KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Towards large-scale physiological stress detection in an ambulant environment

Elena Smets

Supervisors:
Prof. Dr. Ir. Chris Van Hoof
Prof. Dr. Ilse Van Diest

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

October 2018

Towards large-scale physiological stress detection in an ambulant environment

Elena Smets

Examination Committee:

Prof. Dr. Ir. Yves Willems, Chair
Prof. Dr. Ir. Chris Van Hoof, supervisor
Prof. Dr. Ilse Van Diest, co-supervisor
Prof. Dr. Ir. Robert Puers
Prof. Dr. Stephan Claes
Ir. Walter De Raedt
Prof. Dr. Ir. Piet Demeester
Prof. Dr. Ir. Dimitrios Fotiadis

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

In collaboration with imec
Kapeldreef 75
B-3001 Heverlee, Belgium

October 2018

Uitgegeven in eigen beheer, Elena Smets, Kasteelpark Arenberg 10, B-3001 Heverlee (Belgium)

Acknowledgements

The past four years have taught me that studying stress and dealing with your own stress are not quite the same thing. I have learned a thing or two about root causes of stress: papers that only get accepted after a zillion revisions, trial participants who call you in the weekend to notify you that "everything crashed", doing a live demo for the king,... But more importantly I have come to learn that the best way to deal with stress is through the support of colleagues, friends and family. Therefore, I want to start this dissertation by thanking the people who have been there for me along the road and who have contributed to four great years.

First and foremost, I want to thank my promotor, Prof. Chris Van Hoof. To be entirely honest, you have contributed to my PhD both as a stressor and a stress reliever, and I mean this in the best possible way. When at the start of my PhD you subtly added two zeros to the 15 test subjects I was going to measure, and with that tiny effort made it 1,500 test subjects, I knew the next four years were going to be challenging. Thank you for making these four years challenging! Thank you for giving me countless opportunities to learn, improve, present and demonstrate my research both inside and outside imec, with partners, prospective clients as well as the broader public. Also, I want to thank you for the many not work-related talks about running, eating, travelling: the good things in life. You have been a unique promotor, who really made a difference for my PhD and my overall experience at imec during the past four years.

Next, I want to thank my co-promotor, Prof. Ilse Van Diest, and assessor Prof. Stephan Claes. This research has been inherently interdisciplinary, and would have been far less valuable without your contributions. Thank you for giving me so many new insights on stress psychology and physiology and for always being available whether it was for paper revisions or discussions on my results.

A special thank you goes to my daily supervisor Walter De Raedt. From my very first day at imec you have helped me whenever needed and you have made me feel at home at imec. At the start we used to have our weekly

meetings to get me on the right track for my PhD roadmap. These discussions have been extremely valuable and have guided a large part of my PhD. Along the way, I also got to know you better and I have enjoyed our many work and not work-related talks. A special thank you for listening whenever I passed by your office to start a monologue about my frustrations (usually about things not moving fast enough). Those talks have been crucial for my stress relief!

I want to thank all other members of the examination committee. Prof. Robert Puers, thank you for taking up the role as assessor in my committee. I have appreciated the time and effort you have taken to always critically analyze the progress of my PhD, to carefully read my thesis and to provide me with relevant feedback. Further, I want to thank Prof. Piet Demeester and Prof. Dimitrios Fotiadis for accepting their role in my examination committee and providing me with important feedback to improve the text of my dissertation. I also want to thank Prof. Yves Willems for being the chair of my committee.

I am grateful for the funding I received from the "Agentschap Innoveren & Ondernemen" (Vlaio) to execute this PhD.

A major thank you goes to my colleague Jan Cornelis. Without your help the data collection of the SWEET study would simply not have been possible. Thank you for always being available, for never complaining, and for always being helpful to search for a constructive solution when things were going wrong. In the same vein, I want to thank Giuseppina and Emmanuel for working together on the data analysis of the study. I am extremely proud of the work we achieved all together: the dataset, the analysis and hopefully the many results and publications that will follow. We have worked closely together, and often under stressful circumstances, which have not always been easy. You have taught me a great deal on many different aspects of data science, paper writing and also on a personal level. I have enjoyed the many fruitful discussions, with often very different opinions, which I think have led to better insights and results for the entire team. Thank you for being part of this journey with me.

I also want to thank Ellie and Imen for your support with the Exascience team. Imen, thank you for getting all the scripts structured and turning them into an

efficient pipeline to extract the data from the SWEET study. But most of all, thank you for helping me deal with the many problems we faced in the smartphone data collection and communication. Our 'animated' talks about how to tackle these issues and answering e-mails have been of great value to me!

Further, I want to thank the imec.ichange team: Annelies, Matthias, An, Sofie, Olivier, Tanguy, Giuseppina and Emmanuel. I have truly enjoyed our collaborations over the past years. Matthias, I think this is an appropriate time to apologize for the stress I have caused you with the demos. I have very much appreciated your and your team's effort to always go the extra mile to deliver great demos! I really believe in the imec.ichange program and I am convinced it will continue to grow in the next years. I am proud to have been part of this interdisciplinary, ambitious team and I look forward to seeing the future achievements.

A big thank you goes to my colleagues of the BECHSRD team. I must admit that at the start of my PhD I wasn't sure if I would ever fit in the team. I could only understand half of the topics: everyone was being anxious about tape-outs (never heard of it, searched it on Wikipedia, conclusion: it's a deadline, please guys just call it a deadline), talking about front-ends and back-ends, neuroprobes,... I felt a bit lost. But that quickly changed when Tom gave me a short introductory course explaining all topics mentioned. Later we also started the running team with Tom, Jan, Carolina, Ivan and Shuang. Although we might not go as regularly as initially foreseen, I have truly enjoyed our runs! Further, I have learned a lot from all my fellow PhD students Neide, Eric, Erika, Ivan, Bogdan, Yun-Hsuan and Rachit both on work and not work-related topics. I have also enjoyed the many barbecues, team events and especially the weekend in the Ardennes, where I have been surprised by the great karaoke skills of the team, especially the Chinese and Koreans! Thank you guys for this amazing time at imec and for making me feel at home.

Over the years I have had the chance to supervise many thesis and internship students. You all have contributed strongly to my PhD and in a broader sense to the imec.ichange program. Thank you for your enthusiasm and your great work!

If it weren't for GDPR, this acknowledgement section would take up half of the length of my PhD for thanking the more than 1,000 test subjects who have contributed to my trials. It has been a tough cookie to get all of you on board, but it has been extremely motivating to feel all your enthusiasm related to this research and openness towards new technologies.

Further, I want to especially thank my friends and family for their support during the past four years. My friends from Leuven: Ruth, Elise, Sara, Laura, Kaat, Bibi, Eva and Bram, thank you for always being enthusiastic, for all the great weekends, parties, dinners, runs,... we have had the last years. I already say this every year at Thanksgiving, while holding sweaty hands, but I'm truly grateful for having you as my friends! Also my bio-engineering friends: Nore, Elien, Charlien, Alex and Sander and the Brabançonne-ladies Hanne and Tine, thank you for the great times and parties we have shared during our student years and for the dinners and activities we still do. I know schedules are difficult sometimes, but I really hope we can continue doing those things still for a long time! Also a special thanks to my Portugal/Brasil/Leuven/Suits friends: PJ, Willem, Laurens, Joke and Lena. I always enjoy our dinners and I love our trips to Portugal. I hope that tradition will still continue for a long time.

I want to thank my parents and sister for their unconditional belief and support in me, my studies and everything I do. I am grateful to have grown up in such a warm, enthusiastic (maybe sometimes too enthusiastic), adventurous (maybe sometimes too adventurous), and caring family.

Finally, and most importantly, I want to thank my husband Jasper. Thank you for always supporting me, for always having my back, for always putting a smile on my face when I come home, and for always helping me to put things in perspective. I am looking forward to the years ahead of us!

Abstract

In the 21st century, stress and mental health have become major concerns worldwide. Yet, a continuous, quantitative measurement technique, allowing just-in-time interventions to reduce stress, is lacking. Therefore, research has focused on exploiting the sympathetic nervous system's (SNS) fight-or-flight response, by investigating physiological signals for monitoring stress. Research has focused on developing machine learning models for stress detection, based on physiological signals such as heart rate (HR), skin conductance (SC), skin temperature (ST) and respiration. These have shown to be reliable indicators of stress in well-controlled laboratory conditions, but large-scale ambulatory validation is missing.
The goal of this research is to identify physiological sensing priorities and machine learning techniques for physiological stress detection and next, to deploy these on a large population in real-life, ambulatory conditions.
To this end, this dissertation focuses on three objectives.

First, we aim to **identify the most suitable markers for physiological stress detection**. We set up a trial, including stress-inducing stimuli in a controlled, laboratory environment with 20 healthy subjects. The data is used to identify physiological sensing priorities, compare machine learning techniques and investigate inter-person variability. We conclude that, on average, SC and HR related features are more important than ST and respiration related features. However, on a personal level, physiological sensing priorities differ across subjects, favoring a multi-sensor approach. Based on the comparison of six machine learning techniques, we conclude that for generalized models (i.e. including all subjects), support vector machines (SVMs) perform best, for personalized models (i.e. based on one subject), dynamic Bayesian networks perform best. Overall, personalized models outperform generalized models. The selection of the most optimal technique depends on the context of the application.

Second, we aim to **differentiate between healthy subjects and patients based on their physiological stress response**, towards disease prevention

and interception. We repeat the previous experiment with 12 patients with stress-related complaints and use an exploratory methodology to classify healthy subjects and patients. Our results show the potential of using physiological signals for the interception of stress-related diseases (e.g. burnout) and contain large value towards prevention.

Third, we aim to **investigate the physiological stress response on a large scale in ambulatory conditions**. We present the SWEET study: world's largest ambulatory stress detection study, including 1,002 subjects who are continuously monitored during 5 days. We present a protocol including physiological sensing, baseline psychological information, self-reported stress and contextual sensing based on smartphone information. Results highlight the need for personalized models to detect stress, based on the development of digital phenotypes, i.e. personas for stress detection based on digital information including physiological, contextual and psychological baseline data. Further, we present a methodology to use subject-specific information, based on the physiological response to a specific stress task in an ambulant environment, towards a personalized calibration for ambulant stress detection models.

The results of this dissertation provide a first step towards personalized stress detection, and more generally **towards precision medicine and personalized healthcare**. In the future, physiological stress detection, including context information, could enable just-in-time adaptive intervention strategies, towards early detection and prevention of stress-related diseases and cause a paradigm shift from treatment to disease prevention and interception.

Samenvatting

Stress en mentale gezondheid zijn in de 21ste eeuw wereldwijd een belangrijke bekommering geworden. Toch is er een gebrek aan continue, kwantitatieve meettechnieken, die stressreductie interventies op het juiste moment toelaten. Daarom worden fysiologische signalen, die veranderen ten gevolge van de vecht-of-vluchtreactie van het sympathische zenuwstelsel, onderzocht voor stressdetectie. Eerder onderzoek heeft gefocust op de ontwikkeling van machine learning modellen voor stressdetectie, gebaseerd op fysiologische signalen zoals hartslag (HR), huidgeleiding (SC), huidtemperatuur (ST) en ademhaling. In gecontroleerde labo-omgevingen zijn deze signalen betrouwbare indicatoren voor stress gebleken, maar grootschalige, ambulante validatie ontbreekt.
Het doel van dit onderzoek is om de belangrijkste fysiologische signalen en machine learning technieken voor stressdetectie te identificeren en vervolgens toe te passen op een grootschalige populatie in ambulante condities.
Daartoe beoogt deze thesis drie doelstellingen.

Vooreerst beogen we de **meest geschikte markers voor fysiologische stressdetectie te identificeren**. We zetten een onderzoek op met 20 gezonde deelnemers, die we onderwerpen aan stress inducerende stimuli in een gecontroleerde labo-omgeving. De data wordt gebruikt om de belangrijkste fysiologische signalen en machine learning technieken te bepalen en om interpersoonlijke variabiliteit te onderzoeken. We besluiten dat, gemiddeld genomen, SC en HR gerelateerde variabelen belangrijker zijn dan ST en ademhaling gerelateerde variabelen. Echter, op een persoonlijk niveau verschilt het belang van fysiologische signalen, waardoor een multimodale sensor aanpak aangewezen is. We vergelijken zes machine learning technieken en besluiten dat voor algemene modellen (waarbij alle deelnemers inbegrepen zijn), support vector machines (SVMs) het beste presteren, en voor gepersonaliseerde modellen (gebaseerd op de data van slechts een deelnemer), dynamische Bayesiaanse netwerken het beste presteren. Algemeen presteren gepersonaliseerde modellen beter dan algemene modellen. De keuze van de machine learning techniek is afhankelijk van de context waarin deze gebruikt zal worden.

Ten tweede, beogen we **gezonde mensen en patiënten te onderscheiden op basis van hun fysiologische stressrespons**, met als doel ziekte preventie en interceptie. We herhalen het vorige experiment met 12 patiënten die last hebben van stress gerelateerde klachten en gebruiken een exploratieve methodologie om gezonde mensen en patiënten te onderscheiden. Onze resultaten tonen het potentieel van fysiologische signalen voor de interceptie van stress gerelateerde ziektes (bv. Burnout) en zijn van grote waarde in de richting van preventie.

Ten derde, beogen we een **grootschalig onderzoek naar de fysiologische stressrespons in ambulante condities**. We stellen de SWEET studie voor: 's werelds grootste studie rond ambulante stressdetectie, bij 1002 deelnemers die gedurende 5 dagen continu gemonitord worden. We presenteren een protocol met fysiologische metingen, baseline psychologische informatie, zelf-gerapporteerde stress en context metingen op basis van smartphone informatie. De resultaten benadrukken de nood aan gepersonaliseerde modellen voor stressdetectie, op basis van de ontwikkeling van digitale fenotypes, dit zijn personas voor stressdetectie op basis van digitale informatie waaronder fysiologische, contextuele en psychologische baseline data. Verder, presenteren we een methodologie om persoonsgebonden informatie, gebaseerd op de fysiologische respons op een stresserende taak in een ambulante omgeving, te gebruiken voor gepersonaliseerde kalibratie voor ambulante stressdetectie.

De resultaten van deze thesis bieden een eerste stap richting gepersonaliseerde stressdetectie en meer algemeen richting **precisie geneeskunde en gepersonaliseerde gezondheidszorg**. In de toekomst, zou fysiologische stressdetectie, gecombineerd met contextuele informatie, gebruikt kunnen worden om adaptieve interventiestrategieën aan te bieden op het juiste moment. Dit zou vroege detectie en preventie van stress gerelateerde ziektes mogelijk kunnen maken en een paradigmaverschuiving teweeg kunnen brengen van ziekte behandeling naar ziekte preventie en interceptie.

