selection. The feature space is growing upwards. Features are selected based upon the training data performance.*

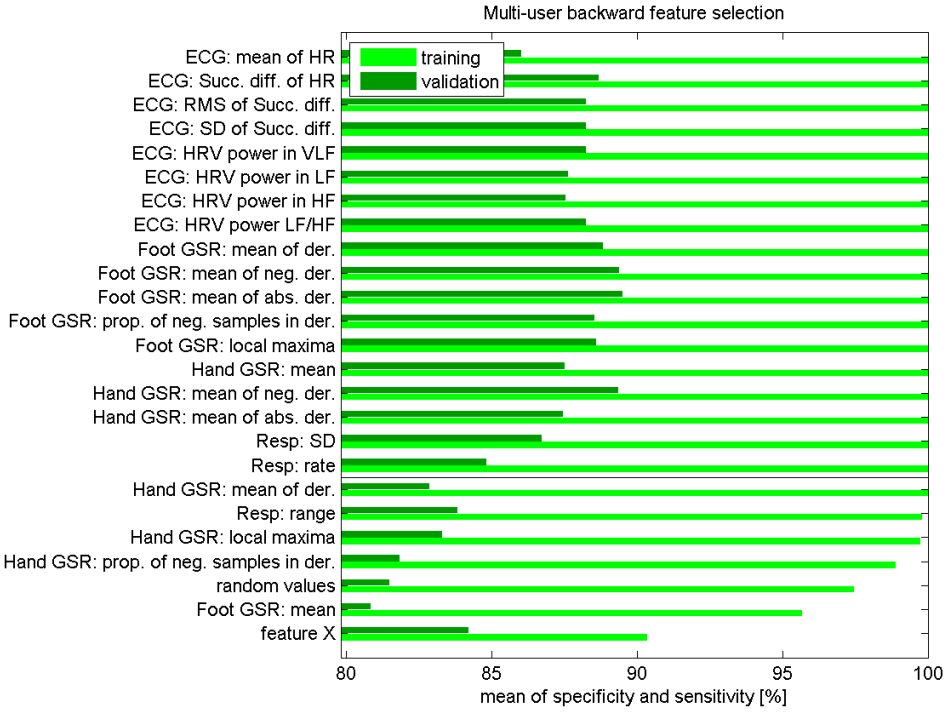

Figure 3.13: *Multi-user backward feature selection. The feature space is growing upwards. Features are selected based upon the training data performance.*

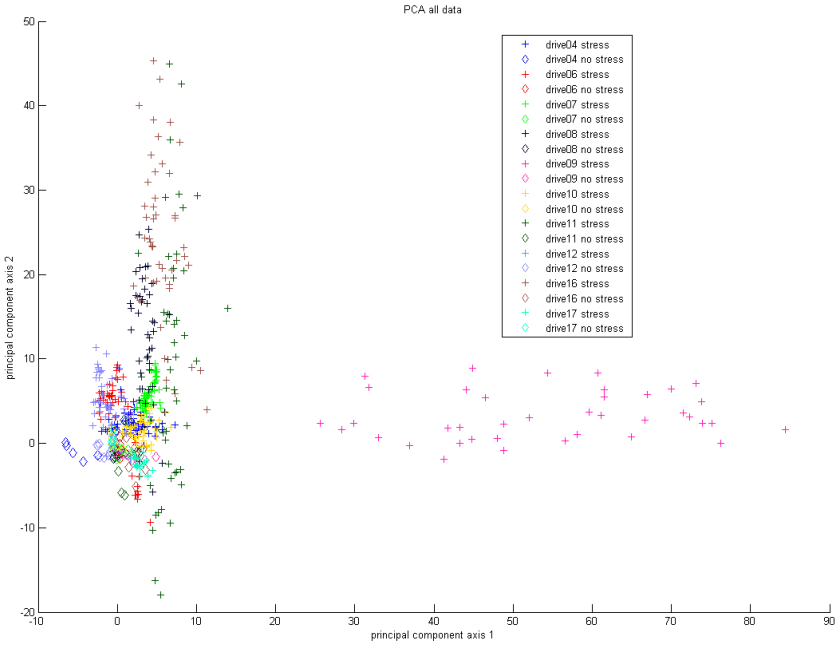

Figure 3.14: *The two principal components, all data sets. "stress class" data points are represented by a "+" and "stress" class is represented by diamonds.*

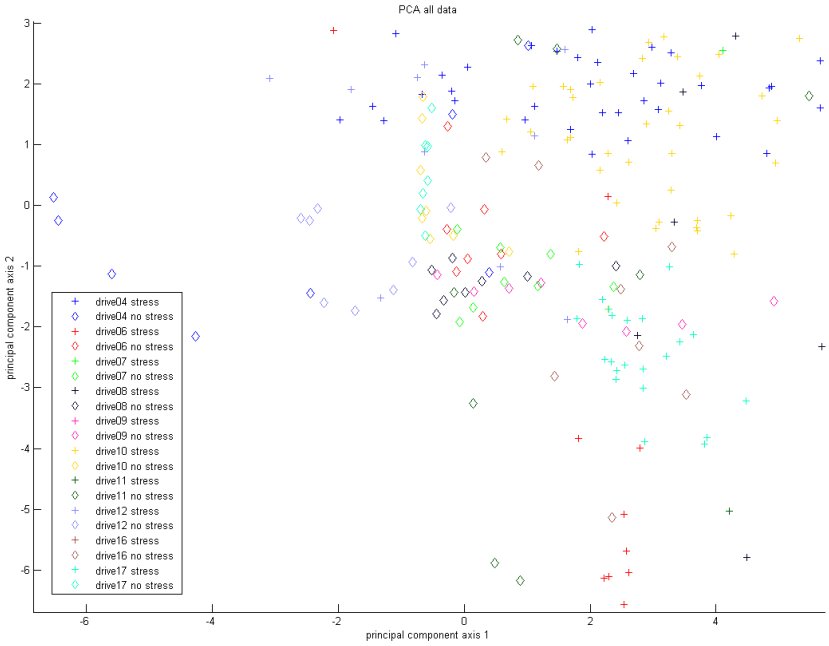

Figure 3.15: *The two principal components, all data sets. "stress class" data points are represented by a "+" and "stress" class is represented by diamonds. Magnified version of Figure 3.14.*

features	user	# feat.	SVM	NB	KNN	PNN
all	single	24	91.9 ± 2.3	93.1 ± 2.2	97.9 ± 1.2	96.5 ± 1.2
PCA		1 comp.	86.0 ± 3.0	91.8 ± 2.3	50.0 ± 4.3	85.3 ± 4.3
PCA		2 comp.	85.1 ± 3.0	95.5 ± 1.8	93.7 ± 2.1	92.8 ± 2.1
PCA		3 comp.	78.9 ± 3.5	95.1 ± 1.8	94.8 ± 1.9	91.5 ± 1.9
forward		3	96.1 ± 1.6	96.6 ± 1.6	98.2 ± 1.1	92.9 ± 1.1
backward		4	93.4 ± 2.1	94.4 ± 2.0	97.4 ± 1.4	93.7 ± 1.4
i.c.		3	96.8 ± 1.5	96.3 ± 1.6	99.5 ± 0.6	93.8 ± 0.6
all	multi	24	85.9 ± 3.0	87.9 ± 2.8	91.6 ± 2.4	90.4 ± 2.4
PCA		1 comp.	80.2 ± 3.4	65.3 ± 4.1	50.0 ± 4.3	78.5 ± 4.3
PCA		2 comp.	78.9 ± 3.5	80.0 ± 3.4	81.8 ± 3.3	84.3 ± 3.3
PCA		3 comp.	84.3 ± 3.1	81.6 ± 3.3	87.1 ± 2.9	86.3 ± 2.9
forward		5	88.3 ± 2.7	94.6 ± 1.9	93.1 ± 2.2	88.6 ± 2.2
backward		6	78.4 ± 3.5	83.5 ± 3.2	83.6 ± 3.2	86.3 ± 3.2
i.c.		3	90.7 ± 2.5	93.9 ± 2.0	90.7 ± 2.5	85.7 ± 2.5

Table 3.2: *MIT Driver database classification results. The performance is calculated as the mean of specificity and sensitivity, with margin of error E_{95}, both in %.*

rates. The PCA feature spaces gain a lot just by adding one or two more principal components. The "intuitive choice" space can compete quite well with the other more mathematical and algorithmic feature selection methods. The highest performance (99.5 %) is achieved by the KNN classifier in the single-user case using this intuitively chosen feature space. It is however important to note that the "intuitive choice" features are selected by a combination of the different score metrics and the selection algorithms, which are all calculated using all the data in different ways. This makes the results from this choice slightly misleading, since it is not a "true" validation.

3.3 Discussion

In this section the results of Section 3.2 and the methodology presented in Section 3.1 are discussed. The purpose is to highlight interesting findings compared to Chapter 2.

3.3.1 Results

Classic features such as means of HR and GSR levels appear as expected. A feature that was never inspired by the state of the art study in Chapter 2, but was rather derived by visual inspection of the data was the "feature x". It has high scores from Fisher, linear correlation and the t-Test analysis, as well as being generally well placed in both forward and backward feature selection. A feature

that appears less than expected is the LF/HF ratio, which is quite common in the literature.

In terms of feature space, the forward and backward algorithms perform badly in generalization, but a cause might be that their decisions are based upon the training performance, which in the single-user case reaches perfect prediction after just a couple of features. This might be a contributing cause to that the author's "intuitive choice" of three features generalizes better. An interesting extension of this work would be to make an exhaustive feature selection of three features. This would mean trying all possible combinations of three features, to find the best ones. A hypothesis is that this would result in the three features from the "intuitive choice". As expected, the algorithms choose more features in the more complicated, multi-user mode (5 in forward and 6 in backward) compared to the single-user mode (3 in forward and 4 in backward). Figure 3.12 shows a large difference between the training data results and the validation data results of the mean of ECG feature. It is chosen last based upon the training data, but the validation result (dark green) shows that it actually reaches a classification rate closer to 100 %. Since the feature selection choices are based upon the training data, this gives a misleading feature space. However the option is to base the feature selection upon the validation, but this would not leave any data from the comparing the classifiers, done in Table 3.2.

Quite high classification rates are achieved compared to what is found in Table 2.4. The reason for this might be that we are actually distinguishing between driving and rest in this study, which is not necessarily the same thing as distinguishing between "stress" and "non-stress". There is no data available to a corresponding stress free driving task (control task), which allows the classifier to adapt to drive-specific features rather than stress-specific features if such exist. It is also important to note that these previous studies are based upon different sensors and data, however the SVM performs generally well compared to other classifiers. Slightly surprising is the result of achieving 99.5 % classification rate with the KNN using three features from three different sensors in single-user mode. This is both a simple classifier and few features, which is promising for real-time diagnostics, however the KNN can be memory-consuming. The KNN also enjoys a great improvement from each additional principal component, and an interesting continuation would be to add even more of them to maximize its potential.

[20], who also analyzes the MIT database, has more details on the database than the public version studied in this Master's thesis. Thus it has the possibility of defining three classes: low, medium and high stress. Using 22 features and LDA, these classes turn out having a recognition rate of 100 %, 94.7 %, and 97.4 % respectively. This is slightly higher than what is achieved in Section 3.2.2.

Another point to consider is that all the data used is down-sampled to 15.5 Hz, which is quite low for the physiological signals studied. This can of course influence the reliability of these results.

3.3.2 Method

An important methodological topic to discuss is the fact that the "rest" and "driving" states have been directly converted into "non-stress" and "stress" class. With access only to raw signal data and no questionnaires, this is however the most reasonable procedure. To deduce that driving is stressful makes sense due to the increased need for attention and reactiveness while in traffic, along with pressure from other drivers and time. But without having the answers whether the subjects felt stressed or not this step is a bit risky.

For a study with even higher validity, the data should first of all be available in its original sampling frequencies, along with information of whether the driver is on the highway or in the city. Optimally, a control task should be performed as well, e.g. where the driver is simply driving calmly in an open area free of other vehicles and pressure.

The choice of the time window T_f is also delicate, since it is a compromise between the data available and the features to be calculated. It is hard to know whether the choice is the best or not, and due to it being a very early choice it might influence all the steps following it. For example, [20] sets T_f to 5 minutes in one analysis and to 1 second in another one.

An alternative to the filtering methods described in Section 3.1.1, which find artifact data points and replaces them with the preceding value, could be integrating this preprocessing into the machine learning methods. If the artifacts are more common (or even unique) in a certain class, this could even be helpful for the classification problem. There is however the risk of overfitting the models to certain sensors or sensor models (if the artifacts are more common and distinctive to them). This would make the model good for the sensor used to train it, but not generalizable to other systems.

The choice of the performance could be made in other ways, which could cause the forward and backward feature selection algorithms to take different decisions. Examples include choosing the accuracy or the precision as performance measure. The g_{mean} could be interesting to try as an option to duplicating samples for class imbalance.

Other classification techniques that would be interesting to examine include the decision trees (Section 2.6.2), for their see-through and visualization possibilities. Different versions of the neural network classifiers than the PNN could also be interesting. Another classifier that appears in the literature is the nearest centroid classifier, which reminds of the simpler algorithms, such as the NB and the KNN. The LDA, used in [20], is of course interesting, since it based upon the same data, and allows a closer comparison. Of course the tuning parameters of the classification can be altered even further, mainly for the SVM, the PNN and the KNN. Different cost parameters, Gaussian widths and number of neighbors could be tempered with even more for optimizing the models.

Ideas for further feature analysis include more methods that extract feature sim-

ilarly to the PCA, e.g. independent component analysis or partial least squares. PCA can also be modified and tried in its multi-linear or kernel versions.

Furthermore, additional features could be added and compared, e.g. temporal and rhythmic measures of the HR, and slower features of the GSR, e.g. its rise time, latency and related metrics. This would require a finer GSR signal than what is found in the MIT Driver Database, where it is available with a sampling rate of 15.5 Hz.

3.3.3 Further Perspectives

Affective signal processing in general and analyzing the stress of drivers specifically is becoming an increasingly important subject in the automotive industry. Applications include preventing accidents, e.g. by soothing interventions like calm music. Cars are becoming more and more computerized, and with the right set-up of effective sensors and algorithms this could become standard in future vehicles.

Ethical issues that might occur is that physiological sensors and computing of affective states might be imposing for certain persons. One must also keep in mind to protect private physiological data from being spread out publicly against the person's wishes. For public databases, this data should be coded, and never be tagged using details that might reveal the identity of the subject.

4

Experiments

As a second study, laboratory stress experiments were performed in order to acquire more data for comparison with the results on the MIT driver database presented in Section 3.2.

4.1 Method

Between May and July 2015 an experimental campaign was performed, with the purpose of collecting physiological signals for stress analysis.

4.1.1 Experiment Procedure

9 subjects were recruited and asked to come four mornings: three times for different tasks and once for performing similar tasks without stress elements (control tasks). Their age varied in the range of 21 to 23 years, and among them were 7 females and 2 males.

Three classical stress tasks were chosen: the TSST, the MST and the SECPT, along with one task not usually found in the literature: the d2 test. A summary of these stress tasks can be found in Table 4.1. The task, its control task and the associated types of stress are presented. The d2 task is not classically used for inducing stress in the literature, however the other three (TSST, MST and SECPT) are all common. All of these were attributed a corresponding control task, supposed to be identical but without stress elements. For the TSST, the control task consists of reading a text out loud followed by simply counting downwards from 1,000 by steps of one. These two control tasks lasted 5 minutes each, and were both performed by the subject alone in the experiment room, without any video recording. The SECPT control task meant submerging the hand in lukewarm in-

test	stress task	control task	stress type
TSST	job interview (with video recording and jury)	read text out loud alone	• social
MST	count backward steps of 13 from 1,022 (with video recording and jury)	count backwards steps of 1 from 100 alone	• social • mental
SECPT	put hand in 0 °C water (with video recording)	put hand in 37 °C water	• social • physical
d2	find all d:s with two lines among other symbols under time pressure	find all d:s with two lines among other symbols	• mental • time pressure

Table 4.1: *The tasks used to induce the stress during the experiments, and their respective control tasks. The type of stress is also indicated.*

stead of freezing water, also without any video recording. The d2 control task was identical to the d2 task, but without any time pressure. The subject is asked to simply find the symbols, taking his or her time.

Upon arrival to the first session, the subject was asked to sign a consent document. At the start of each experiment, a brief explanation of the experiment was given, without revealing the true purpose of inducing stress. Subsequently, the subject was equipped with the sensors and their signals were controlled.

A TSST session was performed according to:

1. rest (ca 10 minutes)
2. questionnaires (ca 2 minutes)
3. explaining presentation task
4. subject prepares presentation task (ca 10 minutes)
5. presentation task (5 minutes)
6. explaining MST
7. performing MST (5 minutes)
8. questionnaires
9. rest in waiting room - saliva samples but no signals recorded (30 minutes)

A d2 session was performed according to:

1. rest (ca 10 minutes)

2. questionnaires (ca 2 minutes)
3. explaining d2 task
4. performing d2 task (4 minutes 40 seconds)
5. questionnaires
6. rest in waiting room - saliva samples but no signals recorded (30 minutes)

An SECPT session was performed according to:

1. rest (ca 10 minutes)
2. questionnaires (ca 2 minutes)
3. explaining submersion task
4. submerging hand (max. 3 minutes)
5. if the hand is still submerged after 1 minute - questionnaires
6. questionnaires
7. rest in waiting room - saliva samples but no signals recorded (30 minutes)

A control task session can be summarized as:

1. rest (ca 10 minutes)
2. questionnaires (ca 2 minutes)
3. explaining presentation control task
4. presentation control task - reading text alone (5 minutes)
5. explaining MST control task
6. MST control task - counting backwards alone (5 minutes)
7. rest (ca 10 minutes)
8. questionnaires (ca 2 minutes)
9. explaining d2 control task
10. d2 control task (no time limit)
11. rest (ca 10 minutes)
12. questionnaires (ca 2 minutes)
13. explaining task SECPT control task
14. SECPT control task (max. 3 minutes)
15. if the hand is still submerged after 1 minute - questionnaires
16. questionnaires

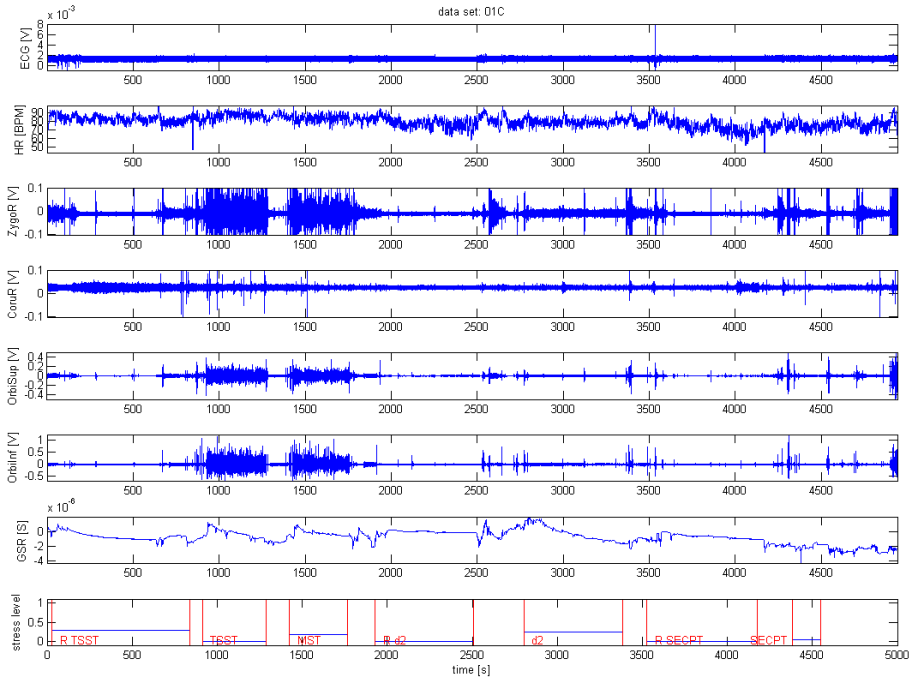

Figure 4.1: *The control tasks for subject 01. "R task" means rest period before a task.*

The questionnaires filled at each session were VAS and STAI-Y on various emotions perceived by the subject during the rest period or task preceding the questionnaire. During the SECPT and its control task, a PANAS questionnaire on the perceived pain level was also distributed. During the first session of each subject, questionnaires on personality type and more were filled out. The order of the tasks within a control session was varied between the subjects, as well as the order of the stress task sessions.

In Figure 4.1 the raw data recorded during the control tasks of subject can be observed. The stress level is derived from the answer at the VAS questionnaire at the end of the rest period or the task.

4.1.2 Sensors

During the experiments, three types of physiological signals were recorded: ECG, GSR, and facial EMG. An overview of the used systems and signals can be found in Figure 4.2.

For recording ECG, AD Instruments Disposable ECG Electrodes were used, and

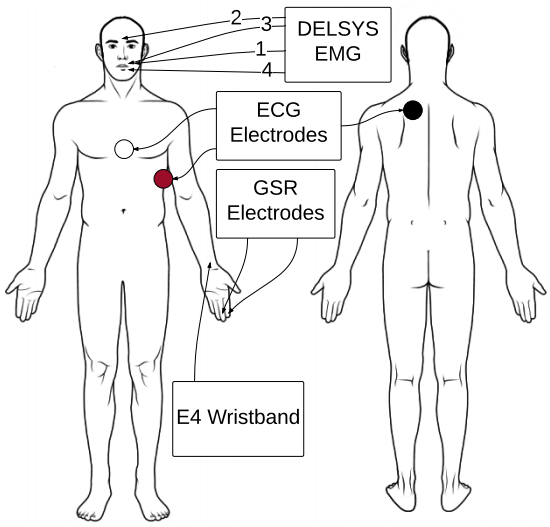

Figure 4.2: *Overview of the used sensors used and the physiological recorded signals. The EMG signals: 1. zygomaticus major 2. currogator supercilii 3. orbicularis oris superior 4. orbicularis oris inferior.*

connected to the Ad Instruments FE132 Bio Amp. Figure 2.2 shows the electrodes used for recording ECG. The white one (to the left) was placed in the middle of the chest, the red one (in the middle) on the side below the chest (to center the heart between the white and the red electrodes). The black one (to the right) was placed upon a flat area on the back, as ground reference.