Nomenclature

AbsDiff2	Absolute second difference of the skin conductance signal
ACC	Acceleration
ACTH	Adrenocorticotropic hormone
ADR	Average detection rate
AI	Artificial intelligence
ANN	Artificial neural network
ANS	Autonomous nervous system
AVP	Arginine vasopressin
BMI	Body mass index
BN	Bayesian network
BP	Blood pressure
BVP	Blood volume pulse
CI	Confidence interval
CK	Cohen-Kappa
CNS	Central nervous system
CRF	Corticotropin-releasing factor
DASS	Depression anxiety stress scale
dBN	Dynamic Bayesian network
DT	Decision tree
EB1	Energy band 0-0.1 Hz
EB2	Energy band 0.1-0.2 Hz
EB3	Energy band 0.2-0.3 Hz
EB4	Energy band 0.3-0.4 Hz
ECG	Electrocardiography
EMA	Ecological momentary assessment
EMG	Electromyography
ERI	Effort reward imbalance
HPA	Hypothalamic pituitary adrenal
HF	High frequency band of the RR intervals (0.15-0.4 Hz)
HR	Heart rate
HRV	Heart rate variability
JDC	Job demands-control

JDCS	Job demands-control-support
JDR	Job demands-resources
JITAI	Just-in-time adaptive intervention
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LPDS	Leuven postprandial distress scale
LF	Low frequency band of the RR intervals (0.04-0.15 Hz)
LFHF	Low frequency over high frequency bands of the RR intervals
LR	Logistic regression
meanRSP	Mean respiration frequency
mHR	Mean heart rate
mPhasic	Phasic component of the skin conductance signal (0.16-2.1 Hz)
mT	Mean skin temperature
mTonic	Tonic component of the skin conductance signal (0-0.16 Hz)
MINI	Mini international neuropsychiatric interview
MIST	Montreal imaging stress task
ML	Machine learning
NrPeaks	Number of peaks of the skin conductance signal
OLS	Ordinary least squares
OPD	Ohmic perturbation duration
PAM	Partitioning around medoids
PCA	Principal component analysis
pNN20	Proportion of the successive normal to normal beat intervals that differ more than 50ms
pNN20	Proportion of the successive normal to normal beat intervals that differ more than 20ms
PNS	Parasympathetic nervous system
PPG	photoplethysmography
PSS	Perceived stress scale
PSQI	Pittsburg sleep quality index
PTSD	Posttraumatic stress disorder
PVN	Paraventricular nucleus
QI	Quality indicator

RDR	Rest detection rate
RF	Random forest
RMSSD	Root mean square of the successive RR differences
RSA	Respiratory sinus arrhythmia
SAM	Sympathetic adrenal medullary
SAM	Self-assessment manikin
sBN	Static Bayesian network
SC	Skin conductance
SCL	Skin conductance level (mean)
SCL-90	Dutch symptom checklist-90
SD	Standard deviation
SDNN	Standard deviation of the normal to normal beat intervals
SDR	Stress detection rate
SNS	Sympathetic nervous system
ST	Skin temperature
slopeT	Slope of the skin temperature
stdT	Standard deviation of the skin temperature
SVM	Support vector machine
SVR	Support vector regression
SWEET	Stress in the work environment
TSST	Trier social stress test

Contents

List of Figures

Chapter 1: Introduction

In the 21st century, stress and mental health at work have become major concerns for organizations worldwide [1]. The American Psychological Association states that in 2015 in the US 24% of adults reported extreme stress [2]. In 2013 in Europe 51% of the working population reports that cases of work-related stress are common in their workplaces, with most important causes of stress being job reorganization or job insecurity and hours worked or workload [3].

Research has already extensively discussed the negative consequences of stress [4, 5, 6, 7]. For example, observational data suggest an average 50% increased risk for coronary heart disease among employees with work stress [8]. Besides these personal health effects, also at the level of organizations increased stress levels can have a negative impact. It has been shown that people perform worse under excessive stress [4]. Many studies have tried to estimate the cost of stress and while quantitative data is scarce, a report of the European Agency for Safety and Health at Work states that in 2002 the cost of work-related stress for Europe was estimated at €20 billion [9]. A more recent study in 2013 estimates the cost of work-related depression in Europe at €617 billion annually [10]. This total is a combination of costs due to absenteeism and presenteeism (€272 billion), loss of productivity (€242 billion), health care costs (€63 billion) and social welfare costs (€39 billion). Although the accuracy of these numbers is debatable as outcomes vary heavily on measurement techniques, they show that prevention and detection of stress at work in an early stage are of utmost importance for both welfare and economy.

Due to this large cost of stress, already in 1999 in Belgium a legislation to prevent stress at work has been drafted ('Collectieve Arbeidsovereenkomst nr. 72' [11]). Also at European level a framework directive was drafted in 1989 to guarantee minimum safety and health requirements throughout European member states (Directive 89/391 EEC [12]). In both legislations, companies are encouraged to detect and evaluate employees' work-related stress.

Currently the most widely used method and current gold-standard to assess stress is by means of questionnaires, e.g. the Perceived Stress Scale [13].

However, these questionnaires are qualitative, time-consuming and reflect subjective responses collected during spot-checks. Therefore, research has focused on finding objective, continuous and quantitative physiological markers of stress [14], which change due to the sympathetic nervous system's (SNS) fight-or-flight response [15], i.e. the bodily response to stress.
The most-frequently investigated physiological signals for monitoring stress are the skin conductance (SC) (changes in SC by sweat-gland innervation), the electrocardiogram (ECG) (changes in heart rate (HR) and heart rate variability (HRV)), the electromyogram (EMG) (electrical activity of skeleton muscles), blood pressure (BP) and skin temperature (ST) [16]. These have shown to be reliable indicators of stress in laboratory conditions [16].
In recent years, the growing availability of wearable sensors has led to increased research towards continuous, ambulatory monitoring of stress. However, still many gaps in research can be identified. Most ambulatory studies have been executed on a small population (i.e. 20-50 participants). Although these studies provide valuable insights, in order to develop models that are generalizable on a large scale, large datasets are needed. Further, in the majority of ambulatory trials, participant background knowledge is not taken into account. It should be investigated how demographics (e.g. gender) and psychological baseline information (e.g. self-reported anxiety and depression levels) contribute to physiological stress detection. Also context information (e.g. location) is often ignored, although it could be used to provide much more actionable feedback [17].

Baring these observations in mind, the **goal of this research** is to **gain more insight in physiological sensing priorities and machine learning techniques for physiological stress detection and next to deploy these on a large population in real-life, ambulatory conditions.**
Three objectives for this thesis have been formulated:

1. *The identification of the **most suitable markers for physiological stress detection** in a controlled laboratory environment on healthy subjects, to translate this knowledge to the ambulant environment.*
    a. *Identify the best performing machine learning techniques, in terms of accuracy, for physiological stress detection*

b. *Evaluate if personalized models (i.e. one model per subject) outperform generalized models (i.e. one model for all subjects)*

First, we focus on the detection of stress in a controlled environment. We submit 20 healthy subjects to a stress test in laboratory conditions, allowing us to control stress and relax periods. We investigate the use of different physiological signals for stress detection and identify the most relevant physiological features.
Further, we compare several machine learning techniques for physiological stress detection and investigate if personalized models (i.e. one model per subject) outperform generalized models (i.e. one model for all subjects). These findings will aid in the selection of physiological signals for an ambulatory trial and provide insights in machine learning techniques for stress detection.

2. *The **differentiation between healthy subjects and patients** based on their physiological stress response, towards disease prevention and interception.*

The same controlled experiment on 20 healthy subjects is repeated for 12 patients with stress complaints. This is needed because in the future, it might be the goal to use continuous, ambulatory stress detection for disease prevention and interception. Therefore, first, the difference in physiological response between healthy subjects and patients must be established. We propose a new exploratory methodology that can be used to differentiate healthy subjects and patients based on their physiological response to a stress task in laboratory settings.

3. *The **large-scale** investigation of the physiological stress response in **ambulatory conditions**, including demographics and context information towards digital phenotypes for personalized and continuous stress detection.*
    a. *Set-up of a large-scale study (>1000 subjects) in the work environment, to grasp inter-subject variability*
    b. *Differentiate subjects according to their digital phenotypes for stress detection, towards personalized physiological stress detection*
    c. *Identify a personalized physiological calibration methodology using a short stress test, towards ambulant model performance improvement*

We present the SWEET study: world's largest ambulatory stress detection study. Over a period of more than two years, we included 1,002 subjects who were continuously monitored during 5 days using a wristband and chest patch for physiological sensing and a smartphone for annotations. A protocol is presented including physiological sensing, baseline psychological information, self-reported stress and contextual sensing based on smartphone information. In terms of size, this study is at least 10 times larger than any previous study related to ambulatory psychophysiological stress detection. In terms of scope, it features an unprecedented combination of multi-sensor data sources, allowing more insight into the link between physiological stress and subjects' context and background information.
Based on this information, we develop models for ambulatory, continuous stress detection and investigate digital phenotypes for physiological stress detection, i.e. how subjects with different demographics and context information differ in their physiological stress response.
Last but not least, we propose a methodology to use a short stress test to increase personalization and improve stress detection models performance.

By conducting these three studies, we aim to provide the research community with more insights on a) the capability of machine learning techniques for stress detection and physiological sensing priorities, b) the physiological difference in stress response between healthy subjects and patients and c) the potential of digital phenotyping for personalized stress detection in ambulatory conditions.
We hope these insights will form the basis for future research to enable highly personalized, just-in-time interventions for preventive health and stress reduction.

1.1. Chapter-by-chapter overview

In Chapter 2 we discuss the background related to the problem of stress and stress detection. We focus on definitions for stress and its physiological pathway. Further, we discuss different machine learning techniques and their advantages and disadvantages for physiological stress detection.

In Chapter 3 we focus on the first objective of the thesis: the identification of the most suitable markers for physiological stress detection in a controlled laboratory environment on healthy subjects. We present the study protocol, compare multiple machine learning techniques and investigate inter-person variability.
In Chapter 4 we focus on the second objective of the thesis: the differentiation between healthy subjects and patients based on their physiological stress response. We present the study protocol, introduce an exploratory methodology to calculate physiological features and develop a model to differentiate between healthy subjects and patients.
In Chapters 5-7 we focus on the third objective of the thesis: The large-scale investigation of the physiological stress response in ambulatory conditions, including demographics and context information towards digital phenotypes for personalized and continuous stress detection. In Chapter 5 we present the protocol of the SWEET study and results regarding compliance and data quality.
In Chapter 6 we use the data collected in the SWEET study, including demographics, psychological background information, physiological and contextual information to infer behavior patterns and digital phenotypes for stress detection. In Chapter 7 we introduce a methodology to use subject-specific information, based on the physiological response to the Montreal Imaging Stress Task (MIST), a short stress test which was conducted as part of the SWEET study, to improve ambulant classification performance.
In Chapter 8 we discuss the conclusions of this thesis and propose relevant future work.

Chapter 2: Background

This chapter explains the problem of stress and stress detection and discusses background research in this domain. First, we focus on the definition of stress. Second, the physiological pathway of stress is explained. Based on this pathway three stress detection techniques are identified and discussed: questionnaires, biochemical indicators and physiological indicators. Third, the most frequently used machine learning techniques for stress detection based on physiology are discussed. The content of this chapter is partially submitted to IEEE Journal of Biomedical and Health Informatics.

2.1. Definition of stress

Stress has first been described by Hans Selye [18] based on an experiment with rats. Selye found that rats, when exposed to a critical situation (e.g. cold or injury), express a typical reaction which he called the 'general adaptation syndrome' and later renamed as the 'stress response' [19]. Since then, many definitions of stress have been formulated. The Oxford dictionary currently defines stress as "a state of mental or emotional strain or tension resulting from adverse or very demanding circumstances". This definition however only focuses on the negative attributes of stress. To emphasize the fact that stress can be both positive and negative, Selye additionally defined 'eustress' or good stress and 'distress', depending on differences in the subject's perception and emotional reaction [19]. According to Selye, the individual defines whether the stressor causes eustress or distress. Another approach called the 'Yerkes-Dodson Law' states that increasing stress leads to increasing performance until some maximum point is reached, after which performance will decrease with increasing stress, also known as the inverted-U diagram [20] as can be seen in Figure 2-1.

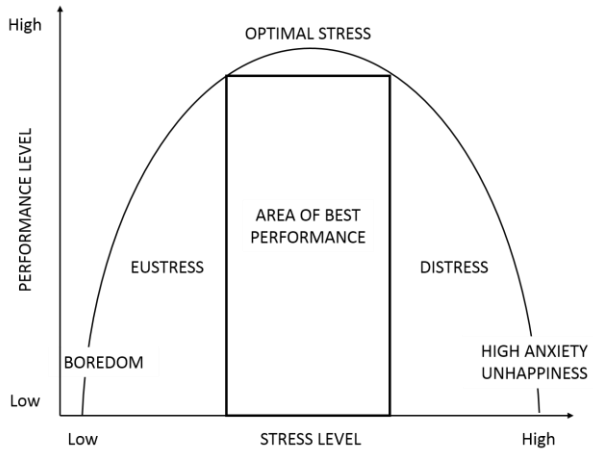


Figure 2-1: Yerkes-Dodson Law [21]

It is interesting to note that Yerkes and Dodson [22] in their original model never mentioned stress and performance levels. They observed an inverted-U relation between stimulus strength and habit formation of mice. Only later, Eysenck [23] hypothesised the relationship would hold true between anxiety and task performance in humans. However, there is no empirical evidence confirming this hypothesis, yet this model is often used in managerial psychology [24]. Research suggests the relationship of the Yerkes-Dodson law is too simplistic to account for the complex relationship between cognitive functions (e.g. performance) and emotional arousal (or stress) [25]. It is argued that context can play an important role and should be taken into account.

Besides these general definitions, also several models have been developed that aim to define stress in the work environment. A commonly used model is the Job Demands-Control (JDC) model of Karasek [26], also called the Job strain model (Figure 2-2). This model states that the combination of high job demands and low decision latitude results in high strain or stress (line A in Figure 2-2). Additionally, high demands combined with high levels of decision latitude result in learning and development of new behavior (line B in Figure 2-2). The model hypothesizes that job decision latitude (control) can moderate the negative effects of high demands on well-being (also called the 'buffer hypothesis').

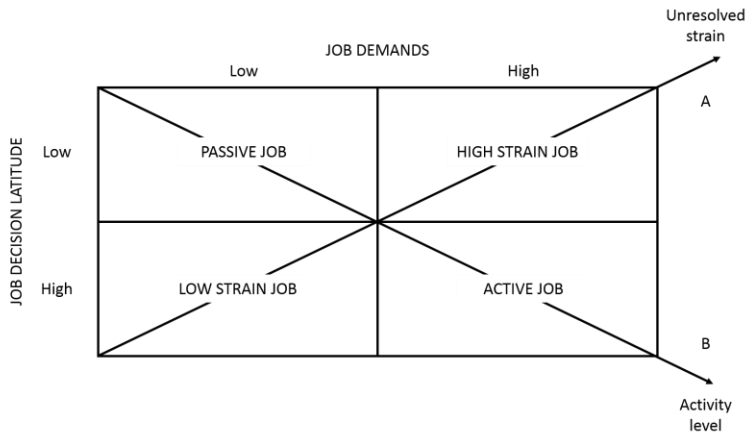


Figure 2-2: Job Demands-Control model of Karasek [26]

Later Johnson and Hall [27] added a social dimension to the model, resulting in the Job Demands-Control-Support (JDCS) model, stating that increased control buffers the negative effects of high job demands best under conditions of high social support [28]. The review of Häusser *et al.* [28] investigated 237 studies between 1998-2007 using the JDC model and found support for the model in 80 % of the studies. Additionally they compared 144 studies using the JDCS model, for which they found support in 61 % of the studies. It can be concluded that there is a strong evidence base for the JDC model.

Another widespread model is the Effort-Reward Imbalance (ERI) model of Siegrist [29] (Figure 2-3), which puts emphasis on rewards rather than on control. Rewards can be given in three ways: money, esteem and job security/career prospects. Additionally, also personal characteristics are included and the term 'overcommitment' is introduced. People who are characterized by overcommitment, have a strong desire of being approved and tend to exaggerate their efforts and underestimate the challenge [30]. A cross-sectional study with more than 11,000 participants shows independent cumulative effects of both the JDC model and the ERI model on employee well-being [30]. Additionally, the study shows that overcommitted people have higher risks at poor well-being.