The AD Instruments GSR Finger Electrodes MLT116F were used for recording GSR. Figure 2.4 shows these electrodes, along with their placement. They were placed on the non-dominant hand of the subject, on the index finger and the middle finger.

An Empatica E4 wristband was ordered and arrived during the middle of the experiment phase. It was subsequently used on all remaining subjects for recording heart rate and inter-beat intervals (both via photoplethysmogram), GSR, skin temperature and 3-axis accelerometer. The placement of the wristband, along with the GSR electrodes, can be studied in Figure 2.4.

For measuring EMG, the DELSYS Trigno Wireless EMG System [24] was used. It was placed at four different locations on the face of the subject, Figure 4.2:

- zygomaticus major muscle, related to smiling
- currogator supercilii muscle, related to frowning
- orbicularis oris superior muscle

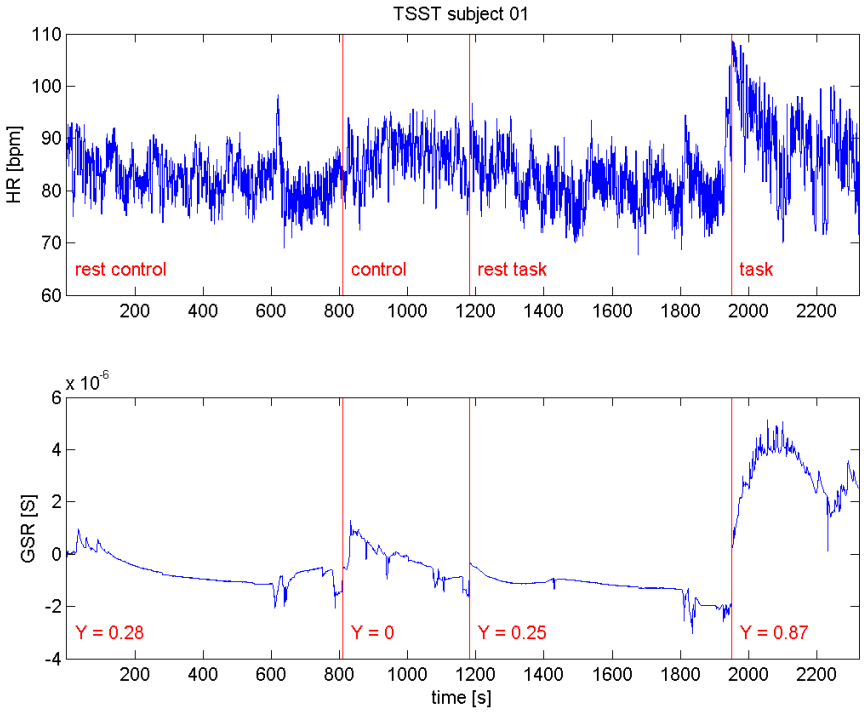

Figure 4.3: *The rest periods and tasks related to the TSST of subject 01. The questionnaire stress level for each period is indicated by "Y" (0-1).*

- orbicularis oris inferior muscle

In Table 4.2 the recorded experimental database is presented. The E4 wristband is missing in some cases since it was not available before the middle of the experimental phase, and the GSR sensor disconnected during the d2 task of subject 03 and the TSST task of subject 05.

Figure 4.3 shows an example of the data collected during a TSST experiment.

4.1.3 Preprocessing

After inspection of the acquired data, it was found that the HR values calculated from the ECG sometimes resulted in "not a number" (NaN). As soon as such a value was found, it was replaced by the preceding HR value. Furthermore, artifacts in the form of unreasonably low heart rates (< 25 bpm) were found in some cases. To deal with this, a filter finding heart rates lower than 3 times the mean value of the task or the rest period was programmed. If this occurred, the heart rate value of the preceding sample was used instead.

subject	experiment	ECG	HR	GSR	E4 wristband
01	control	x	x	x	
	d2	x	x	x	
	SECPT	x	x	x	x
	TSST	x	x	x	
02	control	x	x	x	x
	d2	x	x	x	x
	SECPT	x	x	x	x
	TSST	x	x	x	
03	control	x	x	x	
	d2	x	x	!	
	SECPT *	x	x	x	
	TSST	x	x	x	x
04	control	x	x	x	
	d2	x	x	x	x
	SECPT	x	x	x	
	TSST	x	x	x	x
05	control	x	x	x	
	d2	x	x	x	
	SECPT	x	x	x	
	TSST	x	x	!	x
06	control	x	x	x	
	d2	x	x	x	x
	SECPT *	x	x	x	
	TSST	x	x	x	x
07	control	x	x	x	x
	d2	x	x	x	x
	SECPT	x	x	x	x
	TSST	x	x	x	
08	control	x	x	x	
	d2	x	x	x	x
	SECPT	x	x	x	
	TSST	x	x	x	
09	control	x	x	x	x
	d2	x	x	x	x
	SECPT	x	x	x	x
	TSST	x	x	x	x

Table 4.2: *The experimental database. The signals successfully recorded in each experiment are indicated by "x". "!" indicates a sensor problem occurred during the data acquisition. * subject 03 and subject 06 both removed the hand from the cold water after around 30 seconds, which just enough for one complete sample for the feature calculation (T_f = 30 s). The SECPT for subject 03 and subject 06 were thus not possible to analyze.*

For the experimental data a time window $T_f = 30$ s was chosen, since the SECPT sometimes only lasts about 30 seconds to a minute (the subject has the right to remove her or his hand prematurely if the pain is too high). The remaining tasks (d2, MST and TSST) all last for around 5 minutes, which results in around $\frac{5 \cdot 60\,\text{s}}{30\,\text{s}} \approx 10$ samples for each task and subject.

The experiments are designed to produce data with balance between classes (since the stress task and control are supposed to last the same amount of the time, with the d2 task as an exception). To ensure class balance, the duplication described in Section 3.1.3 was used.

4.1.4 Features

Based upon the HR and hand GSR signals, the following features were chosen:

From the ECG and HR signals:

1. mean of HR
2. successive differences of HR
3. root mean square of successive differences
4. standard deviation of successive differences
5. HRV power in LF (0.04 – 0.15 Hz)
6. HRV power in HF (0.15 – 0.50 Hz)
7. ratio between HRV power in LF and HF

The HRV power in VLF (0.01 – 0.04 Hz) feature was excluded due to shorter time window T_f, which does not offer enough samples for VLF calculations. Neither did this feature prove itself important in Chapter 3.

From the GSR signals measured at the hand:

1. mean
2. mean of derivative
3. mean of negative derivative
4. mean of absolute derivative
5. proportion of negative samples in the derivative vs all samples
6. number of local maxima

These features add up to a total of 13. As in Section 3.1.2, a 14th feature, "random values", consisting of random values between 0 and 1, was also added for comparison purposes.

After computation, each feature was normalized using (3.10). The window for normalization was taken from the rest period preceding each control or stress task.

4.1.5 Comparison: Laboratory Equipment Versus E4 Wristband

A comparison was made to get an understanding of how well a wearable sensor (such as the Empatica E4 wristband) performs compared to stationary laboratory equipment regarding HR and GSR signal recording. To do this, synchronized signals from the E4 wristband and from the laboratory sensors were superimposed and compared.

For the experiments with subject 02, the E4 wristband was available for 3 sessions: control tasks, SECPT and the d2 task. To compare the laboratory equipment and the E4 wristband, a classifier was trained using the signals acquired by the laboratory equipment, while validating on the signals acquired by the E4 wristband. Note that this is a difficult task for a classifier, due to the training and the validation data originating from different sources.

4.1.6 Comparison: Control Task Versus Task

To discover changes in the physiological signals between stressful tasks and stress-free tasks, each task was compared to its corresponding stress task. The method used is to consider the control task as one class and the stress task as another class, and analyze how separable these two classes are, and in that case what signal features that are most relevant for this separability.

4.1.7 Comparison: Different Stress Tasks

We ask ourselves "can these tasks all be considered to induce the same kind of stress, do the same physiological changes appear?"

For comparing the different stress tasks between each other, a classifier was trained using three of the four tasks, while the remaining one was used for validation. The idea is that if the validation task indeed induces similar physiological reactions as the others, it should be possible to predict it quite well.

4.1.8 Continuous Stress Models

In order to also take into account the perceived stress levels given by the subjects in the questionnaires, (which are not available in Chapter 3), continuous models with the purpose of predicting the stress level were made. The VAS values Y were transformed to the interval $[0, 1]$. As earlier, the rest period preceding each task was used for normalization, and this was also done for the stress levels. This means that the relative stress values all appear on the interval $[-1, 1]$. All predictions outside this interval were transformed to -1 (for predictions ≤ -1) or 1 (for predictions ≥ 1).

Four studies were made:

1. linear correlations between features and the stress level, (2.1)
2. linear regressive model, (Section 2.7.1)
3. support vector regression, (Section 2.7.2)

4. variational multiple Bayesian linear regression, (Section 2.7.3)

The linear correlations were calculated using (2.1). This allows distinction between different tasks, which are not necessarily perceived as equally stressful by the subjects. It also gives a hint about how correlated each feature is for changes in stress level. For this calculation, all the four stress tasks were included, and the linear correlation of their features were calculated against the stress level from the corresponding questionnaire.

The variational multiple Bayesian linear regression model uses the software from [4]. This implementation does not transform the data with a basis function, thus $\Phi(X) = X$. The prior parameters were set as $a_0 = 2$, $b_0 = 0.2$, $c_0 = 10$ and $d_0 = 1$, the default settings of the software.

The support vector regressive model was calculated using LIBSVM:s ν-SVR function, with an RBF kernel. The cost parameter c was set to 1 and the ν parameter to 0.5.

4.2 Results

The results from the analysis made in Section 4.1 are presented here.

4.2.1 Comparison: Laboratory Equipment Versus E4 Wristband

The E4 wristband was used in as many data acquisitions as possible. Figures 4.4 and 4.5 compare the derived HR and the GSR signal recorded from the E4 wristband to the signals of the laboratory equipment. The E4 HR is basically a filtered version of the laboratory equipment's signal, with lower sampling rate. The GSR is harder to compare, since it is measured at the wrist and the fingers, respectively.

The classification experiment with the purpose of determining whether a model could be trained with the laboratory equipment sensors and validated using the E4 wristband was done for both the SECPT and the d2 tasks, Table 4.3. The d2 task is not possible to predict better than pure guessing, however the SVM succeeds to distinguish the SECPT from its control task quite well (91.7 % performance). This is probably related to the fact that the SECPT induces clearer and more similar changes in the subjects than the d2 task. However, due to lack of data acquisitions with all the sensors, including the E4 wristband, this is only done for a two single data sets. This makes any stronger conclusions hard to draw on this particular experiment. The 6.3 % result from the naive Bayes classifier is odd, it means that inverting the model would result in 93.7 % accuracy! With the margin of errors at around ± 20 % these result are hard to draw conclusions upon. A possible explanation for these varying results can be found in Figures 4.4 and 4.5, which show that the output signals of two sensor systems do not really correspond very well to each other, and can not be used interchangeably in this way. Compared to the literature, this method of model validation is very rare, however finding a model that works well even for different sensors would be interesting.

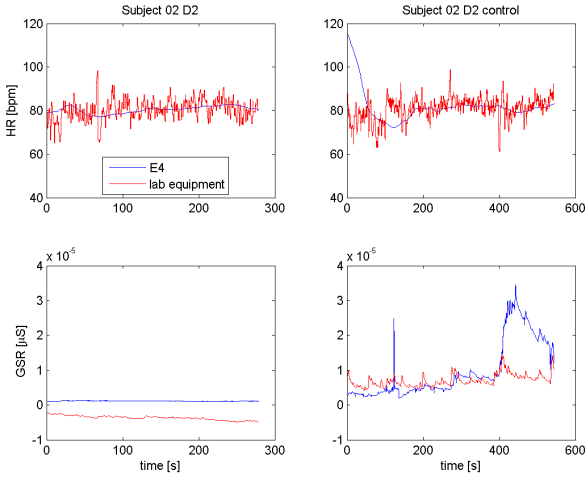

Figure 4.4: *A comparison of the E4 wristband signals and the corresponding signals recorded by the laboratory equipment. The data is from the d2 of subject 02.*

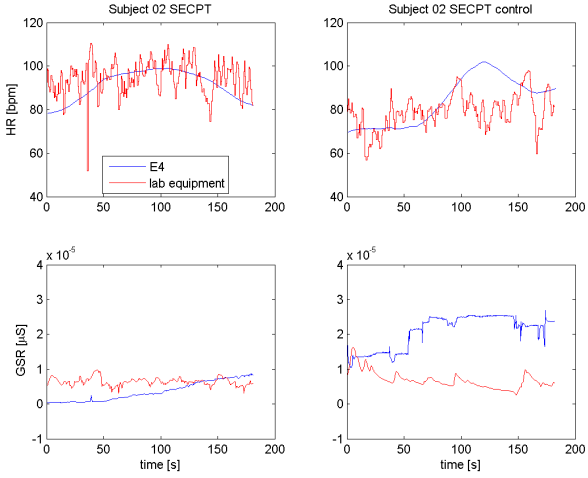

Figure 4.5: *A comparison of the E4 wristband signals and the corresponding signals recorded by the laboratory equipment. The data is from the SECPT of subject 02.*

stress task	SVM	NB	KNN	PNN
d2	44.1 ± 19.5	6.3 ± 9.5	53.7 ± 19.6	41.2 ± 19.3
SECPT	91.7 ± 16.3	45.0 ± 29.4	70.0 ± 27.1	70.0 ± 27.1

Table 4.3: *Different classifiers trying to separate a stress task from its control task. The laboratory equipment signals are used for training and the E4 wristband is used for validation. The performance is calculated as the mean of specificity and sensitivity, with margin of error E_{95}, both in %.*

task	user	# feat.	SVM	NB	KNN	PNN
d2	single	13	90.9 ± 4.1	94.0 ± 3.4	86.5 ± 4.9	88.0 ± 4.7
d2	multi	13	38.4 ± 7.0	61.7 ± 7.0	44.9 ± 7.1	39.0 ± 7.0
MST	single	13	82.8 ± 3.8	92.1 ± 2.7	82.4 ± 3.9	79.3 ± 4.1
MST	multi	13	74.4 ± 4.4	59.3 ± 5.0	63.3 ± 4.9	59.8 ± 5.0
TSST	single	13	91.4 ± 2.4	95.0 ± 1.8	86.5 ± 2.9	86.8 ± 2.9
TSST	multi	13	80.3 ± 3.4	63.6 ± 4.1	70.6 ± 3.9	63.7 ± 4.1
SECPT	single	13	79.7 ± 3.2	80.4 ± 3.1	76.0 ± 3.4	71.0 ± 3.6
SECPT	multi	13	66.4 ± 3.7	60.8 ± 3.9	58.5 ± 3.9	65.3 ± 3.8

Table 4.4: *Classifier performance comparison between different stress task versus their respective control task. The performance is calculated as the mean of specificity and sensitivity, with margin of error E_{95}, both in %.*

4.2.2 Comparison: Control Task Versus Task

Based upon the results in Section 3.2.2, an SVM classifier with a radial basis function was chosen, along with NB, KNN and PNN classifiers, tuned as in Section 3.1.5. The classifiers were trained for each stress task, with the purpose of distinguishing between the stress task and its corresponding control task for each subject. The results are presented in Table 4.4. As in Section 3.2.2, the single-user case is easier to predict than the multi-user case. The most distinguishable tasks for both cases is the MST, the TSST and the SECPT. The d2 is however harder to distinguish, especially in the multi-user case. Identically to the results in Section 3.2.2, the SVM proves to be the most accurate classifier over different tasks and user cases, but not by great margins.

In Figures 4.6, 4.7, 4.8, and 4.9 the t-Test (2.2) has been performed between the stress tasks (d2, MST, SECPT, TSST) and their respective control tasks, all after normalization using the rest period. The d2 task, Figure 4.6, shows very different reactions for different subjects. The outliers for subject 04 (with $T < -25$) in the HRV power features in LF and HF and the successive differences only appear for the d2 test. For example, some experience a decreased mean HR while it increases for other persons. The MST, Figure 4.7, shows more tendencies, with mean of HR

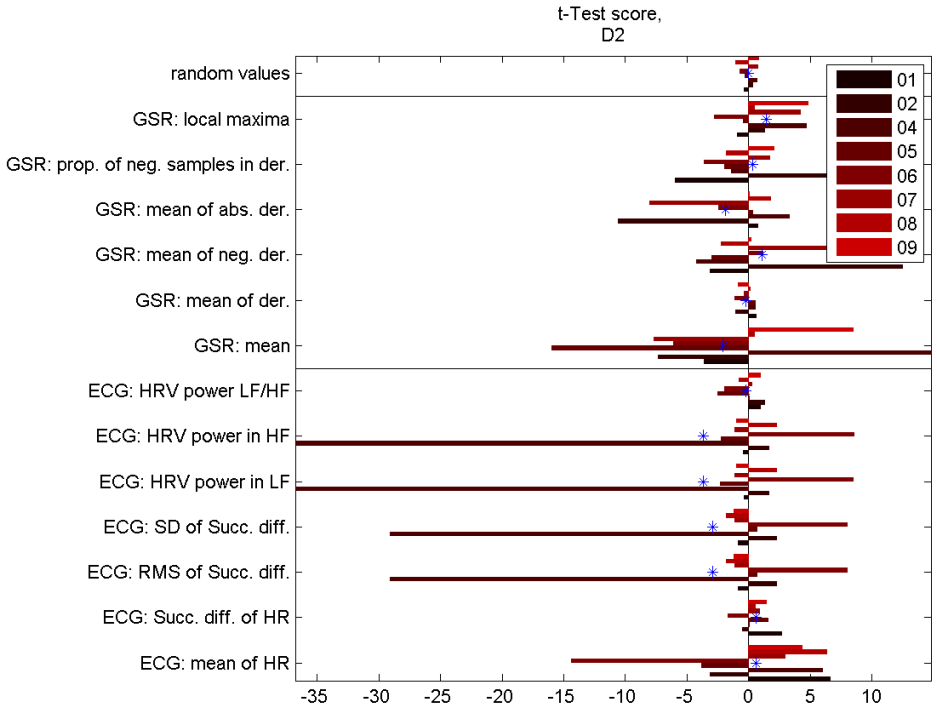

Figure 4.6: *The t-Test scores for each feature for the d2 control task class versus the d2 task. Each subject is presented, and the mean of all subjects is indicated by "*".*

and GSR increasing for almost all subjects, along with the number of local maxima. The SECPT and the TSST, (Figures 4.8 and 4.9) show even more significant changes, mainly in the same features but also in the successive differences of HR.

4.2.3 Comparison: Different Stress Tasks

In Table 4.5 the results of the classification experiment described in Section 4.1.7 are presented. Most classifiers have trouble predicting the d2 task using the other tasks, while e.g. the MST and the SECPT has greater prediction results, staying above the random prediction result of 50 %, even when subtracting their lower margin of error. This might indicate that the physiological reactions of the d2 test is not comparable to the ones of the other tests (i.e. it not being stressful enough or not inducing stress of a comparable kind). Due to a low number of samples from the data (around 10 for each class) and low classification results in general, the margins of error are quite high (15 - 25 %). This is somewhat similar to Table 4.3, which has problems of the same kind. This is also a difficult task for

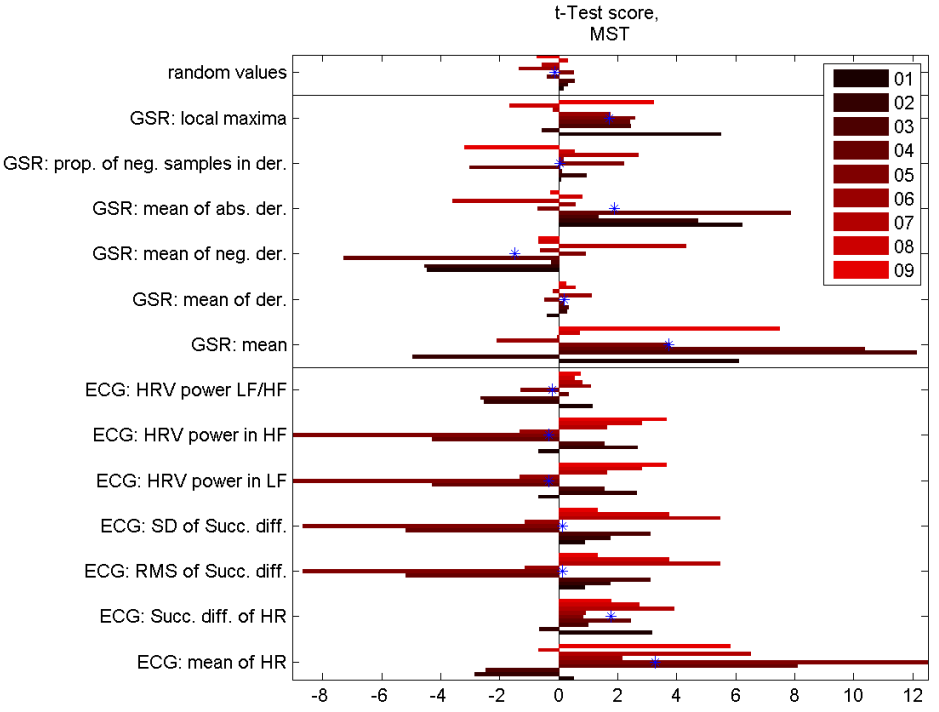

Figure 4.7: *The t-Test scores for each feature for the* MST *control task class versus the* MST *task. Each subject is presented, and the mean of all subjects is indicated by "*".*