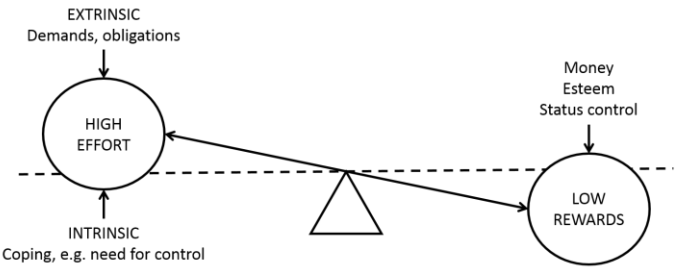


Figure 2-3: Effort-Reward Imbalance model of Siegrist [29]

Finally, Demerouti *et al.* [31] defined the Job Demands-Resources (JDR) model based on strengths and weaknesses of the JDC and ERI models. It is stated that a major weakness of the JDC and ERI models is that they are static, for the JDC model only autonomy is included as resource, while for the ERI model only money, esteem and status control are included. However, it could be possible that in different jobs, different types of resources might be more important. Therefore the JDR model was developed under the assumption that every occupation might have their own risk factors associated with job stress and their own resources [32]. The model states that high job demands lead to strain and impaired health, and that high resources lead to increased motivation and higher productivity (Figure 2-4). The results of 16 cross-sectional studies show strong evidence for the model's assumptions [33].

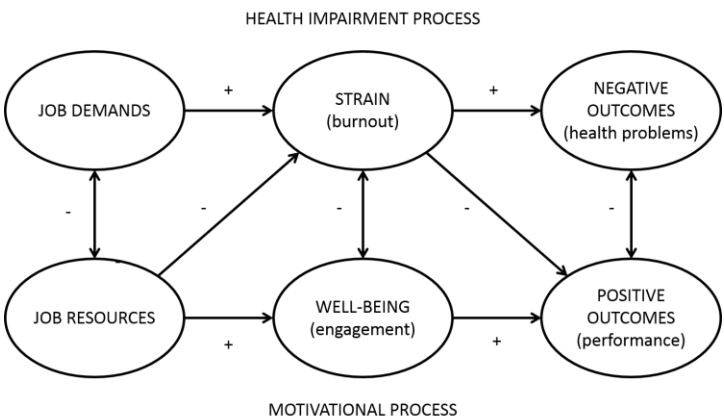


Figure 2-4: Job Demands-Resources model of Demerouti [31].
Signs indicate the direction of the effect, (+) indicates an increase, (-) a decrease, e.g. high job demands lead to higher strain which leads to more negative outcomes.

Overall, these models and definitions of stress can be subdivided into three categories. The first category mainly focuses on stressors as part of the environment, e.g. the JDC model of Karasek. The second category focuses on the individual response, including the physiological and subjective levels of distress, e.g. the definition of Selye. The third category defines stress as an interaction between subject and environment, e.g. the ERI model of Siegrist, including both extrinsic and intrinsic aspects of effort and rewards.
Literature is not conclusive on which model or category is most correct. In this thesis, the JDC model of Karasek is used to identify the subjective stress response, i.e. we investigate the physiological and subjective effect of different stressors, changing the environmental demands, on a subject.

2.2. The physiological stress response

The human body responds to stress using two response systems: a fast response to sudden stress following the sympathetic adrenal medullary (SAM) axis and a slow response to chronic stress following the hypothalamic pituitary adrenal (HPA) axis [34].
The autonomic nervous system (ANS) exists of three subsystems: the sympathetic nervous system (SNS), the parasympathetic nervous system (PNS) and the enteric nervous system. Most tissues are innervated by both SNS and PNS with opposing effects as can be seen in Figure 2-5.

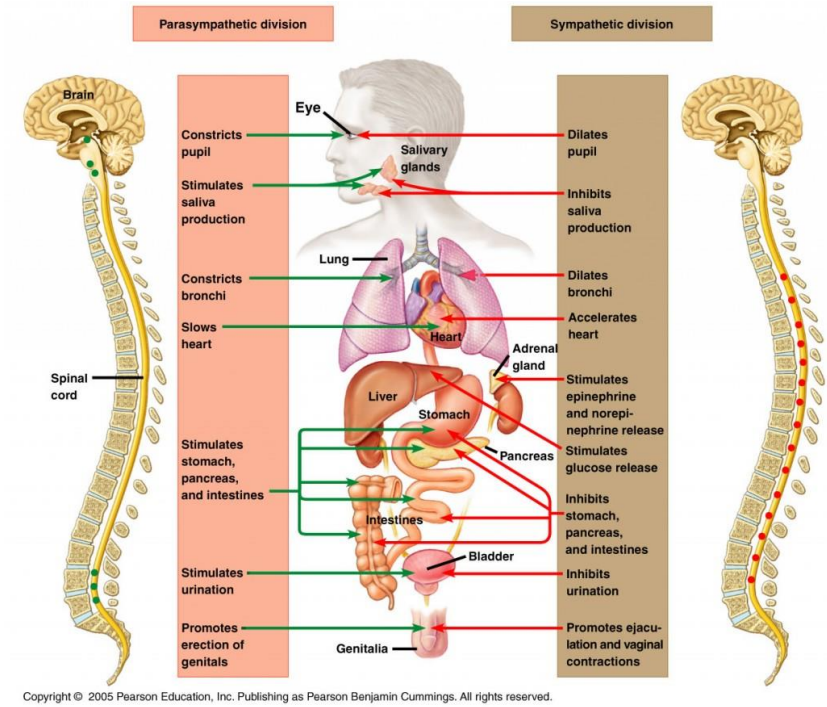


Figure 2-5: The Autonomic Nervous System (ANS) [35]

When the brain encounters a stressor, i.e. an internal or external stimulus that disrupts the body's internal balance, the paraventricular nucleus (PVN) of the hypothalamus will activate the SNS. The SNS in its turn will signal the adrenal medulla to secrete epinephrine and norepinephrine (SAM axis) [15]. This activation results in an increase of heart rate, blood pressure, pupil dilation, etc., which is sustained by the presence of epinephrine and norepinephrine. The body is preparing for a 'fight-or-flight' response. Simultaneously, the PVN will release two hormones: corticotropin-releasing factor (CRF) and arginine vasopressin (AVP) [36]. Both hormones are sent to the pituitary gland (hypophysis) through blood vessels, where they stimulate the production and secretion of adrenocorticotropic hormone (ACTH). ACTH in its turn induces the synthesis and release of glucocorticoids from the adrenal cortex (HPA axis). In humans the most important glucocorticoid is cortisol, which has a wide array of regulatory influences. It plays a key role in the central nervous system (CNS), where it is involved in learning, memory and emotion regulation; in the metabolic system, where it regulates the use and storage of

glucose; and in the immune system, where it regulates the magnitude and duration of the inflammatory response [37]. Cortisol levels reach a peak in the blood about 30 minutes after acute stress exposure [38].
The stress response system is regulated by a negative feedback loop at the level of the pituitary, where it reduces the release of ACTH; the hypothalamus, where it reduces the activity of the PVN; and the hippocampus, which has a stimulating effect on the production of CRF in the absence of cortisol, this activity is depressed in the presence of cortisol [15]. Through the reduced activity of the PVN both SAM and HPA axes are attenuated and levels of epinephrine will decrease. It is important to notice that the main driver of the fight-or-flight response is the SAM axis and presence of epinephrine. The HPA axis and presence of cortisol do not stimulate the stress response, but rather regulate it. Without cortisol no negative feedback loop would be possible and stress responses would have damaging effects on the body [15].

2.3. Methods for stress detection

2.3.1. Questionnaires

Currently the most widespread method and de facto the gold standard to measure stress is by means of questionnaires, e.g. the Job Content Questionnaire [39]. These questionnaires are qualitative, time consuming, conducted on spot-check basis only and their answers are subjective. Emotional arousal does not necessarily reach the level of consciousness and the extent to which it does and whether one wants to share this information can vary from person to person [40]. Therefore the answers of employees to these questionnaires are not always a good representation of their wellbeing in the organization [41]. Verkuil *et al.* [42] suggest that persons with lower emotional awareness might have to rely on different indicators of stress compared to persons with higher emotional awareness. Furthermore, even when a person is consciously aware of and willing to share his stress levels, questionnaires are subject to recall bias ("a systematic error due to differences in accuracy or completeness of recall to memory of past events or experiences" [43]) and in repeated assessments often within-person variability over time and across contexts is lost [44]. A possible solution to this problem is provided by ecological momentary assessments (EMAs). EMAs capture real-

time data on momentary states in the natural environment, often using electronic diaries, e.g. on a smartphone [44]. These questionnaires are sent at random or specific times throughout the day to prompt the subject for immediate feedback. However, these prompts may be experienced as highly interruptive and become a source of stress itself [45].

2.3.2. Biochemical stress indicators

Research has focused on finding objective, non-intrusive, continuous and quantitative ways to detect stress. Based on the physiological pathway of stress (see 2.2) both biochemical and physiological indicators can be used for stress detection.

The most common biochemical stress indicator is cortisol, which can be measured in the blood, sweat, tears, urine or saliva for acute stress detection [46]. Measurement techniques are tedious and time-consuming. To obtain saliva samples, subjects need to swap their mouths or spit into a container, once up to eight times a day, depending on the experimental set-up of the study [47]. Next, these samples are analyzed in a laboratory using immunoassays, i.e. biochemical tests [48] to detect cortisol through the use of antibodies. This is a challenging approach since many steroids are structurally similar to cortisol, which makes the selection of entirely specific antibodies difficult [48]. For example, in saliva the presence of cortisone can significantly reduce the specificity.

In several studies salivary cortisol is also used to assess chronic stress, however, long-term exposure is difficult to evaluate due to inherent circadian variations of cortisol (i.e. morning increase and evening decrease), therefore hair cortisol (i.e. cortisol extracted from the roots of the hair) is suggested for measurement of chronic stress [46]. Less frequently epinephrine and salivary alpha amylase, both markers for SNS activation, are measured for acute stress detection [49]. Review analyses have shown that acute stressors can elicit a cortisol response [50]. However, results are rarely consistent and there is a substantial degree of variability in the magnitude of the cortisol effect, which varies depending on the characteristics of the stressor. A review of 208 studies has shown that stressors high in uncontrollability as well as social threats cause a stronger cortisol response [50]. Further, the cortisol stress reactivity can also vary across different psychiatric disorders and gender, e.g. women with a

major depressive disorder or an anxiety disorder show blunted cortisol stress responses to psychological stress, whereas men show increased cortisol responses [51]. Although cortisol is often suggested as stress detection technique, it still cannot be measured in a continuous, non-intrusive way. Therefore research has investigated physiological stress detection.

2.3.3. Physiological stress indicators

Research focusing on the physiological detection of stress has been conducted under different types of conditions. Historically, most research has been conducted in laboratory settings (Figure 2-6a), where both stressor (timing, frequency, duration) and context can be rigorously controlled [52]. With the increasing use of wearables, many opportunities are emerging for a continuous, ambulatory monitoring of stress and research in this field has increased substantially over the last five years. In ambulatory monitoring two types of conditions are used, either context specific (Figure 2-6b), e.g. stress monitoring while driving a car [53] or in a call center [54], or daily living settings (Figure 2-6c) in which different types of stressors can influence the subject and the context of the stressors is unknown.

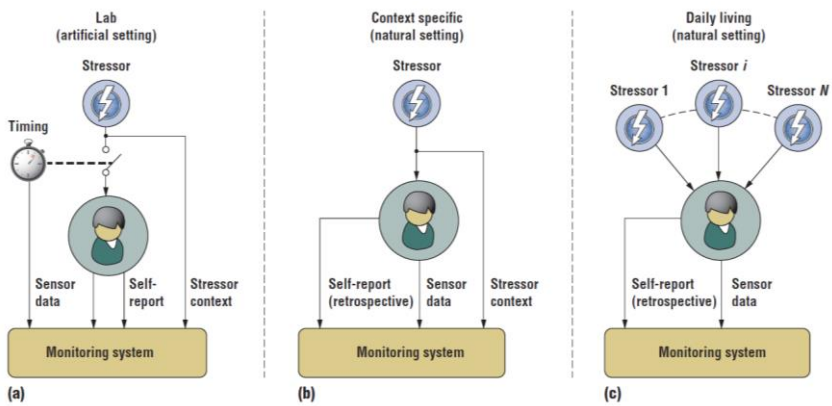


Figure 2-6: Conditions for physiological stress detection [52]

In controlled settings the potential of physiological signals to detect psychological stress has already widely been demonstrated. Several stress inducement stimuli have been used, e.g. the Stroop color word test, mental

arithmetic, public speaking, computer work or a cold pressor test [55]. Also multiple physiological signals have been investigated, the most commonly studied signals are explained in further detail below. In Table 2-6 an overview of current research on physiological stress detection is presented.

Electrocardiogram

The electrocardiogram (ECG) measures the electrical activity of the heart. A sample from a typical ECG is depicted in Figure 2-7. The P-wave is associated with the contraction of the atria. The QRS complex is associated with the contraction of the ventricles. The T/U waves are associated with the repolarization of the ventricles [56]. The R-R distance (or R-R interval) is the time between two R peaks and is used in the calculation of the heart rate (HR) and heart rate variability (HRV). ECG signals from different individuals can exhibit personalized traits such as the relative timing of the peaks but can also exhibit responses to stress and activity.

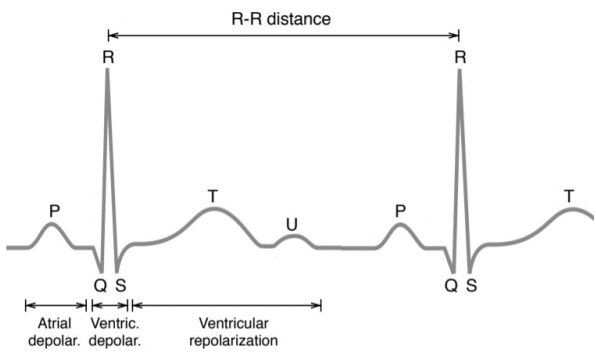


Figure 2-7: Sample of a normal ECG [56]

In resting conditions the heart is under inhibitory control of the PNS. This results in a low HR and high HRV, since the PNS adapts the HR to the breathing phase (inspiration versus expiration), i.e. Respiratory Sinus Arrhythmia (RSA) [57]. In stressful situations, PNS control is decreased resulting in a disinhibition of the SNS and an increased SNS activation through the SAM pathway (see 2.2), which causes the HR to increase and the HRV to

decrease since SNS modulation of HR reacts too slow to respond to the respiratory phase.
HR and HRV have already widely been investigated in stress-related research. Karthikeyan *et al.* [58] found a 79.17% classification accuracy for binary stress classification (i.e. rest versus stress) based on HRV. Gomes et al. [59] found a 68% association of HRV with self-perception of stress for fire fighters in real emergency situations and Michels *et al.* [60] and Rieger *et al.* [61] found significant correlations of HRV with perceived stress. Widjaja *et al.* [62] showed that by removing the influence of the RSA (i.e. the influence of respiration on HR and HRV), the classification accuracy for a binary stress detection could be improved from 57.13% to 97.88%.
HR and HRV are also influenced by physical activity [59], drugs (e.g. beta-blockers, estrogens, vitamin E and coenzyme Q10 increase HRV, progestins and nifedipine decrease HRV [63]), nutrient uptake (e.g. fasting increases HRV [63], sodium restriction decreases HRV [64]), personal attributes (e.g. smoking and alcohol intake reduce HRV [65], evening types have a lower HRV than morning types [66]) and time of the day (HRV is influenced by a circadian rhythm with high HRV before waking up and lower HRV after [67]).
In general, the measurement of HR and HRV is noninvasive and easy to perform, it has a relatively good reproducibility and there is already a large body of research supporting the link with psychological stress [65].