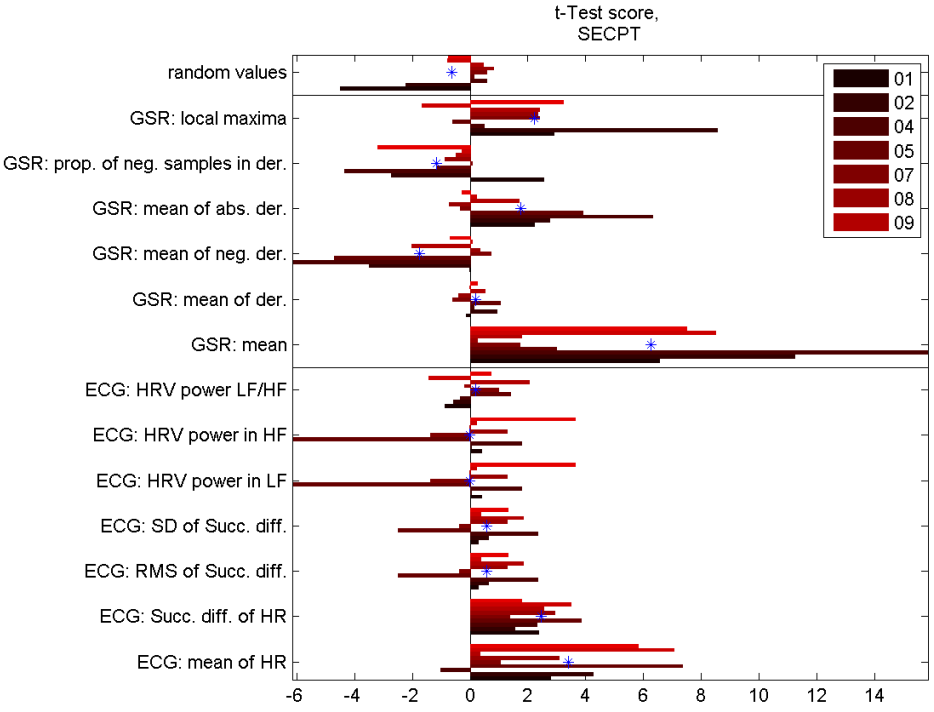

Figure 4.8: *The t-Test scores for each feature for the* SECPT *control task class versus the* SECPT *task. Each subject is presented, and the mean of all subjects is indicated by "*".*

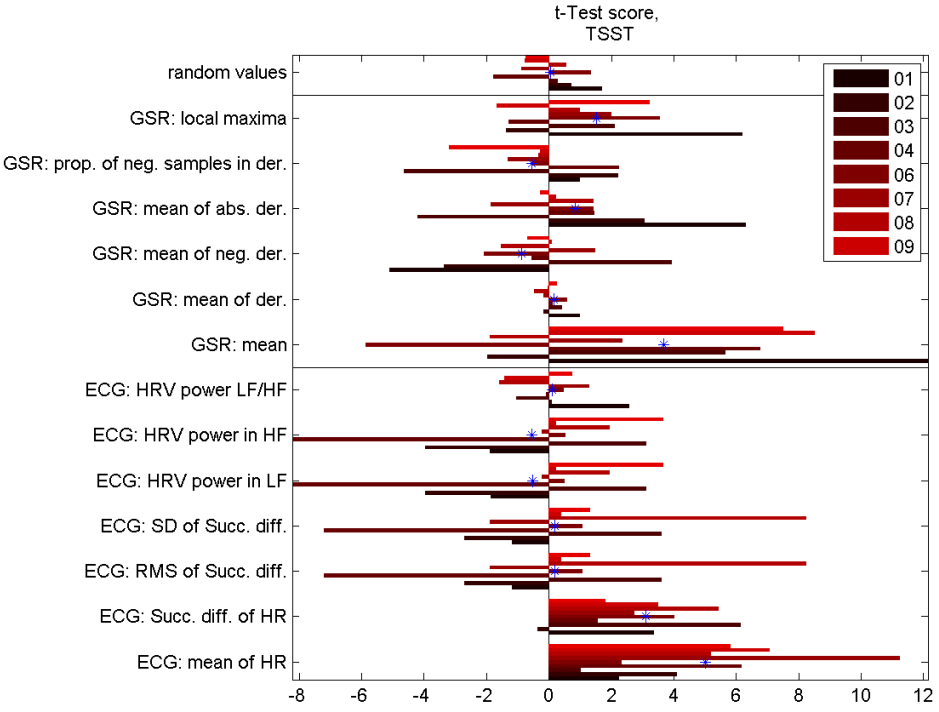

Figure 4.9: *The t-Test scores for each feature for the TSST control task class versus the TSST task. Each subject is presented, and the mean of all subjects is indicated by "*".*

validation task	SVM	NB	KNN	PNN
d2	57.1 ± 14.3	65.6 ± 14.6	53.4 ± 17.1	54.4 ± 15.1
MST	63.7 ± 17.0	86.1 ± 9.1	62.3 ± 18.3	54.8 ± 19.5
SECPT	72.1 ± 15.4	76.5 ± 15.3	72.3 ± 16.8	67.0 ± 17.0
TSST	65.5 ± 22.2	57.1 ± 26.9	57.1 ± 23.3	51.2 ± 26.5

Table 4.5: *The mean classifier performances over all subjects when using one stress task as validation data and the remaining three as training data. The performance is calculated as the mean of specificity and sensitivity, with margin of error E_{95}, both in %.*

task	linear RMS error	SVR RMS error	VBML RMS error
d2	0.1211	0.0501	0.1273
TSST	0.0696	0.0329	0.0661
MST	0.1908	0.1185	0.1773
SECPT	0.0951	0.0214	0.0982

Table 4.6: *The RMS errors of different regressive models predicting the stress levels given by the subjects in the questionnaires. The task and its control task are used in each data, using cross validation.*

classifiers, since it treats the signals from all experiments in the same way, while the t-Test from Figures 4.6, 4.7, 4.8, and 4.9 shows different results, especially in d2 task. This further suggests that the d2 task is inducing different or weaker reactions in the subjects compared to the other stress tasks.

4.2.4 Continuous Stress Model

The linear correlations between features and the stress level can be studied in Figure 4.10. This time, the number of local maxima in the GSR shows to be quite correlated with the stress level. Also the mean of GSR and HR can be found, as found previously in Section 3.2.1.

In Table 4.6 the results of a leave-out-one cross validation of the predicted continuous stress level is presented. In each validation, a stress task and its control task are included for one user. The three models (linear, SVR and VBML) then try to predict the stress value related to each sample using the other samples to learn the model. The SVR model has a smaller RMS error than the linear and the VBML ones, which are similar in performance. The SECPT, MST and d2 are the easiest ones to predict, followed by the TSST, which has a greater error.

In Figure 4.11 the results from a cross validation within each user and task are presented. The three continuous models are compared, where the SVR is more accurate and the linear and the VBML models follow each other closely.

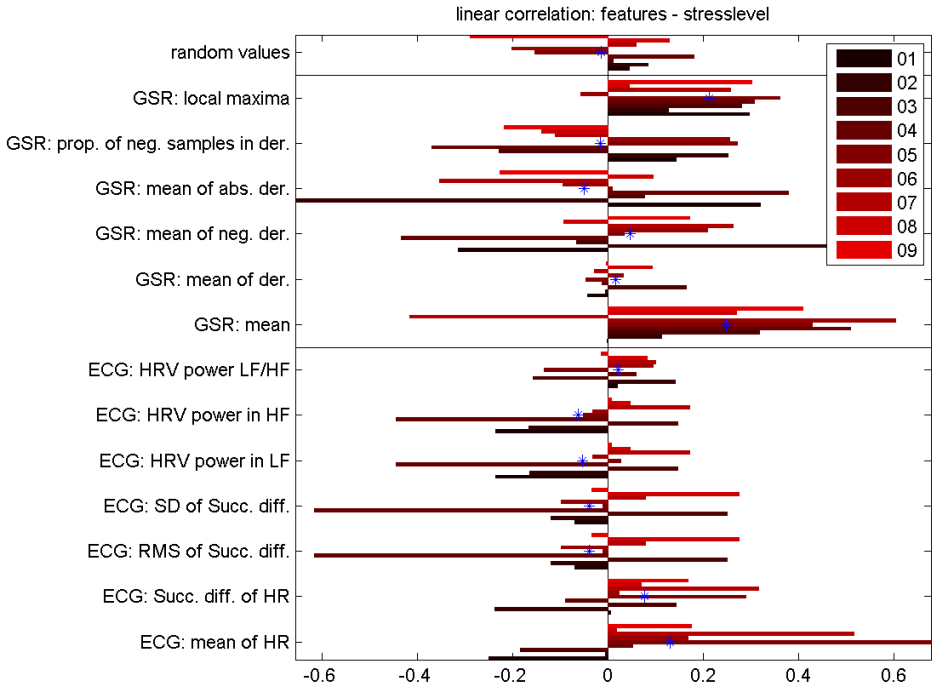

Figure 4.10: *The linear correlations between all experiments and their corresponding stress level.*

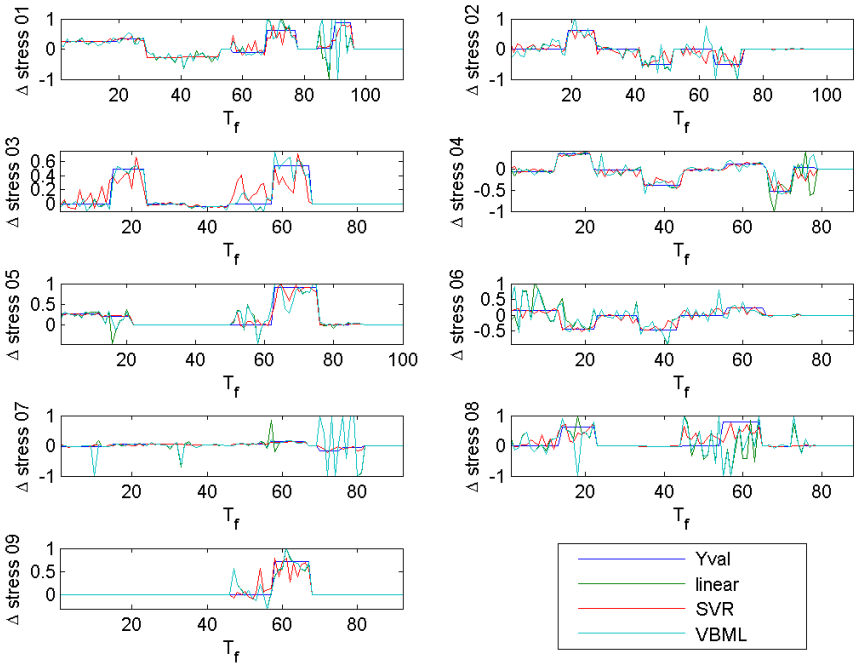

Figure 4.11: *The predicted stress levels for each subject, using cross validation within each task and subject. SVR, linear regression and VBML are compared.*

subject	linear RMS error	SVR RMS error	VBML RMS error
01	0.381	0.374	0.382
02	0.278	0.350	0.318
03	0.201	0.250	0.202
04	0.323	0.315	0.324
05	0.544	0.434	0.544
06	0.244	0.279	0.247
07	0.159	0.120	0.162
08	0.354	0.372	0.357
09	0.229	0.291	0.225

Table 4.7: *The RMS errors of different regressive models predicting the stress levels given by the subjects in the questionnaires. The indicated subject is used for validation and the other subjects are used as training data.*

The results of the calculated regressive model in multi-user multi-task mode is presented in Table 4.7. Generally they have a similar error, except for when the linear model and the VBML model have RMS errors greater than 0.5, when trying to predict subject 05:s stress levels. In this case the SVR stays on more reasonable error levels. Disregarding this problem, the linear model has the lowest RMS error in general.

Figure 4.12 presents the results from a cross validation, using eight subjects for training models and the last one as validation. The three continuous models are compared, however none of them output an accurate prediction. This prediction is evidently too hard; more subjects could help increasing the performance.

4.3 Discussion

In this section the results of Section 4.2 and the methodology presented in Section 4.1 are discussed. The purpose is to highlight interesting findings compared to Chapter 2.

4.3.1 Results

Also for the experimental data, quite high classification results are achieved, at least for the single user mode, when using SVM methods. The stress tasks all share common elements, except for the d2 task which seems to invoke a certain amount of stress, but not of a magnitude comparable to the other three more conventional stress tasks. As expected from the literature, the TSST invokes the highest stress reactions, and it is also what the subjects perceive as the most stressful moment of the experiments. Some machine learning experiments tested have shown ambiguous results due them simply being too hard for the classifiers to generalize well upon the limited amount of data. However a preliminary general

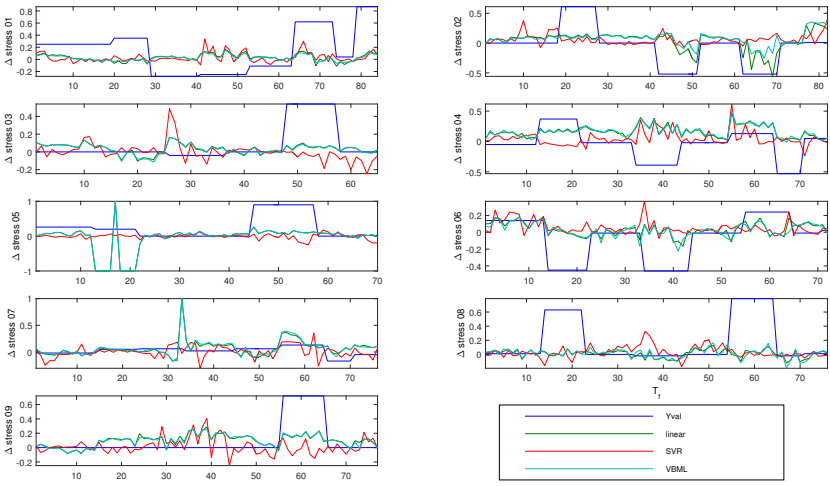

Figure 4.12: *The predicted stress levels for each subject, using the other subjects as training data. SVR, linear regression and VBML are compared.*

indicator of what features and sensors that are useful have been found.

Table 4.3 shows several results where the classification rate is below random guessing, i.e. 50 %. Probable reasons include that the models are learned on one sensor system and validated upon the same signals from another one, and that the sensor signals do not really resemble each other very well (Figures 4.4 and 4.5). Classification rates around 90 % and 5 % are both achieved, which could be motivated by chance, since very limited amounts of data were used for this particular experiment. This is the only case where a simple learning followed by a validation is used, contrary to the other cases where cross validation is applied. This indicates that one single model cannot efficiently be used for different sensor systems measuring the same signals.

Table 4.4 also shows results of classification rates is below 50 %, for the d2 task in the multi-user mode. This is probably explained by the fact that the features calculated from the d2 task all change differently depending on the subject (Figure 4.6). These classification rates have a value of approximately 40 %, and the one only one above 50 % (the NB classifier) reaches just above 60 %. This indicates that the d2 test induces different reactions depending on the subject, which is further suggested by the fact that in the single-user case classification rates at around 90 % are reached.

In this work, only two sensors and two signals were compared, but many other sensors could be tried in this way. If a wearable sensor manages to well classify stress using models learned from the laboratory equipment, it would be a good in-

dicator that a generalizable model has been found, and that the sensor performs well. This was unfortunately not the results for the E4 wristband, but more experiments with more subjects could change this. To give a fairer challenge to the sensor, using data from it to both learn and validate would be a good idea, but it would require it to be present at more experiments.

For the continuous model results presented in Table 4.7, the linear one and the VBML output RMS errors very close to each other, which is reasonable due to the similarity of the techniques, which are both linear (one also including probability estimates). For 4 out of 9 subjects, the linear RMS error is higher than the one of the SVM, which makes it hard to distinguish which model is better in the case of multi-user multi-task validation. Important to note is that the errors are found between 0.2 and 0.4, while they in the single user single task case (Table 4.6) take values of 0.05 - 0.2.

4.3.2 Method

In this work three methods were chosen for the continuous stress level estimation. Ideas for other regression methods for continuous stress estimation include logistic regression, tree regressions or non-linear methods (e.g. non-linear least squares). Among the tested methods, SVR had the best performance, but one must remember that it is significantly more complex and computationally demanding than e.g. the linear regression, which was not far behind in RMS error. Additionally, the tuning parameter options of the SVR could be be explored even more for further comparison and improvement of the models.

For a better comparison with the MIT driver database, a respiration sensor could be added to the experimental setup. Other ideas include EMG placed elsewhere than the face (e.g. shoulder, neck or foot for measuring the muscle tension). Like in Section 3.1.4, a feature selection optimization and comparison could be performed for each task, including forward and backward feature selection algorithms. But since the "intuitive choice" based upon the t-Test among other indicators worked well in Chapter 3, the decision was taken to focus more on task comparisons and continuous modelling of the experimental data.

Compared to the stress tasks used in Chapter 3, the laboratory tasks can be considered to be less influenced by other factors, since the subject is isolated and undisturbed from other events during the data acquisition.

An advantage of performing four things associated to each task (rest before control task, control task, rest before stress task and stress task) makes the work of classifying stress more reasonable than in the case of Chapter 3. The only difference between the tasks is the stress element, otherwise they are identical. This permits distinction from signal influence caused by the activity in itself, and allows focus on what difference the stress element makes.

Of course, when analyzing nine subjects the differences between persons cannot really be considered negligible. At least 25 - 50 subjects of different age, gender, backgrounds etc. would be needed if one wants to be available to draw strong

conclusions on the experiments. In this case, virtually all the subjects are young, healthy students.

Linear regressive models available to predict any stress level are rare in previous literature, which makes its performance hard to compare. The SVR model however is not completely off, at least when trying to predict a person's perceived stress level based upon the data of other persons performing the same tasks. In many cases, the subjects denote a higher stress level on the VAS questionnaire during the rest period preceding the control task, than during the control task itself. This might be due to the subject not knowing what comes next in the experiment, which might influence the result. One can then discuss if it is not more reasonable to continuously estimate stress than to divide it into classes. This seems like a scenario more closely related to the real world and what humans really experience.

4.3.3 Further Perspectives

Hopefully this study can serve to help future work choosing sensors, signals, features and modelling methods for stress detection. A possible application is e.g. a real-time application using the data of the E4 wristband to inform the user of her or his stress level via the mobile phone. This could be useful for both medical diagnosis and for informing other applications to adapt for the user's stress level to calm him or her down. Another idea is to help autistic children who have troubles communicating their emotional state to people in their vicinity.

Of course, one must always keep in mind whether a person wants to have his or her emotional state measured, which can be an intimate thing. In the case of stress it is less sensitive, but applications with other states and emotions are similar.

5
Conclusion

The purpose of this chapter is to get back to research questions stated in Chapter 1. It will also analyze the impact of this study and recommendations for similar work.

The research questions asked in Chapter 1 were:

1. Which sensors are most relevant for detecting stress?
2. Which signal properties are most relevant for detecting stress?
3. Which signal properties and features are common for different types of stress?
4. Which machine learning techniques are most relevant for modelling stress?

Regarding Question 1, the most interesting sensors related to stress has been found to be the ones measuring heart rate (ECG or PPG), skin conductance measures (GSR electrodes) and also respiration to a certain extent. The EMG data was chosen not to be analyzed due to the signal also being correlated to basic muscle movements. The skin temperature signal was excluded due to it being constant across experiments, and the accelerometer signals because they can primarily be used in combination with other sensors to further understand what the subject has been doing. The E4 wristband sensor is impressive for its size and simpleness, but GSR measurements from two points at the wrist cannot really compare to the ones measured at the fingers in terms of signal resolution. Also the calculated heart rate is a lot less exact compared to the one from an ECG at $F_s = 1$ kHz.

To answer Question 2, mainly simple features have proven to be important, such as the mean of HR and GSR values, which both tend to increase with stress. Furthermore some features on the derivative of the GSR have appeared, such as the

mean of the absolute derivative and the mean of the negative derivative. A feature of the respiration signal that is not traditionally used in the domain of stress detection, "feature x", has also appeared to be performing well.

The conclusions on Question 3 are that for all four laboratory experiments and the driver stress, the mean of the HR, the successive differences of the HR, the mean of the GSR and the number of local maxima in the GSR signal turn out to be common for different stress types. Other features show results with greater variation between subjects and stress types, and do not seem to generalize well.

Question 4 is basically answered by SVM:s which in almost every case perform better than other classification methods (such as NB, KNN and PNN). This is not extremely surprising, given the previous results presented in Section 2.6. Also in the case of continuous stress modelling, the SVR achieves a greater performance than simple linear regression and VBML. For a real-time application capable of combining both classification techniques and continuous estimation, SVM and SVR could be an idea, due to their similarity. This would be heavier computationally, thus if a weak CPU is the only option NB could be better options due to its smaller requirements on computation and due to their relatively high performance during found in this work. All of these are also differently sensitive to the chosen feature space, which for some work well with only few features (e.g. the SVM) while for others this choice is more critical (e.g. the KNN).

Since two different databases have been studied, the conclusions on the HR and GSR measures should generalize well. The results from Chapter 3 alone are a bit weaker since the study only compares one type of task and the classification is actually made between "doing nothing" and "driving", which is not identical to "stress" and "not stress".

Experience gained from this work is that machine learning requires a great amount of preprocessing of the data, "know your data". It is one of the most important steps, and if one fails here the created models will be useless in terms of generalization. In the case of physiological data, this means scanning the signals for unreasonable values and artifacts, and knowing the time constants of the changes in the signals. This is important to correctly choose and calculate signal features. Other examples is the knowledge of choosing a method to balance classes, knowing what kind of tuning the classifiers and probabilistic distribution they assume on the data for example.

5.1 Future Work

Future work includes expanding the experimental database with more subjects and possibly more sensors (e.g. a respiration signal), for drawing stronger conclusions. Given more time, a real-time application for detecting stress using the Empatica E4 wristband could be developed, to be used in experiments for validating the algorithms and the models. This sensor could use e.g. a KNN or an NB classifier, which are relatively simple to implement, and for more simplicity the

mean HR and the mean GSR and the number of local maxima in the GSR could be used as feature space. According to the results found in this work, there is hope for this setup working quite well, mainly if one adapts the model for each user, although this would require a calibration period.