Skin conductance

Skin conductance (SC), also called, galvanic skin response or electrodermal activity, reflects the production of sweat caused by activation of the SNS. Three types of sweat glands exist: eccrine, apocrine and apoeccrine [68] [69] [70]. The eccrine glands are already present at birth and are located over the entire body, although most concentrated in forehead, palms, soles, axilla and scalp (e.g. 600-700 glands/cm^2 on palms compared with 64 glands/cm^2 on the back [69]). The apocrine glands are small and inactive until puberty, then they become large and secrete a solution thicker than sweat, localized at axilla, areola of the nipples and perineum. The apoeccrine glands combine the characteristics of eccrine and apocrine glands. There are two types of sweating: thermoregulatory, in which eccrine glands over the whole body are

involved, and emotional, in which apocrine and eccrine glands in face, axilla, palms and soles are involved [68]. Emotional sweating is controlled by the SNS, which normally has norepinephrine as a postganglionic neurotransmitter, but for sweat production this is acetylcholine (normally the postganglionic neurotransmitter of the PNS). Wilke *et al.* [70] stated that eccrine sweat glands are also triggered by norepinephrine but that this effect is much smaller than acetylcholine. In Figure 2-8 the pathway of thermoregulatory and emotional sweating is presented. The left part of the figure represents the input of thermoregulatory sweating, for emotional sweating the input is stress at the level of the brain. These inputs are both processed in the brain and a response is sent to the sweat glands through the neurotransmitter acetylcholine.

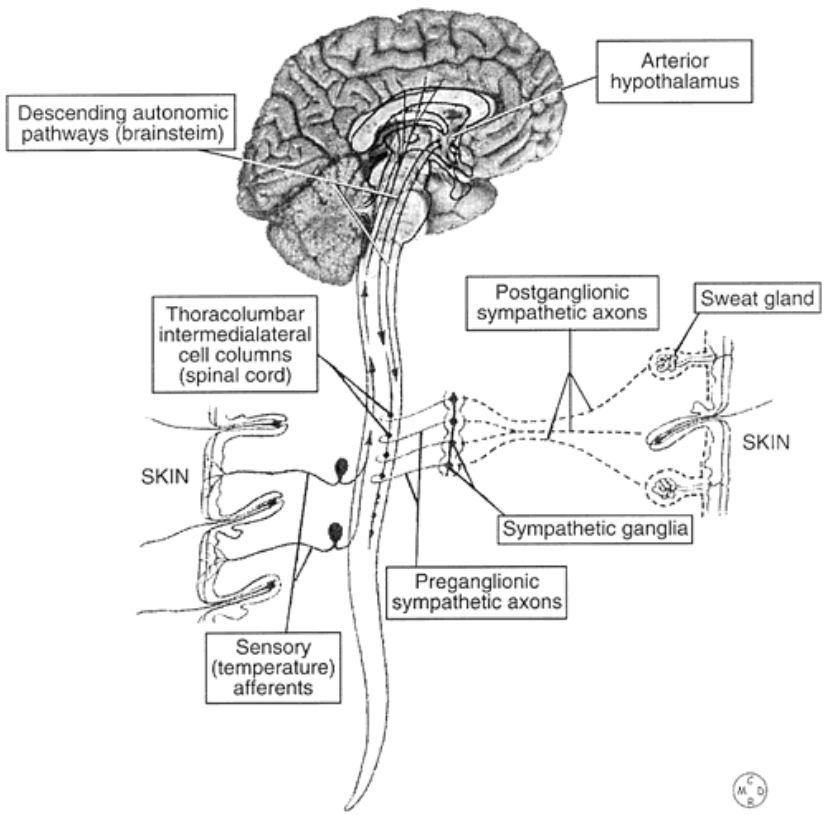


Figure 2-8: Pathway of thermoregulatory and emotional sweating [71]

Van Dooren *et al.* [69] stated that the best locations to measure SC are at the fingertip, foot and shoulders. However, these locations are not practical in

terms of user comfort for ambulatory studies. Therefore, in practice mostly the wrist is used to measure SC in ambulant conditions.

Multiple studies have already indicated the correlation between SC measurement and stress. Svetlak *et al.* [72] showed that 93.4% out of 106 subjects had a higher SC during a Stroop task on left and right hand compared with the rest situation. 88.7% of the subjects had a higher SC on the left hand and 89.6% on the right hand. Villarejo *et al.* [73] found a 76.56% success rate on detecting different stress states based on a newly developed sensor using SC. A comparison of healthy subjects and subjects with posttraumatic stress disorder (PTSD) showed that both groups respond to a stressor with increased SC with larger responses for subjects suffering from PTSD [74].

Several contextual conditions can influence the SC signal. Contradicting findings exist on the influence of temperature on SC. Vetrugno *et al.* [71] stated that high ambient temperatures increase the effect of stress on SC and low ambient temperatures can erase the effect of stress on SC. However, Harrison *et al.* [75] found no significant difference in SC between infants based on temperature. Further several drugs can block the effect of stress on SC (e.g. Botulinum toxin [71]), but as opposed to HR and HRV, beta-blockers have no significant effect on SC [76], which makes it an interesting signal for people suffering from a heart condition. Further, absolute values of SC can differ largely among persons [77] and depending on the handedness of a person the SC signal can differ significantly between left and right hand [78], although this is contradicted by the findings of Svetlak et al. [72]. Finally, although Vertrugno *et al.* [71] indicated that different sweat glands are activated due to physical activity as compared to stress, Villarejo *et al.* [73] stated that in practice it is difficult to differentiate a sweat response as being due to stress or due to physical activity.

In general, SC has widely been used to detect stress and a positive link has been established. It is a relatively easy to measure signal with a good reproducibility. Due to its independence of circulatory regulations it can provide complementary information with HR and HRV, especially for subjects suffering from heart conditions.

Electromyography

The electromyography (EMG) measures the electrical activity of the muscles. During the stress response the SNS will be activated and prepare for the fight-or-flight response, therefore energy will be mobilized to the muscles. The EMG shows a higher muscle activity during stress as compared to rest situations. A typical location to measure muscle activity due to stress is in the upper trapezius muscle (between neck and shoulder) [79].
Wijsman *et al.* [80] showed that the EMG amplitudes of the trapezius muscles are 2.6% higher during stress compared to rest and there is 14.3% less time between EMG gaps during stress. Further, EMG correlated significantly with subjectively indicated stress levels [80]. Karthikeyan *et al.* [79] found a classification accuracy of 71.25% on distinguishing binary stress levels in a laboratory environment, based on trapezius muscle EMG. Finally, Schleifer *et al.* [81], Lundberg *et al.* [82] and Krantz *et al.* [83] found a significant increase in trapezius muscle EMG activity when mental stress is present.
Not all muscles are equally responsive to stress, with trapezius and facial muscles being the most sensitive locations [84]. Further also physical activity [80] and fatigue [85] have a strong influence on EMG. Veldhuizen *et al.* [85] found that EMG activity decreases during the workday due to fatigue. Finally, also personal attributes such as gender and trait anxiety can influence EMG.
Although EMG has less been studied than HR and SC, it has been shown to be a reliable measure for physiological stress in laboratory conditions. It is especially an interesting measure since neck pain is a frequent cause of absenteeism in office workers. 30% of office workers report back pain, of which neck and shoulder are the most affected areas [86]. Since the main causes of pain include biomechanical exposure and stress [87], the measurement of EMG at the trapezius muscle and the reduction of stress could provide benefits in prevention of these musculoskeletal disorders [80].

Other physiological signals

Blood pressure (BP) increases when the SNS is activated. Under normal conditions BP is controlled through the baroreflex feedback system, as represented in Figure 2-9. When the BP increases in the vessels (1), this is sensed by the baroreceptor (2) and the signal is sent to the vasomotor center

in the brain (3). This causes a decrease in activity in the sympathetic nerve (4) which leads to a reduced HR and therefore a reduced BP in the vessels (5) [88]. In stressful situations the SNS is activated and the baroreflex is, temporarily, overruled. After some time (4 min according to Zhao *et al.* [89]) sympathetic activity decreases and the baroreflex is again in control of BP.

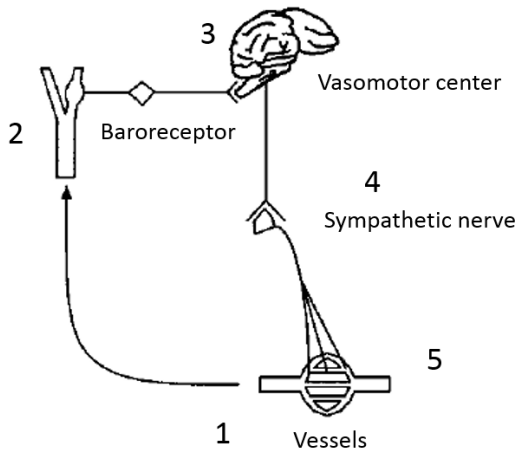


Figure 2-9: Baroreflex control of BP [88]

Carroll *et al.* [90], Zhao *et al.* [89] and Matthews *et al.* [91] all found a significant increase in systolic (pressure when the heart beats) and diastolic (pressure between heart beats) BP to a mental stress task. However, most of these studies are longitudinal (e.g. Carroll *et al.* [90]), to investigate the negative effect of stress on high BP and therefore high risk at cardiovascular diseases. Few studies focus on the effect of acute stress on BP. A recent study of Ottaviani *et al.* [92] showed a significant increase in BP after a perseverative cognition task, i.e. rumination about the past and worrisome thoughts about the future, and after a problem solving task, using a continuous BP cuff. The main reason why BP is not frequently measured is the lack of a practical measurement system. In ambulatory conditions, subjects still have to wear a device on the waist connected via tubing to an inflatable cuff on the arm [93]. Another less frequently used signal, although easy to measure, is skin temperature (ST). When the SNS is activated, ST will decrease due to vasoconstriction. Kistler *et al.* [94] state that when fingertip temperature is initially higher than 32°C and the vasoconstriction due to sympathetic activation lasts longer than 5 s, fingertip temperature decreases with a lag

phase of 15 s. Karthikeyan *et al.* [95] found a 75.32% classification accuracy based on armpit ST measurements in stress and no-stress classes.
Also pupil diameter and eye blinks can be used to detect stress. This is relatively easy to measure when subjects perform computer work (by use of a camera), but it is less straightforward to measure in daily life settings. Partala and Surakka [96] state that pupil size increases significantly in both negative and positive emotional states compared with neutral situations. Zhai and Barreto [97] found a stress classification accuracy based on measurements of blood volume pulse, SC, pupil diameter and ST of 90.10%, leaving pupil diameter out this drops to 61.45%. Wilson [98] found that eye blink rate tends to decrease with increased visual demands in a dynamic flight environment for pilots (i.e. a stressful situation).

Combining physiological signals

Previous research discussed mainly laboratory studies using only one physiological signal for stress detection. However, in recent years the focus has shifted towards multi-sensor, ambulatory research.
First, the advantage of the multi-signal approach has been investigated in laboratory environments. In many studies HRV and respiration rate are measured together to control the HRV signal for respiratory influences [53] [99] [100] [101]. In many other studies a combination of HR, SC and/or EMG is used to improve classification accuracy. Sandulescu *et al.* [102] used SC and HRV to detect binary stress levels of subjects during the Trier social stress test (TSST) (i.e. a public speaking task) and reached a classification accuracy of 79% in 0.1 s time intervals. De Vries *et al.* [103] combined SC, HR and respiration signals to classify binary stress levels in 5 min intervals with a classification accuracy of 88%. De Santos Sierra *et al.* [104] even found classification accuracies of more than 99% for a binary stress classification using HR and SC signals in 10 s intervals. However, this large accuracy is probably due to the choice of stressor, i.e. a hyperventilation task, which is rather a physiological challenge than a psychological challenge such as the TSST.
Also promising classification accuracies have been obtained in ambulatory settings. The research of Healey and Picard [53] in 2005 was one of the first studies to leave the laboratory and do measurements in a real-world driving task. Although participants still had to drive on a set route, this was a first step

towards physiological stress detection in uncontrolled conditions. They reached a classification accuracy of 97% in a 5 min interval based on HR, SC, EMG and respiration. The main challenge to detect stress in ambulatory conditions is the presence of physical activity which can mask the effect of stress on the physiological signal(s) [105]. Additionally, ambulatory measurements are more susceptible to (motion) artefacts, which imposes the need for accurate signal processing techniques and signal quality indicators [106]. A new approach, combining information from laboratory and ambulatory settings, was suggested in the cStress model [107]. In this approach, first a model per subject is trained based on physiological data (HR and respiration) during a laboratory stress test. Then, this model is applied on ambulatory data, using the acceleration signal to include the physiological changes due to physical activity. An average classification accuracy for 20 subjects of 72% was reached, classifying into binary stress levels in 1 min time intervals.

In recent research not only physiological signals, but also smartphone information is used to increase classification performance. Muaremi *et al.* [17] measured HRV and smartphone features (e.g. GPS location, microphone, calls, battery status, etc.) of 35 employees in their daily lives during 4 months. They used self-reported stress from EMAs to classify each day into three stress levels and found a classification accuracy of 55% using only smartphone-related features, 59% using only HRV features and 61% using a combination of both.

2.4. Overview of machine learning techniques

Different machine learning (ML) techniques have been used in literature to predict mental stress based on physiology. Below the most frequently used techniques are described.

2.4.1. Logistic regression

In logistic regression (LR) the probability of the outcome of the stress vs rest classification is modeled as a function of the features weighed by coefficients [108]. A classical linear regression aims to explain a dependent variable as a function of multiple independent variables. The model is estimated with

ordinary least squares (OLS), a technique to select the model parameters which minimize the sum of squared errors between true and predicted outcome [109]. However, OLS can only be used under certain assumptions such as homoscedasticity (i.e. "the error term is the same across all values of the independent variables" [110]), linearity and normality. These assumptions are violated when the output variable has only two or three response categories (e.g. stress vs rest). Therefore, LR is a variation of linear regression, using the maximum likelihood estimation, transforming the dependent variable into a logit (log of the odds of falling into the "1" category, i.e. stress) [111]. Where the "maximum likelihood estimates yield parameters such that the likelihood of being able to use those parameters to replicate the actual data is maximized" [109]

The difference between the linear model, where b_0 and b_1 are estimated using OLS, and a logistic model, where b_0 and b_1 are estimated using maximum likelihood, is presented in Figure 2-10. The y-axis represents the probability of a point with independent variable x to belong to the "1" class (i.e. stress). The weights $b_0, b_1, ..., b_n$, where n represents the number of features, can be used to identify feature importance. The higher absolute value of the feature weight, to more important the feature. This information can provide additional insight in which features are more or less influenced by stress.