Appendix

A

Stress Generating Tasks and Tests

This appendix summarizes tasks and tests that can be and has been used for generating stress.

A.1 Trier Social Stress Test

[27] presents the "Trier Social Stress Test" [51], [9], [39], [36], which has the purpose of inducing psycho-social stress in its participants using highly standardized methods. The test consists of three components, lasting about 5 minutes each. Prior to the beginning of the test, an IV and a heart rate monitor are mounted on the participant. The test starts by taking the participant to a room where three judges, a video camera and an audio recorder are waiting. The first part consists of 5 minutes of anticipatory stress, where the participant is asked to prepare a presentation lasting 5 minutes, commonly explained to be a job interview. The participant has the possibility to use a paper and pen to prepare the presentation, but these aids are removed when it is time to present. The judges should stay neutral during the test, avoiding to comment the presentation. The participant is asked to continue the presentation if it is finished before the 5 minutes have passed. Directly after the presentation, a 5 minute mental arithmetic test is performed. The task of the participant is to count backwards from 1,022, subtracting 13 in each step. They must start again from 1,022 if a mistake is made in the calculations. After this, a recovery period is performed, followed by a debriefing. The participant is informed that the only purpose of the test was to create stress. Samples are collected for a while even after the stress tasks have ended.

A.2 Socially Evaluated Cold Pressor Test

The SECPT [26], [32], [41] combines the physiological stress from cold water with the psychological stress from being filmed. The test should be performed between 2 p.m. and 5 p.m. to avoid influence from the circadian cortisol rhythm. First of all, blood pressure, ECG and saliva are sampled during 5 minutes for baseline data. The participants are then moved to another room, and asked to submerge their right hand in cold water (0 °C - 4 °C), until the wrist. This will continue for a maximum of three minutes, but the participant can remove the hand earlier. They are however asked to maintain the hand as long as possible. The participants are also informed that they will be filmed in order to evaluate their facial expressions, and are asked to look into the camera. During the test, ECG and blood pressure are are recorded during the immersion. Immediately after the submersion, the participant is asked to judge the discomfort, stress and pain level, on a scale from 0 to 10. The experiment leader watches the participants during the whole test, and asks them to remove their hand if the full 3 minutes are reached. After the test, saliva and blood pressure are sampled again during 5 minutes, in a new room. The cortisol in the saliva is measured at 10, 20, 30, 45 and 60 minutes after the end of the test. Between the samples, the participant is allowed to read. Finally, the participant is debriefed. Only men participate in this test due to bias from menstrual cycles in cortisol level. Participants should be in good health, non-smoking and with normal BMI (19 $\frac{\text{kg}}{\text{m}^2}$ - 27 $\frac{\text{kg}}{\text{m}^2}$). 3 hours before the test there should be no intesive physical activity and no alcohol, caffeine or food consumption.

A.3 d2 Test

[53] uses the d2 Test of attention, which contains 658 items (14 lines of 47 characters). The letters "d" and "p" are present, along with dashes. Between one and four dashes, individually or in pairs are placed above or below each character. The task of the subject is to find all "d:s" with two dashes, thus the name d2. If the subject has spent more than 20 seconds scanning the same line, the words "stop, next line" are pronounced to proceed the test. The subject is asked to scan at the highest rate possible without making any mistakes. The consistency and reliability of this attention test have been reported as good [53, p. 200].

A.4 Mental Arithmetic Stress Test

[59] uses a mental arithmetic stress task [18], [37], where it is explained to the participant that a quick mental test is performed in order to compare the result with other participants. The subject is also explained that other participants experienced the test as an easy task. The test consists of counting backwards from 2,193 in steps of seven, as rapidly and correctly as possible. This resembles the mental arithmetic tests in the Trier Social Stress Test, A.1. The participant is told

to start directly after the instructions are given unless there are questions. If a mistake is made, the researcher says “That was incorrect, please start again.” If the participant hesitates or pauses, the researcher says “Please continue”. This continues for five minutes, then the participant is asked to relax.

A.5 Other Methods

[10] presents the Montreal Imaging Stress Task, where a computer program displays arithmetic tasks to be solved by the participants under stress.

The Stroop color word [48] task uses the Stroop effect to generate stress. The Stroop effect is the finding that the reaction time when naming a colour is higher when the word is written in a colour not corresponding to the word.

[58] presents an academic task for induction of stress. It consists of verbal analogy trials. The participants had to choose the correct word to complete an analogy by pressing a key at the right time.

Bibliography

[1] Yaser Abu-Mostafa. Learning From Data, 2012. URL http://work.caltech.edu/telecourse.html. Cited on page 20.

[2] Jorn Bakker, Mykola Pechenizkiy, and Natalia Sidorova. What's Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 573–580, 2011. ISBN 978-0-7695-4409-0. Cited on pages 8, 11, and 20.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738. URL http://www.rmki.kfki.hu/~banmi/elte/Bishop%20-%20Pattern%20Recognition%20and%20Machine%20Learning.pdf. Cited on pages 31, 32, and 33.

[4] Kay H. Brodersen. Variational Bayesian linear regression, March 2013. URL http://people.inf.ethz.ch/bkay/downloads/Readme_vblm.pdf. Cited on page 72.

[5] Kay H. Brodersen, Jean Daunizeau, Christoph Mathys, Justin R. Chumbley, Joachim M. Buhmann, and Klaas E. Stephan. Variational Bayesian mixed-effects inference for classification studies. *NeuroImage*, 76:345 – 361, 2013. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2013.03.008. URL http://www.sciencedirect.com/science/article/pii/S1053811913002371. Cited on page 33.

[6] Chih-Chung Chang and Chih-Jen Lin. Training ν-Support Vector Regression: Theory and Algorithms. *Neural Computation*, 14(8):1959–1977, August 2002. doi: 10.1162/089976602760128081. URL http://www.csie.ntu.edu.tw/~cjlin/papers/newsvr.pdf. Cited on page 31.

[7] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*,

2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Cited on page 47.

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. URL https://www.jair.org/media/953/live-953-2037-jair.pdf. Cited on pages 28, 29, and 30.

[9] Frances S. Chen, Julian Schmitz, Gregor Domes, Brunna Tuschen-Caffier, and Markus Heinrichs. Effects of acute social stress on emotion processing in children. *Psychoneuroendocrinology*, 40:91 – 95, November 2014. ISSN 0306-4530. doi: http://dx.doi.org/10.1016/j.psyneuen.2013.11.003. URL http://www.sciencedirect.com/science/article/pii/S030645301300406X. Cited on pages 8 and 93.

[10] Katarina Dedovic, Robert Renwick, Najmeh Khalili Mahani, Veronika Engert, Sonia J. Lupien, and Jens C. Pruessner. The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *J Psychiatry Neurosci*, 30(5). Cited on pages 8 and 95.

[11] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, second edition, 2000. ISBN 0471056693. Cited on pages 16, 19, 24, 27, and 47.

[12] Centre d'Études sur le Stress Humain. Recette du stress, 2015. URL http://www.stresshumain.ca/le-stress/comprendre-son-stress/source-du-stress.html. Cited on page 5.

[13] Inc Empatica. Empatica E4: the Wearable Device for Researchers that Need Access to Real-world Physiological data, 2015. URL https://empatica.app.box.com/E4-TechSpecs. Cited on page 13.

[14] K. Frank, P. Robertson, M. Gross, and K. Wiesner. Sensor-based identification of human stress levels. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 127–132, March 2013. doi: 10.1109/PerComW.2013.6529469. URL http://www.csee.usf.edu/~labrador/Share/workshops/papers/p127-frank.pdf. Cited on pages 8, 12, 13, 14, 20, and 21.

[15] Andrea Gaggioli, Federica Pallavicini, Luca Morganti, Silvia Serino, Chiara Scaratti, Marilena Briguglio, Giulia Crifaci, Noemi Vetrano, Annunziata Giulintano, Giuseppe Bernava, Gennaro Tartarisco, Giovanni Pioggia, Simona Raspelli, Pietro Cipresso, Cinzia Vigna, Alessandra Grassi, Margherita Baruffi, Brenda Wiederhold, and Giuseppe Riva. Experiential Virtual Scenarios With Real-Time Monitoring (Interreality) for the Management of Psychological Stress: A Block Randomized Controlled Trial. *J Med Internet Res.*, 8. Cited on page 6.

[16] C. Godin, F. Prost-Boucle, A. Campagne, S. Charbonnier, S. Bonnet, and A. Vidal. Selection of the Most Relevant Physiological Features for Classifying Emotion. In *Proceedings of the 2nd International Conference on Physiological Computing Systems*, pages 17–25, 2015. ISBN 978-989-758-085-7. doi: 10.5220/0005238600170025. URL http://www.researchgate.net/publication/271470987_SELECTION_OF_THE_MOST_RELEVANT_PHYSIOLOGICAL_FEATURES_FOR_CLASSIFYING_EMOTION. Cited on pages 16 and 17.

[17] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8. Cited on page 31.

[18] Skjalg S. Hassellund, Arnljot Flaa, Leiv Sandvik, Sverre E. Kjeldsen, and Morten Rostrup. Long-Term Stability of Cardiovascular and Catecholamine Responses to Stress Tests. *Hypertension*, 55(1), January 2014. URL http://hyper.ahajournals.org/content/55/1/131.full.pdf. Cited on pages 8 and 94.

[19] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, February 2009. URL http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf. Cited on pages 19, 22, 23, 24, 25, 29, and 31.

[20] J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, (2):156–166, June . ISSN 1524-9050. doi: 10.1109/tits.2005.848368. Cited on pages 1, 7, 8, 17, 31, 60, and 61.

[21] Jennifer Healey and Rosalind W. Picard. Driver Stress Data, MIT Affective Computing Group, 2002. URL http://affect.media.mit.edu. Cited on page 35.

[22] Javier Hernandez, Robert R. Morris, and Rosalind W. Picard. Call Center Stress Recognition with Person-Specific Models. In Sidney K. D'Mello, Arthur C. Graesser, Björn Schuller, and Jean-Claude Martin, editors, *ACII (1)*, Lecture Notes in Computer Science, pages 125–134. Springer. ISBN 978-3-642-24599-2. Cited on pages 8, 13, and 21.

[23] Andreas Holzinger, Manuel Bruschi, and Wolfgang Eder. On Interactive Data Visualization of Physiological Low-Cost-Sensor Data with Focus on Mental Stress. In *Availability, Reliability, and Security in Information Systems and HCI - IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2013, Regensburg, Germany, September 2-6, 2013. Proceedings*, pages 469–480, 2013. Cited on pages 8, 13, and 14.

[24] Delsys Incorporated. TRIGNO Wireless System User's Guide, October 2010. URL http://www.biomedicale.parisdescartes.fr/pf-sensorimotricite/wp-content/uploads/2014/04/Trigno-Users-Guide.pdf. Cited on page 67.

[25] Cornelia Kappeler-Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410–417, March 2010. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5325784&isnumber=5431105. Cited on pages 6, 8, 11, 13, 14, and 21.

[26] Valerie L. Kinner, Serkan Het, and Oliver T. Wolf. Emotion regulation: exploring the impact of stress and sex. *Front Behav Neurosci*, 8, November 2014. Cited on pages 8 and 94.