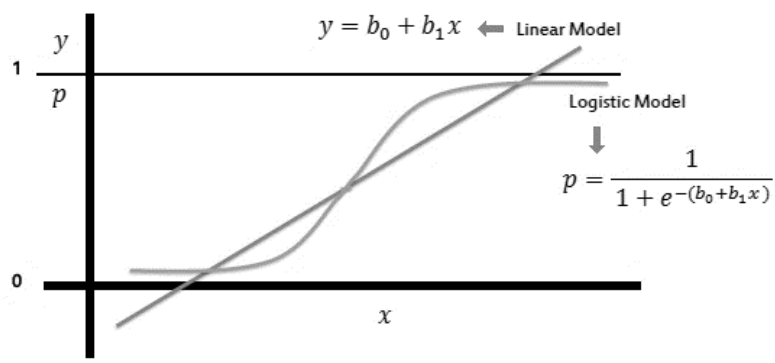


Figure 2-10: Logistic regression [112]

LR is one of the most common approaches for classification problems. They are easily interpretable, but can underperform when decision boundaries are non-linear or more complex [113].

Table 2-1: Strengths and weaknesses of logistic regression [113]

Strengths	Weaknesses
- Provides estimates of the strength of the relationships among features and the outcome	- Only works with numeric features, so categorical data requires extra processing - Tends to underperform when there are multiple or non-linear decision boundaries.

2.4.2. Support vector machines

Support vector machines (SVMs) search for an optimal hyperplane to separate the data between features of stress and rest [114]. SVM uses a geometrical transformation that projects the features into an infinite dimensional space where a linear separation is found. The optimal hyperplane separates the positive from the negative examples with the largest margin (see Figure 2-11). The training points that lie exactly on the edges of the margin and whose removal would change the solution found, are called support vectors (indicated with arrows in Figure 2-11) [114].

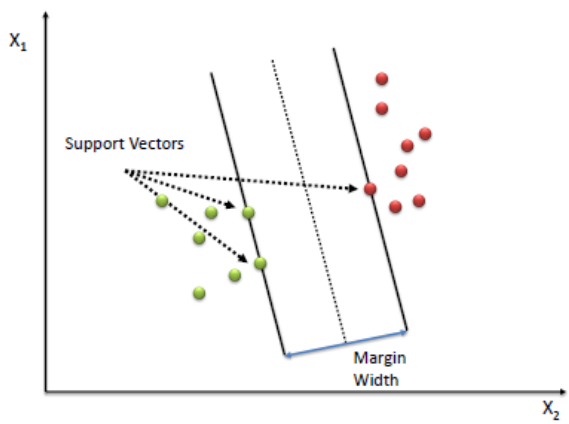


Figure 2-11: Support vector machines [115]

For linearly separable data quadratic optimization is used to find the largest margin [113]. If the relationship between variables is non-linear, kernels can be used, to map them into a higher dimensional space into linearly separable

classes. Essentially, the kernel trick involves a process of adding new features that express mathematical relationships between different variables [113]. These are powerful classifiers, although they have some drawbacks such as high computational complexity and being black box models, which means the structure behind the model and the importance of features cannot be interpreted [113]. An overview of advantages and drawbacks is presented in Table 2-2.

Table 2-2: Strengths and weaknesses of kernel-based SVMs [113]

Strengths	Weaknesses
- Not overly influenced by noisy data and not very prone to overfitting - Gaining popularity due to its high accuracy and high-profile wins in data mining competitions	- Finding the best model requires testing of various combinations of kernels and model parameters - Can be slow to train, particularly if the input dataset has a large number of features or examples - Results in a complex black box model that is difficult if not impossible to interpret

2.4.3. Decision trees and random forests

Decision trees (DTs) learn structures underlying the data using hierarchical partitioning [116]. Nodes of the tree represent splits, which test the value of an expression of the attributes. The final branches (i.e. leaves) represent the outcomes of the test. Each leaf node has a class label associated with it. DTs are built based on recursive partitioning, also known as the 'divide and conquer' approach, because based on the feature values the data is consecutively split into smaller subsets of similar classes [113]. From the root node, representing the entire dataset, the algorithm chooses the feature which is most predictive of the target class to make the first split. It continues this approach until the stopping criterion is reached. This might occur at a node if all of the examples at the node have the same class, all the features have been used or the three has reached a predefined size limit [113].

Essentially, a DT represents a flowchart which makes it highly traceable and applicable in scenarios where the classification mechanism needs to be transparent. Although for stress detection this is no requirement, the insight of features used in the model and importance of the features can generate additional knowledge on the mechanisms of stress. Other strengths and weaknesses of DTs are presented in Table 2-3.

Table 2-3: Strengths and weaknesses of DTs [113]

Strengths	Weaknesses
- Highly-automatic learning process can handle numeric or nominal features, missing data - Uses only the most important features - Can be used on data with relatively few training examples or a very large number - Results in a model that can be interpreted without a mathematical background (for relatively small trees)	- DT models are often biased toward splits on features having a large number of levels - It is easy to overfit or underfit the model - Small changes in training data can result in large changes to decision logic - Large trees can be difficult to interpret and the decisions they make may seem counterintuitive

The classification performance of DTs can significantly be improved by growing an ensemble of trees and letting them vote for the most popular class [117]. This ensemble of trees is also called a random forest (RF), where different techniques can be used to grow the ensemble. A frequently used technique is bootstrap aggregating (bagging), where each tree is built using a random selection of data and features [117]. This way the variance of the results can be reduced. Another often used technique is boosting (e.g. AdaBoost algorithm), which is similar to bagging, but in a sequential approach where data points that were misclassified in the previous tree have a higher chance to be selected to build the next tree.
An important characteristic of the RF is the number of estimators (i.e. trees) that are grown. More trees results in higher accuracy, but also longer

computational time. Furthermore, there usually is a threshold after which the accuracy no longer increases with increasing number of trees. Therefore usually the out of bag classification error is defined for different numbers of trees, using a subset of the training data. The ideal number of trees is at the cut-off point after which the out of bag error does not reduce significantly anymore by adding more trees. Although RFs are less transparent than DTs, still feature importance can be investigated, by averaging the relative contribution of each feature to the different DTs in the RF [118].

2.4.4. Artificial neural networks

Artificial neural networks (ANNs) are based on our biological understanding of how neural networks work in our brain [113]. Where our brain exists of a network of interconnected cells (i.e. neurons), an ANN uses a network of artificial neurons (i.e. nodes). In the brain, incoming signals are processed by the cell's dendrites, through a chemical process which allows the cell to weigh the importance of the impulse. Different impulses accumulate in the cell and when a threshold is reached, it fires an output which is transmitted via an electrochemical process down the axon. At the axon's terminal the signal is processed and passed to the neighbouring neurons through the synapse. Similarly Figure 2-12 represents an ANN where the dendrites (i.e. the nodes) receive inputs X_1, X_2 and X_3, which are weighed by W_1, W_2, and W_3 and summed in the node. The signal is passed on according to an activation function f to generate output y [113].

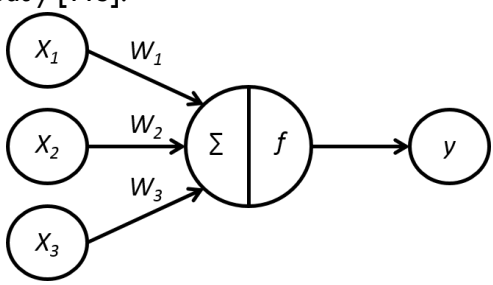


Figure 2-12: Artificial neural network [113]

ANNs are mainly defined by three characteristics: an activation function, which transforms the combination of input signals to one output signal that is transferred further in the network; a network topology, describing the number

of nodes, layers and the manner in which these are connected; and a training algorithm based on which the weights are defined [113]. The most commonly used training technique is back-propagation, which iterates through multiple cycles of a forward phase, in which neurons are activated from input to output signal through the different layers, and a backward phase, in which the output signal is compared to the true value and the error is propagated backwards through the network to update the weights [113]. ANNs are very strong ML techniques, although they also contains some weaknesses such as being computationally intensive and a black box model (Table 2-4).

Table 2-4: Strengths and weaknesses of ANNs [113]

Strengths	Weaknesses
- Among the most accurate modeling approaches	- Reputation of being computationally intensive and slow to train, particularly if the network topology is complex
- Makes few assumptions about the data's underlying relationships	- Easy to overfit or underfit training data
	- Results in a complex black box model that is difficult if not impossible to interpret

2.4.5. Naive Bayes and Bayesian networks

Bayesian methods are based on the idea that the estimated likelihood of an event should be based on the available evidence across multiple trials [113]. Therefore, the relationship between different dependent variables (e.g. stress and HR) can be described using the Bayes' theorem as shown in Eq.(1.1). This theorem states that the probability of an event *A* to occur given that event *B* has occurred (i.e. a conditional probability) equals the proportion of trials in which *A* occurred together with *B* out of all trials in which *B* occurred. For our stress example this means that the probability of having stress (*A*) while having a high HR (*B*) equals the proportion of trials in which subjects had high stress while having a high HR out of all trials in which subjects had a high HR.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{Eq.(1.1)}$$

After applying some algebra, Eq.(1.1) can be reformed to Eq.(1.2). In this equation the probability of the event *B* to occur given that *A* has occurred (i.e. $P(B|A)$) is called the likelihood, which can be defined based on a frequency table in which this combination of events can be counted based on examples (trials). The probability that *A* occurs is known as the prior probability (e.g. the general probability that a subject is stressed). The probability that *B* occurs is known as the marginal likelihood (e.g. the general probability that a person has a high HR).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{Eq.(1.2)}$$

The Naive Bayes algorithm is the most common application of the Bayes' theorem, frequently used in text classification problems (e.g. identifying spam e-mails based on content). The largest weakness is that the method assumes that all features are independent and equally important, which in most real-world problems is not the case [113].
Other strengths and weaknesses of this algorithm are listed in Table 2-5.

Table 2-5: Strengths and weaknesses of Naive Bayes algorithms [113]

Strengths	Weaknesses
- Easy to obtain the estimated probability for a prediction	- Relies on an often-faulty assumption of equally important and independent features
- Does well with noisy and missing data	- Not ideal for datasets with many numerical features
- Requires relatively few examples for training, but also works well with very large numbers of examples	- Estimated probabilities are less reliable than the predicted classes
- Simple, fast and very effective	

Bayesian networks (BNs) are more complex models based on the Bayes' theorem. These are directed acyclic graphs, where each node represents a random variable (e.g. the features and stress levels), and edges represent direct correlations between the variables. Each node is characterized by a conditional probability distribution of the variable given its parents [119]. BNs are static or dynamic. Dynamic BNs are identical to the static BNs, but additionally model the temporal relation of variables [120]. Therefore an additional edge is placed between the stress level at time t-1 and time t. An example of a static BN is presented in Figure 2-13, which represents the BN developed by Sun *et al.* [105] to represent the combined influence of physical activity and mental stress on physiology.

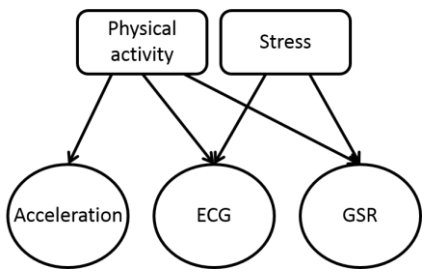


Figure 2-13: Example of a static Bayesian network [105]

2.4.6. Comparison of classification techniques

The results of stress detection, obtained with different classification techniques, are difficult to compare, since across literature also protocol, physiological signals used, etc. might differ. In Table 2-6 an overview of papers focusing on physiological stress detection and machine learning is presented. It is a representative yet not exclusive list of current research using physiological sensors for stress detection up to September, 2018. Most frequently used techniques are SVM [54] [97], BN [99], decision trees and RFs (most frequently AdaBoost) [105]. Less frequently used approaches include Fisher projection and linear discriminant analyses, and neural networks.

Several studies have compared the classification performance of different models. In their review, Sharma and Gedeon [16] concluded that best accuracies could be obtained using SVMs and ANNs. Han *et al.* [121] compared the classification accuracies of four models for a three-class stress classification: SVM, Linear Discriminant Analysis (LDA), Adaboost and K-

Nearest Neighbors (KNN). The highest classification accuracy of 84% was reached for the SVMs, followed by LDA with 80%, Adaboost with 79% and the lowest accuracy of 72% was reached for KNN [121]. Similar models were compared by Mozos *et al.* [122] where highest accuracies for a binary stress classification were reached for Adaboost (94%) and SVM classifiers with a radial basis function kernel (93%). The lowest accuracy was reached for KNN (87%) and SVM classifiers with a linear kernel (85%) [122]. Sun *et al.* [105] compared the classification accuracies for a binary stress classification of SVM, BN and decision trees and found the highest accuracy of 92.4% for decision trees.

Based on these results it is not clear which machine learning technique is most appropriate for physiological stress detection. Additionally, in most comparisons, only classification performance is used to compare different models. However, also other considerations should be taken into account. For example, techniques such as SVMs and ANNs are black box models and do not provide insight in *how* the link between physiology and stress is established. This could be interesting to investigate in order to formulate new hypothesis for psychological research based on artificial intelligence (AI). Further, computational complexity is another important aspect to take into account when deciding on which ML technique to choose. This is often overlooked in laboratory research, but becomes more important when developing models for ambulatory research which needs to make real-time predictions and which might be deployed on smartphones or smart watches instead of on computers. In general SVMs and ANNs, the best performing algorithms for stress detection according to Sharma and Gedeon [16], are also the most computationally intensive techniques. In ambulatory research it might become more important to take this model characteristic into account.

Finally, current overview has only focused on supervised stress detection techniques, where the model is developed based on a training set including examples of features (e.g. physiology) and output variable (e.g. stress). It could also be interesting to investigate unsupervised techniques to identify stress vs. rest, without having to use subjectively reported stress levels. This has been explored by using self-organizing maps, obtaining a stress classification accuracy of 79% [123].

Table 2-6: Current research on physiological stress detection and machine learning techniques.

Ref.	Laboratory (L) Ambulant (A) Semi-ambulant (SA)	Nr. Of participants	Physiological signals	Analysis technique	Nr of classification levels	Classification performance
[53]	SA (driving on a set route)	9 (24 drives)	ECG, SC, EMG, respiration	Fisher projection and linear discriminant	3	97% (accuracy)
[58]	L	10	ECG	KNN	2	94.58% (accuracy)
[59]	A	4 (94h of data)	ECG	Correlation	3	68% (correlation)
[29]	L	40	ECG, respiration	SVMs	2	97.88% (accuracy)
[35]	SA (controlled driving task)	100	SC	Chi-square test	2	/
[105]	L	20	ECG, SC	Decision Tree BN SVM	2	92.4% (accuracy, decision tree)
[79]	L	10	EMG	KNN	2	90.70% (accuracy)

[80]	L	30	EMG	Friedman test Wilcoxon signed rank test	2	/
[81]	L	23	ECG,EMG, CO_2	Friedman test	2	/
[82]	L	17	ECG, EMG	t-test	2	/
[92]	L	65	BP	General linear models	Continuous	/
[94]	L	20	BVP, SC, respiration, ST	t-test	2	/
[95]	L	40	ECG, SC, EMG, ST	KNN Probabilistic NN	2	93.75% (accuracy)
[97]	L	32	BVP, SC, pupil diameter, ST	SVMs	2	90.10% (accuracy)
[122]	L	18	BVP, SC, sociometric badge (e.g. body movement, speech)	SVMs AdaBoost KNN	2	94% (accuracy, AdaBoost)

[121]	L	39	ECG, respiration	SVMs LDA AdaBoost KNN	Binary vs. three-class	84% (accuracy, SVM, three-class) 94% (accuracy, SVM, binary)
[124]	SA (driving on a set route)	13	ECG, SC, respiration, driver events (e.g. GPS)	BN	2	96% (accuracy)
[125]	L + A	21 (Lab) 17 (Ambulant)	ECG, respiration	Decision tree AdaBoost SVMs	2	90% (accuracy, lab, AdaBoost) 0.72 (correlation with self-reported stress, ambulant)
[107]	L + A	21 (Lab) 26 (Ambulant)	ECG, respiration	SVMs (lab) BN (ambulant)	2	95.3% (median accuracy, lab) 72% (accuracy, ambulant)
[126]	A	10	SC, ST	/	5	/

[127]	A	10	ECG, SC, respiration	RF Lasso SVR KNN	continuous	1.5 (mean squared error, SVR)
[17]	A	35	ECG, smartphone (audio, social, physical)	LR	3	61% (accuracy)

2.5. Conclusion

In this literature study we discussed three main topics: stress definition, the physiological stress response and machine learning techniques for stress detection.
Stress can be defined focusing on the subject, the environment, or the interaction between subject and environment. Although there is no consensus on which definition or theoretical framework is most correct, we do not aim to focus on that part of research in this thesis. We choose to use the JDC model of Karasek, defining stress as a combination of high job demands and low decision latitude.
The most commonly studied physiological signals and machine learning techniques for stress detection were discussed. However, important information is lacking in literature. First, it is not clear which physiological signals and features, or which combination of physiological signals and features, provide the highest accuracy for stress detection. Although review articles have tried to provide an answer, results are inconclusive due to differences in protocols, stressors and analysis techniques. Second, most studies have been executed in controlled laboratory conditions. It should be investigated whether insights based on these studies can be extended to the ambulant environment. Third, it is not clear which machine learning techniques are most suitable for physiological stress detection and whether a one-model-fits-all solution is feasible or whether models should be personalized.

Current work aims to provide answers to these problems, which are fundamental towards the goal of continuous, physiological stress detection in daily life. We investigate physiological sensing priorities in both laboratory and ambulant conditions, primarily focusing on physiological signals that can be monitored continuously in an ambulant environment, i.e. SC, ST, HR and HRV. Further, we investigate which machine learning techniques are most suitable, by comparing their performance in a controlled laboratory study on healthy subjects. We compare performances of LR, SVM, DT, RF and BN. Additionally, we compare personalized versus generalized models, both in laboratory and ambulant conditions and investigate the influence of different digital

phenotypes, including personal demographics and baseline psychological stress levels, on ambulant physiological stress detection.
These insights are fundamental towards daily-life stress detection and will form the basis for future research to enable highly personalized, just-in-time interventions for stress reduction.

Chapter 3: Physiological stress detection in a controlled environment

In this chapter we address the first objective of the thesis: the identification of the most suitable markers and machine learning techniques for physiological stress detection in a controlled laboratory environment on healthy subjects.
A protocol was set up and a dataset was collected and analysed by Elena Smets with promotor Chris Van Hoof. The data was used to compare multiple physiological signals and machine learning techniques and investigate inter-person variability. First, we conclude that, on average, SC and HR related features are more important than ST and respiration related features. Second, we conclude that for generalized models (i.e. including all subjects), SVMs perform best (average detection rate = 82.7%), for personalized models (i.e. based on one subject), dynamic BNs perform best (average detection rate = 84.6%). The content of this chapter is based on research presented at the MindCare conference [128][1].

3.1. Problem statement

Previous research has indicated that physiological signals can be used to detect mental stress in laboratory settings (see Chapter 2). There is however no consensus on the optimal algorithm for this detection. Large differences exist among classification accuracies from different studies. This is mainly due to three aspects of the studies being the experimental design, the sensor quality and the analysis methods. The focus of the current study is on the latter aspect. In many research linear discriminants, generalized estimation equations or support vector machines have been used to classify rest and stress states [53] [97] [129] [130]. Other, more recent, research has focused on probabilistic machine learning techniques such as Bayesian networks [124] [120]. Sharma

[1] The final publication is available at Springer via https://doi.org/10.1007/978-3-319-32270-4_2

and Gedeon [16] report an overview of different computational techniques for stress classification based on results from different studies conducted under different experimental designs, sensors and physiological parameters. Although this comparison can provide significant insight in which are good modeling techniques, up to our knowledge there is no direct comparison of modeling techniques that result in the optimal algorithm to employ for stress detection. Furthermore, in most research one general model is developed for all subjects. Literature and experts however agree that physiological responses to a stressor differ among subjects, e.g. the difference according to gender [131].

This study sought to compare several computational techniques for classifying stress based on physiological parameters within the same study design. Additionally, both generic and personalized models are compared.
The main contributions are:

a) We evaluate **physiological sensing priorities** for stress detection
b) We evaluate and compare the results of different **machine learning techniques** for stress modeling in comparison to rest
c) We compare the results of **generalized and personalized models** for stress detection.

3.2. Materials and methods

A controlled experiment was conducted to investigate the effect of stress on physiological parameters. The Medical Ethical Committee of KU Leuven approved the protocol and analysis methods of the experiment (protocol ID: S57066). In this section, the protocol and the sensing modalities are described. Furthermore, the feature list used for detection is described.

3.2.1. Data collection

Experimental Protocol

Twenty healthy participants, 10 males and 10 females volunteered to participate (mean age= 40 ± 10 years). Subjects were recruited in two

companies in Belgium and did not receive any compensation for their participation. Subjects were included if they were healthy employees with a mainly sedentary job. This was evaluated through an intake questionnaires, including, for example, questions related to whether subjects suffer or have suffered from psychosis, hyperventilation, depression, epilepsy, panic attacks, and burn-out. Subjects who answered 'yes' to any of these questions were excluded. Experiments were conducted in a quiet room using a standard desktop computer.

Figure 3-1 presents the timeline of the experiment. During the preliminary phase, the participant completed some general questionnaires and the sensors were attached. The test phase included three stress tasks of two minutes each. As a first task, a Stroop Color-Word test [132] was presented. Color words were written in an incongruously ink color, e.g., the word red was written in the color blue. Participants had to respond with the real ink color, e.g., blue in the previous example. A math test was performed as second task. In the third task, participants were instructed to tell about an emotional or stressful event in their life. To induce additional stress the experimenter could intervene by saying 'wrong' or 'faster' during the first two tasks. To control for the physiological response due to speaking, an additional counting task was included where the participant had to count out loud from zero to hundred. This task was performed twice: once before the Stroop test and once after the stress talk, separated by a two minutes rest phase. All parts during the test phase take two minutes, except for the counting blocks which are dependent of the participant's pace of counting and the first resting block which serves as a baseline and takes four minutes. During the finishing phase the participant completed a retrospective questionnaire where his/her stress levels during each task were rated on a five-point Likert scale.

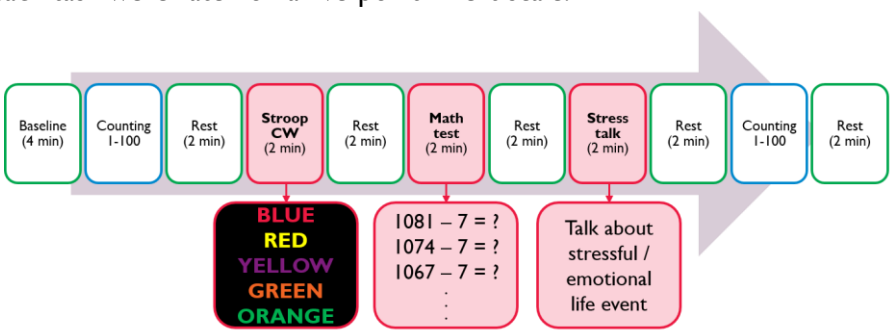


Figure 3-1: Experimental protocol

Physiological recordings

Two sensors were used. The first was the imec Necklace (Figure 3-2a), a wireless electrocardiography (ECG) sensor for research use developed by imec [133]. With this sensor single-lead ECG in a lead-one configuration was recorded at a sampling frequency of 256 Hz. The second sensor was the NeXus 10 – MK II (Mind Media, Herten, The Netherlands) (Figure 3-2b). This is not a wearable device, it is merely portable, though highly accurate. This sensor was used for the measurement of SC (Figure 3-2c) and temperature (Figure 3-2d) at the fingertip and respiration using a chest belt (Figure 3-2e). All NeXus signals were measured at a sampling frequency of 32 Hz.

Figure 3-2: Necklace (a) for ECG recording and NeXus 10 – MK II (b-e) for SC, ST and respiration recordings

3.2.2. Feature computation

Data from Nexus and Necklace were merged based on their timestamps and stored in Matlab files. Next, a comprehensive set of 22 features has been used, corresponding to the most frequently used features in earlier publications on the expression of stress. For each sensing modality, features have been calculated on a sliding window of 30 seconds with 29 seconds overlap. ECG has been characterized with HR and HRV, the latter considered in both time and frequency domain. SC features are based on tonic and phasic responses. ST has been characterized using the mean value and standard deviation for each window, and the corresponding slope. Finally, respiration has been characterized as energy of several frequency bands. The complete list of features is reported in Table 3-1.

Table 3-1: List of features computed for each sensing modality

Nr.	Feature	Abbreviation	Extracted from	Ref.
1	The root of the mean of the sum of the squares of differences between normal to normal beat intervals	RMSSD	ECG	[58]
2	Proportion of the successive normal to normal beat intervals that differ more than 50 ms	pNN50	ECG	[58]
3	Proportion of the successive normal to normal beat intervals that differ more than 20 ms	pNN20	ECG	[58]
4	Mean HR	mHR	ECG	[129]
5	Standard deviation of the normal to normal beat intervals	SDNN	ECG	[129]
6	Low frequency band (0.04-0.15 Hz)	LF	ECG	[129]
7	High frequency band (0.15-0.4 Hz)	HF	ECG	[129]

8	Low frequency over high frequency band	LFHF	ECG	[129]
9	Absolute second difference	AbsDiff2	SC	[129]
10	SC level	SCL	SC	[129]
11	Ohmic perturbation duration	OPD	SC	[129]
12	Number of peaks	Nrpeaks	SC	[63]
13	Tonic component (0-0.16 Hz)	mTonic	SC	[63]
14	Phasic component (0.16-2.1 Hz)	mPhasic	SC	[63]
15	Mean ST	mT	ST	[95]
16	Standard deviation of the ST	stdT	ST	[95]
17	Slope of the ST	slopeT	ST	[97]
18	Mean respiration frequency	meanRsp	Resp	[129]
19	Energy band 0-0.1 Hz	EB1	Resp	[53]
20	Energy band 0.1-0.2 Hz	EB2	Resp	[53]
21	Energy band 0.2-0.3 Hz	EB3	Resp	[53]
22	Energy band 0.3-0.4 Hz	EB4	Resp	[53]

3.2.3. Analysis methods

A binary classification problem was considered with classes corresponding to rest and stress periods. In theory a multi-class classification would also be possible (e.g. low, medium and high stress), but is here not feasible due to the low sample size and the lack of subjective feedback. The reference stress profile contains stress during the three different stress tasks and rest in the remainder of the experiment, including the counting parts. A feature selection methodology based on correlation was used to eliminate features that are not useful but can negatively affect the classification performance. For every feature the correlation with the reference stress levels was calculated and all features with an absolute coefficient higher than 0.5 were retained. The feature selection procedure was performed only on the training set. To guarantee independent predictors, variables with an absolute correlation higher than 0.8 were eliminated.

Six machine learning algorithms were considered for evaluating the classification performance. The selection aims to cover a comprehensive set of algorithms with both conventional, linear techniques and more novel

approaches (for a detailed explanation of the techniques see 2.4). We built six models:

1. logistic regression (LR) (Matlab 'mnrfit')
2. support vector machines (SVMs) (Matlab 'fitcsvm', using a linear kernel),
3. decision trees (DTs) (Matlab 'fitctree'),
4. random forests (RFs) (Matlab 'TreeBagger', using 20 trees based on out-of-bag samples),
5. static Bayesian networks (sBNs) (Matlab 'Bayes Net Toolbox' [134]),
6. dynamic Bayesian networks (dBNs) (Matlab 'Bayes Net Toolbox' [134])

For the LR, SVMs, DTs and RFs, standard hyperparameters available in Matlab were used. To learn the structure of the BN, a greedy search algorithm was employed, the conditional distributions were calculated using maximum likelihood estimation. Junction tree inference was used for classification of the test set.

For every machine learning algorithm, two models were trained, one using data from all subjects, i.e., a *generalized model*, and one using only data of a specific subject, i.e., *a personalized model*.

For the generalized models, a leave-one-out validation procedure was used. The models were trained on the data of all, but one, participant and evaluated on the data of this participant. For the personalized models a different approach was used. Since stress accumulates over time and its physiological response does not return immediately to the original baseline [135], we have trained our models on the first two stress tests and evaluated their performance on the last stress test, including the stress talk. Using this validation approach instead of the usual cross-validation we have been able to take the time-dependent nature of stress into account and to provide more trustworthy performance indicators of the models.

Sensitivity (Stress Detection Rate) and specificity (Rest Detection Rate) were considered as performance measures. These two measures will give a good understanding of the classification performance in case of an unbalanced amount of rest and stress examples. As overall performance measure, the average of these two measures was taken instead of the usual classification accuracy. We define this measure as Average Detection Rate (ADR).

3.3. Results

First a correlation-based feature selection was performed. For generalized models, 4 features were selected, mHR from ECG and SCL, mTonic and mPhasic from SC. For personalized models, the features selected varied according to person. On average 8 features were selected per person. The 5 features selected for most participants are mHR and mPhasic (84% of participants), SCL and mTonic (80% of participants) and AbsDiff2 (74% of participants). For no participant the following 5 features were selected: LF, LFHF, EB1, EB2 and EB4. This indicates that mainly SC and ECG-related features have high importance, and respiration-related features have lower importance.

Figure 3-3 represents the SC of one participant. The orange bars indicate the counting periods, which have been included to control for the response due to speech, the red bars indicate the stress tests. SC reacts in both areas which underlines the importance of including a control for speech. The figure also highlights the time-dependent nature of stress as after each task the SC does not return to baseline. This aspect is shown in further detail for the entire population in the boxplots of Figure 3-4. We show the increasing trend for all subjects in normalized SCL during the consecutive rest phases. This trend was found significant for SDNN (p<.001), HF (p=.03), LFHF (p=.005), SCL (p=.02), EB1s (p=.001) and EB3s (p=.01).

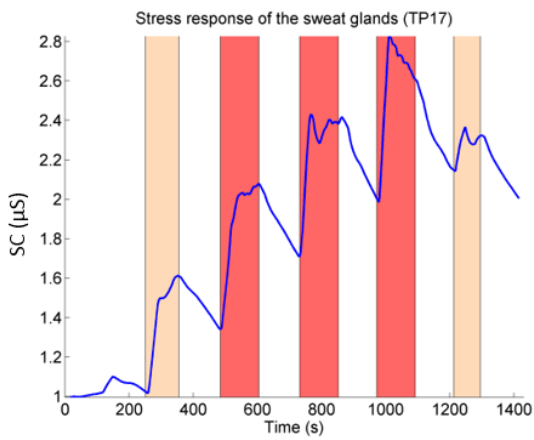


Figure 3-3: SC response of one participant

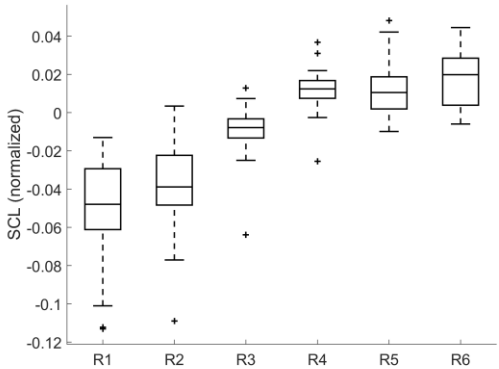


Figure 3-4: Boxplot with increasing rest (R1-R6) response of the normalized SC level (SCL)

The classification results obtained by generalized and personalized models are reported in Table 3-2 and Table 3-3 respectively and are graphically represented in Figure 3-5. The average and standard deviation for rest detection rate (RDR), stress detection rate (SDR) and average detection rate (ADR) are presented. Most of the misclassifications for the rest condition are situated in the counting task, due to the physiological response to speech. Results indicate that overall the highest ADR is reached using personalized dBN (84.6%) and generalized SVM (82.7%). Besides for dBN the personal approach did not perform better than the general.