[27] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. The "Trier Social Stress Test" - A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting. *Neuropsychobiology*, 8, 1993. URL http://p113367.typo3server.info/uploads/media/lit9304.pdf. Cited on pages 8 and 93.

[28] Ivan Kojadinovic and Thomas Wottka. Comparison between a filter and a wrapper approach to variable subset selection in regression problems. In *Proceedings of the European Symposium on Intelligent Techniques (ESIT 2000), Aachen, Germany*, September 2000. Cited on page 17.

[29] D. McDuff, S. Gontarek, and R. Picard. Remote measurement of cognitive stress via heart rate variability. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2957–2960, Aug 2014. doi: 10.1109/EMBC.2014.6944243. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6944243&tag=1. Cited on pages 8, 14, and 21.

[30] Saul McLeod. What is the Stress Response, 2010. URL http://www.simplypsychology.org/stress-biology.html. Cited on page 5.

[31] Nicola Blefari Melazzi. Bonvoyage 2020 Project. URL http://bonvoyage2020.eu/. Cited on page 1.

[32] Nina Minkley, Thomas P. Schröder, Oliver T. Wolf, and Wolfgang H. Kirchner. The socially evaluated cold-pressor test (SECPT) for groups: Effects of repeated administration of a combined physiological and psychological stressor. *Psychoneuroendocrinology*, 45, March 2014. Cited on pages 6, 8, and 94.

[33] Puran and Susanna Bair. Respiratory Sinus Arrhythmia (RSA), 2009. URL https://iamheart.org/dome/201/stimulation/RSA.html. Cited on page 10.

[34] Michael Randall. The Physiology of Stress: Cortisol and the Hypothalamic-Pituitary-Adrenal Axis, 2011. URL http://dujs.dartmouth.edu/fall-2010/the-physiology-of-stress-cortisol-and-the-

hypothalamic-pituitary-adrenal-axis#.VQbDho7A5ph. Cited on page 6.

[35] Alejandro Riera, Aureli Soria-Frischa, Anton Albajes-Eizagirrea, Carles Grau Pietro Cipressob, Stephen Dunnea, and Giulio Ruffini. Electro-Physiological Data Fusion for Stress Detection. In *Studies in health technology and informatics, Volume: 181*. Starlab Barcelona SL., PubMed, 2012. URL http://www.researchgate.net/publication/230810302_Electro-Physiological_Data_Fusion_for_Stress_Detection. Cited on pages 1, 7, 8, 14, and 21.

[36] Mary R.Lee, Kelsey Cacic, Catherine H.Demers, Maleeha Haroon, Stephen Heishman, Daniel W Hommera, David H. Epstein, Thomas J. Ross, Elliot A. Stein, Markus Heilig, and Betty Jo Salmeron. Gender differences in neural–behavioral response to self-observation during a novel fMRI social stress task. *Neuropsychologia*, 53, December 2013. URL http://www.researchgate.net/publication/259246337_Gender_differences_in_neural-behavioral_response_to_self-observation_during_a_novel_fMRI_social_stress_task. Cited on pages 6, 8, and 93.

[37] Morten Rostrup, Arne Westheim, Sverre E. Kjeldsen, and Ivar Eide. Cardiovascular Reactivity, Coronary Risk Factors, and Sympathetic Activity in Young Men. *Hypertension*, 22(6), December 1993. URL http://hyper.ahajournals.org/content/22/6/891.full.pdf. Cited on pages 8 and 94.

[38] Akane Sano and Rosalind 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, pages 671–676, 2013. ISBN 978-0-7695-5048-0. Cited on pages 8, 14, and 21.

[39] Laura R. Saslow, Shannon McCoy, Ilmo van der Löwe, Brandon Cosley, Arbi Vartan, Christopher Oveis, Dacher Keltner, Judith T. Moskowitz, and Elissa S. Epel. Speaking under pressure: Low linguistic complexity is linked to high physiological and emotional stress reactivity. *Psychophysiology*, 51(3), June 2014. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4059522/pdf/nihms-591391.pdf. Cited on pages 8 and 93.

[40] Dr. Saed Sayad. K Nearest Neighbors - Classification. URL http://www.saedsayad.com/k_nearest_neighbors.htm. Cited on page 27.

[41] Lars Schwabe, Leila Haddad, and Hartmut Schachinger. HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology*, 33, March 2008. Cited on pages 8 and 94.

[42] Brian Luke Seaward. *Managing Stress*. Jones and Bartlett Publisher, 2015. Cited on page 6.

[43] Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, 108(3):1287–1301, July 2012. URL http://ac.els-cdn.com/S0169260712001770/1-s2.0-S0169260712001770-main.pdf?_tid=8c13a7aa-cc83-11e4-82eb-00000aab0f6c&acdnat=1426582789_cca9a9718962902583578cc1b1fbedc2. Cited on pages 1, 13, 14, 15, 20, 21, and 24.

[44] Yuan Shi, Minh Hoai Nguyen, Patrick Blitz, Brian French, Scott Fisk, Fernando De la Torre, Asim Smailagic, Daniel P. Siewiorek, Mustafa al' Absi, Emre Ertin, Thomas Kamarck, and Santosh Kumar. Personalized stress detection from physiological measurements. *International Symposium on Quality of Life Technology*, 2010. URL http://www.humansensing.cs.cmu.edu/projects/stress_detect/stress_detect.pdf. Cited on pages 8, 12, 13, 14, 15, and 21.

[45] Strahler Jana Schlotz Wolff Niederberger Larissa Marques Sofia Fischer Susanne Thoma Myriam V. Spoerri Corinne Ehlert Ulrike Nater Urs M. Skoluda, Nadine. Intra-individual psychological and physiological responses to acute laboratory stressors of different intensity. *Psychoneuroendocrinology*, 51(Complete):227–236, 2015. doi: 10.1016/j.psyneuen.2014.10.002. URL http://www.psyneuen-journal.com/article/S0306-4530%2814%2900384-9/pdf. Cited on pages 6 and 8.

[46] Donald F. Specht. Probabilistic Neural Networks. *Neural Netw.*, 3(1):109–118, January 1990. ISSN 0893-6080. doi: 10.1016/0893-6080(90)90049-Q. URL http://courses.cs.tamu.edu/rgutier/cpsc636_s10/specht1990pnn.pdf. Cited on page 27.

[47] Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358 – 3378, 2007. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2007.04.009. URL http://sci2s.ugr.es/keel/pdf/algorithm/articulo/2007%20-%20PR%20-%20Sun%20-%20Cost-Sensitive%20boosting.pdf. Cited on page 30.

[48] J. H. M. Tulen, P. Moleman, H. G. van Steenis, and F. Boomsa. Characterization of Stress Reactions to the Stroop Color Word Test. *Pharmacol Biochem Behav*, 32, February 1988. URL http://www.sciencedirect.com/science/article/pii/0091305789902049#. Cited on pages 8 and 95.

[49] Arizone State University. Algorithms | Feature Selection @ ASU. URL http://featureselection.asu.edu/software.php. Cited on page 17.

[50] E. L. van den Broek. *Affective Signal Processing (ASP): Unraveling the mystery of emotions*. PhD thesis, University of Twente, Enschede, Septem-

ber 2011. URL http://eprints.eemcs.utwente.nl/20812/01/VandenBroek11-Affective_Signal_Processing_ASP2.pdf. Cited on pages 7, 9, 15, and 21.

[51] Bart Verkuil, Jos F. Brosschot, and Julian F. Thayer. Cardiac reactivity to and recovery from acute stress: Temporal associations with implicit anxiety. *International Journal of Psychophysiology*, 92(2):85 – 91, March 2014. ISSN 0167-8760. doi: http://dx.doi.org/10.1016/j.ijpsycho.2014.03.002. URL http://www.sciencedirect.com/science/article/pii/S0167876014000713. Cited on pages 8 and 93.

[52] Katarzyna Wac and Christiana Tsiourti. Ambulatory Assessment of Affect: Survey of Sensor Systems for Monitoring of Autonomic Nervous Systems Activation in Emotion. *T. Affective Computing*, 5(3):251–272, 2014. Cited on page 7.

[53] Renske Wassenberg, Jos G M Hendriksen, Petra P M Hurks, Frans J M Feron, Esther H H Keulers, Johan S H Vles, and Jelle Jolles. Development of inattention, impulsivity, and processing speed as measured by the d2 Test: results of a large cross-sectional study in children aged 7-13. *Child Neuropsychol*, 14(3):195–210, November 2008. ISSN 0929-7049. URL https://atmire.com/dspace-labs3/bitstream/handle/123456789/6919/file14218.pfd.pdf. Cited on pages 8 and 94.

[54] Griffin Weber. A Comparison of Single Lead ECG Data Compression Techniques, 1998. URL http://www.hcs.harvard.edu/~weber/HomePage/Papers/ECGCompression/. Cited on page 10.

[55] Peter D. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *Audio and Electroacoustics, IEEE Transactions on*, 15(2):70–73, Jun 1967. ISSN 0018-9278. doi: 10.1109/TAU.1967.1161901. URL https://www.utd.edu/~cpb021000/EE%204361/Great%20DSP%20Papers/Welchs%20Periodogram.pdfm. Cited on page 40.

[56] J. Wijsman, B. Grundlehner, Hao Liu, H. Hermens, and J. Penders. Towards mental stress detection using wearable physiological sensors. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 1798–1801, Aug 2011. doi: 10.1109/IEMBS.2011.6090512. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6090512. Cited on pages 7, 8, 17, and 21.

[57] Jacqueline Wijsman. *Sensing stress: stress detection from physiological variables in controlled and uncontrolled conditions*. PhD thesis, Enschede, December 2014. Cited on page 10.

[58] C. J. (Christopher) Wilson, Y. Barnes-Holmes, and D. Barnes-Holmes. The

effect of emotion regulation strategies on physiological and self-report measures of anxiety during a stress-inducing academic task. *International Journal of Psychology and Psychological Therapy*, 14(1), 2014. ISSN 1889-1780. URL http://hdl.handle.net/10149/550228. Cited on pages 8 and 95.

[59] Lindsay K. Yamaoka. The Effects of Adherence to Asian Values and Extraversion on Cardiovascular Reactivity: A Comparison Between Asian and European Americans. Brandeis University Senior Honors Thesis, 2014. Senior Thesis. Cited on pages 7, 8, and 94.

[60] Jing Zhai and Armando Barreto. Stress Recognition Using Non-invasive Technology. In Geoff Sutcliffe and Randy Goebel, editors, *FLAIRS Conference*, pages 395–401. AAAI Press. Cited on pages 7, 8, 15, and 21.

[61] Joy Ying Zhang. Confidence interval and the Student's t-test, December 2006. URL http://projectile.sv.cmu.edu/research/public/talks/t-test.htm. Cited on page 16.

[62] Nina Zhou and Lipo Wang. A Modified T-test Feature Selection Method and Its Application on the HapMap Genotype Data. *Genomics, Proteomics & Bioinformatics*, 5(3–4):242 – 249, 2007. ISSN 1672-0229. doi: http://dx.doi.org/10.1016/S1672-0229(08)60011-X. URL http://www.ntu.edu.sg/home/elpwang/PDF_web/07_GPB.pdf. Cited on page 16.