Table 3-2: Classification accuracy for generalized models (RDR = rest detection rate, SDR = stress detection rate, ADR = average detection rate).

	LR	SVM	DT	RF	sBN	dBN
RDR (%)	93.2±2.8	93.4±3.2	88.6±4.4	90.7±4.1	91.2±3.6	58.3±14.8
SDR (%)	68.2±13.6	72.0±10.4	65.4±8.1	67.6±8.4	70.5±14.0	90.2±14.1
ADR (%)	80.7±7.3	**82.7±5.8**	77.0±4.9	79.2±5.1	80.9±7.8	74.3±10.3

Table 3-3: Classification accuracy for personalized models (RDR = rest detection rate, SDR = stress detection rate, ADR = average detection rate).

	LR	SVM	DT	RF	sBN	dBN
RDR (%)	79.5±20.4	77.5±20.2	78.3±18.7	79.1±19.0	81.3±21.2	87.7±10.4
SDR (%)	72.5±25.2	74.8±25.8	69.2±24.4	72.0±25.4	77.0±25.3	81.5±21.9
ADR (%)	76.0±10.7	76.1±11.3	73.7±12.6	75.6±12.9	79.2±13.7	**84.6±9.8**

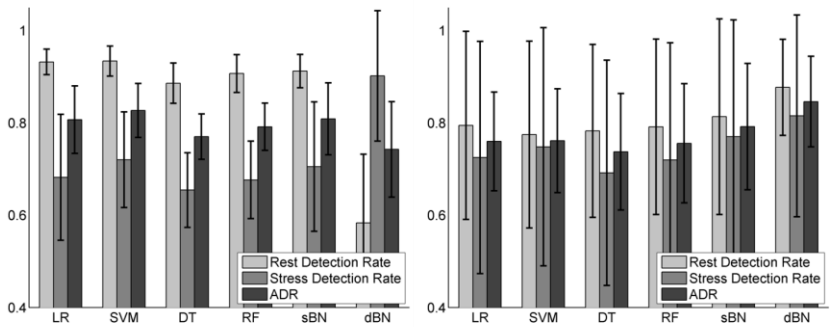


Figure 3-5: Classification accuracy for generalized models (left) and personalized models (right)

3.4. Discussion

To correct for the inherent physiological response due to speech, a counting task was introduced before the first and after the last stress task. Classification performances showed that we could successfully differentiate between rest and stress, although most misclassifications for the rest condition were situated during the speech task. This emphasizes the importance of including regular speech in the experimental protocol for stress detection. Further, the feature selection procedure indicated that mainly SC-based features together with the mHR are interesting with respect to the detection of stress in a controlled environment. We also showed the cumulative effect of different successive stress tasks on physiology, with an increased baseline over the different rest periods. This indicates that the two minutes rest phase which was selected, might not be enough for full physiological recovery of a stress task. Further, it indicates that the physiological stress response is a relative measure and we need intelligent models to differentiate between rest and stress. Models incorporating a time dimension, such as the dBNs, could benefit since these are inherently relative and take previous information into account to predict current stress levels.

Comparison of the results in Table 3-2 and Table 3-3 does not confirm the hypothesis that a personal approach renders higher average detection rates (ADR) than a general approach. This is only the case for dynamic Bayesian networks. However it can be observed that generalized models have relatively low stress detection rate, compared to rest detection rate. In most

applications the main goal is to detect stress. Therefore models with higher stress detection rates should be preferred. Furthermore it can be observed that standard deviations for the personal approach are much higher than for the general approach. This means that for some participants a very high ADR could be reached, where for others the ADR was very low. Further analysis revealed that the datasets with high and low ADR are not the same for different modelling techniques. Future research should therefore investigate whether a further personalization in terms of machine learning algorithm selection could be beneficial. Another improvement could be made by merging the generalized approach with a subject-dependent feature calculation as suggested in [136]. Finally a personalized method is capable of giving more insight into the personal physiological stress response, e.g., the correlation-based feature selection can give an indication of the person's principal stress physiology. This can be interesting for targeted treatment and relaxation techniques in order to overcome the detrimental effects of stress on the human body.

The best classification results, for this experiment, were obtained for the personalized dynamic (dBN) Bayesian networks and the generalized support vector machines (SVM), with ADR of 84.6 % and 82.7 % respectively. It can be expected that dBN profit most from a personal approach, as they are probabilistic, adapting models. On the other hand SVM are models which are more most capable of generalization as compared to the other techniques and therefore can perform better in a general approach. The calculation of the dBN however is quite time consuming and computationally heavy. The SVM method is much less effortful and still gives reasonably good classification results. The downside of this approach however is that it can be considered a complete black-box. This is not a problem in terms of classification, but it becomes a problem when the goal is to gain insight in the physiological stress response. For that purpose BN are much more suited, due to their graphical character.

Therefore in the future a distinction should be made based on the purpose of the analysis. If the goal is to develop a fast algorithm for real-time stress detection, where only information about stress or no stress is required, the SVM technique should be considered the best choice. If the goal is to gain insight into a person's stress response a better option is to use dBNs.

Furthermore future research should investigate whether the conclusions drawn from this controlled study also hold for ambulatory studies.

3.5. Conclusion

The goal of the study was to identify the optimal physiological signals and computational methods for stress detection in a controlled environment. An experiment was conducted in a laboratory environment where participants had to fulfill three different stress tests. To control for the physiological response to speech a counting task was introduced before the first and after the last stress task. Four physiological signals and six machine learning techniques were investigated using a general and personal approach.
First, it can be concluded that SC and HR related features outperform ST and respiration related features. Second, we conclude from this study that personalized dynamic Bayesian networks and generalized support vector machines render the best average classification results with 84.6% and 82.7% respectively. Based on characteristics inherent to the methods, it is suggested to use dynamic Bayesian networks when insight in the model is necessary and to use support vector machines when it is not.

Chapter 4: Comparison of the physiological stress response of healthy subjects and patients

The content of this chapter addresses the second objective of this thesis: the differentiation between healthy subjects and patients based on their physiological stress response. The goal is to develop a model to differentiate between healthy subjects and patients which could serve as a first step towards disease prevention and interception. We repeated the experiment conducted in Chapter 3 with patients with stress-related complaints. The dataset protocol was set up and the analysis was done by Elena Smets with promotor Chris Van Hoof, the data was collected by the therapists of Tumi Therapeutics. We used an exploratory methodology to calculate features and develop a model to differentiate between healthy subjects and patients. We achieved a classification accuracy of 78% based on response-related features, including all physiological signals and using SC-related features only. The content of this chapter is published in Health Science Reports [137][2].

4.1. Problem statement

Many studies have revealed the harmful influence of chronic stress on mental and physical health. For example, Stansfeld and Candy [138] concluded from a meta-analysis, that work stressors are prospective risk factors for common mental health disorders, including depressive and anxiety disorders. Rosengren et al. [139] have shown that psychosocial stressors increase the risk of acute myocardial infarction. Furthermore, associations have been established between psychological stress and depression, cardiovascular disease and the course of HIV/AIDS [34]. Another review concluded that both acute and chronic stress research reveal extensive data concerning the stressors'

[2] The final publication is available via https://doi.org/10.1002/hsr2.60

contributions to deteriorated health, including sudden death and myocardial infarction [140]. Together, these findings highlight the need for affordable and effective early detection of stress problems and preventive interventions of stress-related mental health disorders.

Stress-related health problems can be conceptualized into three areas along the stress continuum [141]: stress-related complaints, overstrain, and burnout. A main differentiator between these three areas is the chronicity of the complaints. For stress-related complaints, the time since the onset of the complaints is less than three months, whereas for overstrain this is more than three months and for burnout it is more than six months [141]. Furthermore, persons categorized in the stress-related complaints group do not yet feel any substantial limitation in their social or professional functioning, whereas this is increasingly the case both for overstrained and burnout patients [141].

Physiological signals such as HR, BP, and SC have been investigated to detect stress-related health problems. Studies on autonomous nervous system (re)activity in the context of stress-related health problems have focused especially on the last stage in the stress continuum, i.e. burnout. May *et al.* [142] found that school burnout was associated with decreased baseline heart rate variability (HRV). Contradictory, Morgan *et al.* [143] showed that persons who score higher on the Maslach Burnout Inventory have significantly higher HRV. De Vente *et al.* [144] found that burnout patients show higher resting HR than healthy controls. Other studies investigating the hypothalamic-pituitary-adrenocortical (HPA) activity concluded that burnout patients and controls do not show differences in HPA outcomes [145]. Although preliminary, such research is promising for the detection of burnout. However, in terms of prevention it could be more valuable to detect stress-related health problems already in an earlier stage of the stress continuum. To date, no validated questionnaires exist to identify individuals with stress complaints, who are vulnerable to develop overstrain and burn-out.

In the current study, we therefore sought to focus on persons with stress-related complaints who are not yet limited in their social or professional functioning, i.e. the first stage of the stress continuum. Analogous to previous studies, focusing on burnout [144], we aimed to investigate the patient's autonomic nervous system responses to and recovery from an acute stressor, as especially these measures and reactivity patterns may have a great potential for ambulatory stress monitoring and dynamically tailored, direct feedback and

just-in-time behavioral interventions. However, in contrast with most studies in this field, we opted for a less conventional, fundamentally different approach of the data. Traditionally, psychophysiological studies are hypothesis driven, which means that a study is specifically designed to answer a question [146]. The analysis therefore is confirmatory rather than exploratory. However, as technology is continuously improving and wearables become widespread, the amount and nature of psychophysiological data that is available has exponentially grown and calls for complementary approaches that allow to maximally explore the wealth of data that is nowadays available. Data scientists have already moved towards more exploratory data-mining techniques to develop classification algorithms that can unravel new knowledge hidden in the data [146]. In this study we will explore and apply this more exploratory approach to analyze the data and differentiate persons with stress-related complaints from healthy subjects.

Previous studies have mainly investigated single physiological parameters independently (e.g. [143] [144]), while combinations of multiple physiological parameters and comparisons between single markers should preferably be investigated. Furthermore, previous studies have focused mainly on static features, i.e. the comparison of mean HR in rest and stress tasks. However, both physical fitness and stress research strongly suggests that dynamic features such as response and recovery time, can provide additional information regarding physical condition determination [147]. Based on the research of McEwen [148] failure to shut off allostatic activity after a stress response, is one type of allostatic load. This could be reflected in a longer recovery time of the physiological signals after a stressor for patients. It is therefore needed to investigate if such dynamic features can also improve the detection of persons with stress-related complaints.

In this study we aimed to explore a multi-parameter classification model that, based on the physiological response to and recovery from three standardized laboratory stress tasks, can differentiate between healthy subjects and persons with stress-related complaints. We also assessed which physiological signal(s) are most suitable for the characterization of persons with stress-related complaints. We included three commonly used physiological signals for stress detection being HR, SC and ST. We hypothesize that a classification model combining all three physiological signals will outperform models based on the individual signals separately. Furthermore, we compared classification

performances based on response and recovery related features. We hypothesize, based on the suggestion of Linden *et al.* [149], that recovery related features could provide additional insight in the difference between healthy subjects and persons with stress-related complaints and therefore can increase classification performance. Finally, we used both static and dynamic features for classification. We hypothesize, based on earlier findings in physical fitness research [147], that dynamic features can improve classification performance. The findings will enhance our understanding of the physiological differences between healthy subjects and persons with stress-related complaints and may advise further strategies to use physiological signals for the early detection of stress-related health problems.

4.2. Materials and methods

A controlled experiment was conducted to investigate the difference in physiological stress response between healthy subjects and patients with stress-related complaints. Therefore, the experiment presented in Chapter 3 for healthy subjects, was here repeated for patients with stress-related complaints. The Medical Ethical Committee of KU Leuven approved the protocol and analysis methods of the experiment (protocol ID: S57469). In this section, the protocol, sensing modalities and the feature list used for detection are described.

4.2.1. Data collection

Participants

A controlled laboratory study was conducted with the approval of the Medical Ethical Committee of the UZ Leuven. All subjects signed an informed consent form before participating in the study. In this study, 32 subjects, of which 20 healthy subjects (10 women, 10 men, , Mean age=39.8 years, age range: 26-57 years) and 12 persons with stress-related complaints (7 women, 5 men, Mean age=38 years, age range: 23-56 years), participated. The data of the healthy subjects is the same as for Chapter 3. An additional data collection was set-up to acquire data from persons with stress-related complaints. The focus of this research is on early detection of stress-related health problems, therefore,

only persons with stress-related complaints, but without formal diagnosis of any clinical mental health disorder were included.
For inclusion and exclusion criteria of the healthy subjects we refer to Chapter 3. Persons with stress-related complaints were recruited at Tumi Therapeutics, a multidisciplinary ambulatory diagnostic and treatment centre specialized in stress-related symptoms and syndromes. In return for participation, patients received the psychophysiological diagnostics, which involved the stress tests, free of charge. In addition to the stress test and as part of the standard intake procedure at Tumi Therapeutics, patients also completed a set of questionnaires. Only patients with stress-related complaints (first phase of the stress continuum) were included. Specifically, the following inclusion criteria were applied: a) the patient experienced somatic complaints, *and* b) the complaints started less than three months before consultation *and* c) the patient did not feel limited in his or her personal or professional life, *and* d) the patient did not suffer from any psychiatric disorder or organic disease. To assess the somatic complaints, the Dutch Symptom Checklist-90 (SCL-90) [150] was used. This questionnaire is often used in clinical practice and research for initial evaluation of patients at intake. The test measures eight primary symptom levels, i.e. sleep difficulties, agoraphobia, hostility, somatization, interpersonal sensitivity, anxiety, cognitive-performance deficits and depression. The results can be compared with a healthy and clinical norm group for female and male subjects separately [151]. The mean results for the selected patients are reported in Table 4-1. Numbers indicate the severity of the complaint, based on patients' answers to questionnaires. The included patients scored higher on the subscales than the healthy norm group, but lower than the clinical norm group, for all scales, except for somatization and sleep difficulties for which they scored higher than the average clinical norm group.

Table 4-1: Average scores for male and female patients on the different scales of the SCL-90, , compared with a healthy and clinical norm population [151]

SCL-90 scale	Male			Female		
	Patients	Healthy population	Clinical population	Patients	Healthy population	Clinical population
Somatization	25.6±7.7	15	24	32.2±7.7	16	26
Cognitive-performance deficits	19.8±9	12	20	20.3±9.3	13	21
Interpersonal sensitivity	28.4±7.4	23	35	35±12.6	23	38
Depression	29.6±6.2	18	37	34±7.1	21	44
Anxiety	17.4±3.1	11	23	22±2.6	13	27
Hostility	9.4±2.8	6	10	15.8±7.3	6	10
Agoraphobia	9±2.2	7	11	11.5±4.4	7	12
Sleep difficulties	6.2±1.5	3	5	8.2±3.7	4	7

The Nijmegen questionnaire for hyperventilation [152] was used to assess several singular stress complaints such as chest pain, being short of breath, blurred vision. Included subjects scored positive on the Nijmegen questionnaire for hyperventilation, having 18 points or more. All subjects confirmed their complaints started less than three months before consultation and all subjects were still capable of fully functioning in their social and professional lives. Further, a clinical interview based on the Mini International Neuropsychiatric Interview (MINI), which is based on DSM-IV criteria [153] [154] was conducted to exclude the existence of any psychiatric disorders. Organic diseases were excluded based on doctor's reports, physical examination, medical tests and self-reporting. The healthy subjects did not report any physical or psychological disease or complaint.

Procedures

The protocol is based on the protocol explained in Chapter 3. It consists of three stress tests of two minutes each: a Stroop Color-Word test, a math test and a stress talk, each separated by rest phases of two minutes. The main difference with the protocol for patients is that the counting task, performed by healthy subjects at the beginning and end of the protocol, was eliminated

for patients. In Chapter 3, we showed that a stressful task with speech can be distinguished from a non-stressful speaking task, i.e. counting. For this reason and to reduce the experimental time, the counting task was removed for the patients. The adjusted protocol is presented in Figure 4-1. To align the two protocols, the two counting tasks executed by the healthy subjects and the first rest phase executed by both healthy subjects and patients were excluded from further analysis.

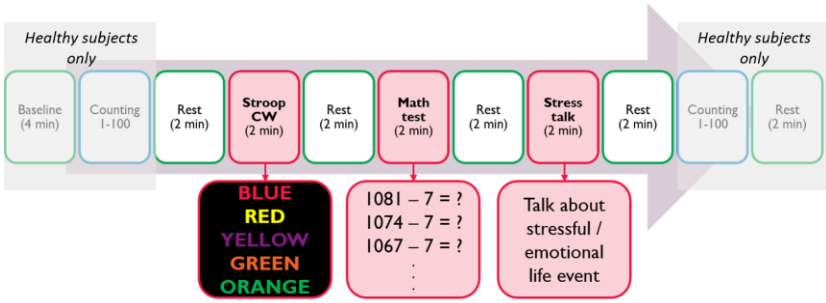


Figure 4-1: Adjusted experimental protocol for healthy subjects and patients (no counting)

Physiological recordings

Three physiological signals were measured using the NeXus 10 – MK II hardware (Mind Media, Herten, The Netherlands) (same as in Chapter 3, see Figure 3-2). SC was recorded at 32 Hz from the distal phalanx of the index and middle finger of the nondominant hand using Ag/AgCl electrodes embedded in Velcro straps. Skin temperature (ST) was recorded at 32 Hz from the distal phalanx of the little finger of the nondominant hand using a thermistor. This is a small point probe, secured by placing tape over the measuring tip to avoid signal contamination by air flow. Heart rate (HR) was measured at 128 Hz using a blood volume pulse (BVP) sensor at the ring finger of the nondominant hand. The sensor used photoplethysmography which is a light-based technology to sense the rate of blood flow as controlled by the heart beats. Based on this signal, instant HR was detected in real-time by the NeXus software. For healthy subjects additionally HR was measured using the Necklace (see Figure 3-2), but in current analysis only NeXus-based HR was used, since it reduced the number of sensors needed for the patients (only using NeXus). Participants were asked to keep the hands still, as all signals are susceptible to motion artefacts. Physiological channels were simultaneously

streamed to disk and displayed on a PC monitor. Offline, all channels were visually inspected to ensure good quality.

4.2.2. Feature computation

Whereas for physiological stress detection already many features have been proven successful in previous research (see 3.2.2. Feature computation), this is not the case for differentiating healthy subjects and patients. Therefore, we aimed to expand the number of existing features by exploring new features, selected based on results in other fields of research, as is explained below. We applied an exploratory approach towards the signal analysis and feature computation, meaning the outcome for each feature is not hypothesized beforehand, but rather explored.

Before feature extraction, data from both populations, i.e. healthy and patient, were merged and stored in data frames in Python (version 2.7), using the Pandas library. The physiological signals were standardized with zero mean and unit variance per subject to obtain time series on the same scale. Then the time series were divided into rest and stress blocks of two minutes each, according to the task performed in each segment. This resulted in a total of seven blocks, four rest blocks (R_1, R_2, R_3 and R_4) and three stress blocks (S_1, S_2 and S_3). The first rest block (R_1) was excluded since for the healthy subjects this task was preceded by a counting task, whereas for the patients this was the start of the experiment. Next, two types of features were calculated: static and dynamic features.

The static features describe the distribution of the physiological signals, e.g. the mean and standard deviation, in each block. For each signal 18 static features were calculated, including the mean and standard deviation, as well as differences of means between pairs of rest or stress blocks (see Table 4-2). These trends were calculated to explore whether healthy subjects and patients differ in the cumulative effect of consecutive stress tasks.

The dynamic features represent the transition between different blocks, e.g. the transition from rest to stress as response features and the transition from stress to rest as recovery features. These type of features have been shown valuable in physical fitness research [147], we investigate whether they can bring additional value for detecting persons with stress-related complaints. For each signal 24 dynamic features were calculated. Previous research has

indicated that HR and SC increase [53] and ST decreases [94] as response to a stressor. Therefore, for each stress block the response time was calculated as the time to reach the maximum value for HR and SC and the minimum value for ST starting from the onset of the stress task. Similarly, for each rest block the recovery time was calculated as the time to reach the minimum value for HR and SC and the maximum value for ST starting from the onset of the resting phase. Additionally, for all the blocks a straight line was fitted through the signal and the slope was calculated. To investigate the cumulative effect of the different stress tasks, also the trends across the slopes and the response or recovery times over the different pairs of rest and stress phases were calculated, e.g. a positive value for the trend of HR slopes means an increase in steepness of response, a positive value for the trend of HR response times means an increase in response time after different stress tests. In Figure 4-2 the SC response to the three stress tests, indicated in red, is shown. The recovery time, recovery slope, response time and response slope are graphically represented. An overview of all the features is presented in Table 4-2.

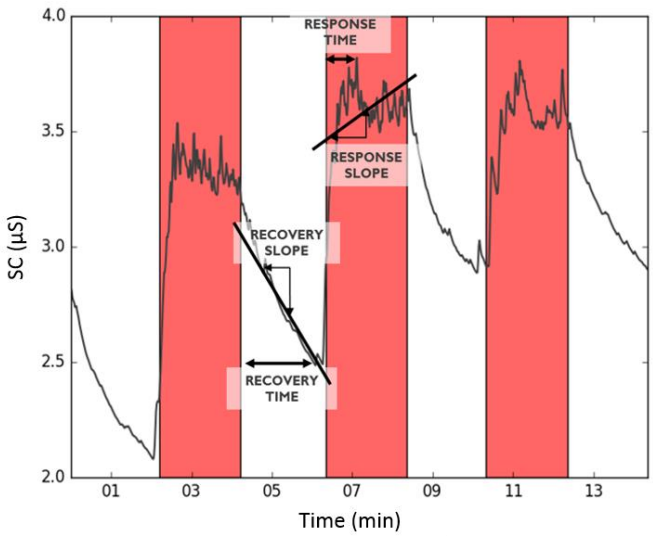


Figure 4-2: Dynamic feature calculation including recovery time, recovery slope, response time and response slope.
Red bars represent stress phases, white are rest. The example signal is SC from one subject, the same features are calculated for ST and HR.

Table 4-2: Overview of the static and dynamic features, calculated for each physiological signal (R_x=stress block x, S_x = stress block x).

Nr.	Feature name	Blocks	Static or dynamic	Explanation of features
1-6	Mean	R_2, R_3, R_4, S_1, S_2, S_3	Static	Mean of the physiological signal in the rest/stress block
7-12	Standard deviation	R_2, R_3, R_4, S_1, S_2, S_3	Static	Standard deviation of the physiological signal in the rest/stress block
13-15	Trend means of stress	$S_4 - S_2$, $S_4 - S_3$, $S_3 - S_2$	Static	Difference between the means of different stress phases e.g. $S_4 - S_2$
16-18	Trend means of rest	$R_4 - R_2$, $R_4 - R_3$, $R_3 - R_2$	Static	Difference between the means of different rest phases e.g. $R_4 - R_2$
19-21	Response time	S_1, S_2, S_3	Dynamic	Time in seconds to reach the maximum (HR and SC) / minimum (ST) starting from the onset of the stress task
22-24	Recovery time	R_1, R_2, R_3	Dynamic	Time in seconds to reach the minimum (HR and SC) / maximum (ST) starting from the onset of the rest phase
25-30	Slope	R_2, R_3, R_4, S_1, S_2, S_3	Dynamic	Slope of a straight line fitted through physiological signal in the rest/stress block
31-33	Trend response times	$S_4 - S_2$, $S_4 - S_3$, $S_3 - S_2$	Dynamic	Difference between the response times of different stress phases e.g. $S_4 - S_2$
34-36	Trend recovery times	$R_4 - R_2$, $R_4 - R_3$, $R_3 - R_2$	Dynamic	Difference between the recovery times of different rest phases e.g. $R_4 - R_2$
37-39	Trend slopes of stress	$S_4 - S_2$, $S_4 - S_3$, $S_3 - S_2$	Dynamic	Difference between the slopes of different stress phases e.g. $S_4 - S_2$

40-42	Trend slopes of rest	$R_4 - R_2$, $R_4 - R_3$, $R_3 - R_2$	Dynamic	Difference between the slopes of different rest phases e.g. $R_4 - R_2$

4.2.3. Analysis methods

The goal of this study is to develop a classifier that can distinguish between healthy subjects and persons with stress-related complaints.
In Chapter 3 we have shown that SVMs and dBNs provide the highest classification accuracies. However, in current research we aim to gain insight in the feature importances and model structure, therefore SVMs are not suitable. Additionally, dBNs are not appropriate because there is no dynamic component in separating healthy subjects and patients (as opposed to detecting rest and stressful events successively). Therefore, logistic regression (LR) using the Scikit-learn library of Python 2.7 with default hyperparameters was used for the analysis [118]. In LR the probability of the outcome of the healthy subjects versus patients is modeled as a function of the features weighed by coefficients obtained with a training set [108].
A total of 126 features were calculated (i.e. 18 static and 24 dynamic features for 3 physiological signals). To avoid overfitting unsupervised feature selection using principal component analysis (PCA) was applied. We calculated the principal components of the features and selected the number of components which explained 95% of the variance of the dataset (20 components). Then, we calculated the correlation of each feature with each principal component and retained the features with a correlation higher than 0.6 with at least one component. This reduced the dataset from 126 to 38 features. Next, to minimize feature redundancy, we calculated the correlation between all features and removed those with a correlation higher than 0.6, reducing the dataset to 26 features.
To compare the classification performance of separate physiological signals and of recovery versus response signals, six feature sets were separated based on the reduced feature set: a) a combination of all features derived from all physiological signals i.e. SC, HR and ST (26 features), b) all features derived from SC (8 features), c) all features derived from HR (8 features), d) all features derived from ST (10 features), e) all recovery-related features derived

from all physiological signals (13 features) and f) all response-related features derived from all physiological signals (13 features).
The performance of each classifier was assessed using a leave-one-out cross-validation. The models were trained on the data of all, but one, participant and evaluated on the data of this participant, this was repeated until all participants had been evaluated exactly once. To evaluate the model performance specificity, or true negative rate (healthy), sensitivity, or true positive rate (patient), and accuracy were calculated.
To further investigate the contribution of separate features of different physiological signals to the model, the feature importance was calculated for the model with the highest performance (accuracy). In a LR model, more important features have higher weights. Therefore, the feature importance was calculated by ranking the weights of the model. For the most important features also a t-test was performed. For features with significant differences, i.e. $p<.05$, also the effect size (Cohen's d) was calculated [155].

4.3. Results

To identify healthy controls and persons with stress-related complaints, classifiers using LR based on six feature sets were developed. After unsupervised feature reduction, based on PCA and correlation analysis, 26 features were retained, 10 static and 16 dynamic. The accuracy, sensitivity and specificity for each set are presented in Figure 4-3. The best performance was obtained for the response and SC feature sets (accuracy = .78, sensitivity = .75, specificity = .80). The worst performance was obtained for the ST feature set (accuracy = .59, sensitivity = .50, specificity = .65) and recovery feature set (accuracy = .63, sensitivity = .50, specificity = .70). An intermediate performance was found for the single-parameter feature set with HR features (accuracy = .66, sensitivity = .50, specificity = .75) and feature set with all features (accuracy = .72, sensitivity = .75, specificity = .70).

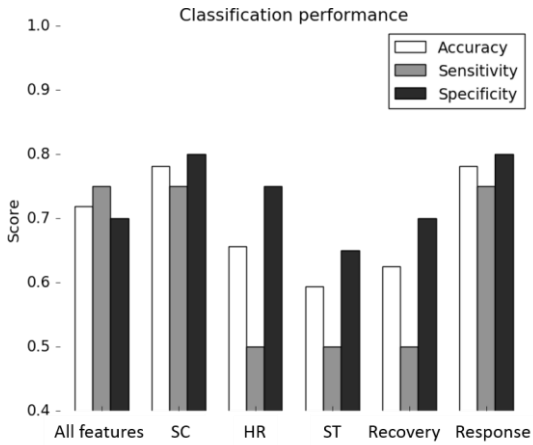


Figure 4-3: Classification performance for each feature set using a LR model.
The performance is evaluated using F-score, sensitivity and specificity. Classification based on the response and SC features give the best performance.

For the model based on the response feature set (highest accuracy, including all physiological signals), the relative feature importance was further investigated. Features were ranked based on their relative contributions to the model predictions. The result is shown in Figure 4-4.

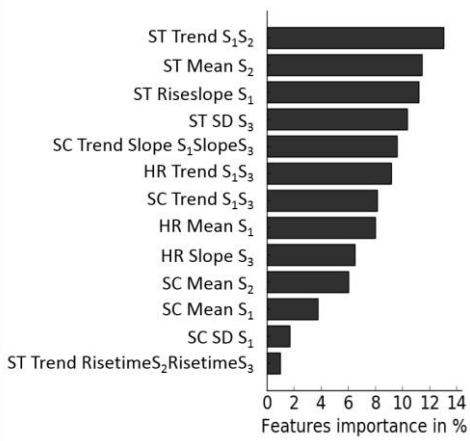


Figure 4-4: Feature importance of the response feature set based on the relative contribution to the LR model (SD = standard deviation).
Feature names contain 3 parts, separated by a space: (1) the physiological signal for which the feature was computed, i.e., HR, SC, or ST; (2) the feature (see Table 4-2); and (3) the stress task(s) for which the feature was computed: S1 = stress task 1 (i.e., Stroop Color-Word test), S2 = stress task 2 (i.e., math test), S3 = stress task 3 (i.e., stress talk).

Significant differences for the t-test and medium to large effect sizes based on Cohen's d were found for the five most important features (others did not show significant differences). These include four ST and one SC related features. The t-test was found significant for $p<.05$ and an effect size $d>0.5$ was considered medium and $d>0.8$ large [155]. In Figure 4-5 the boxplots of these features are shown comparing the standardized feature values of healthy subjects and patients. The stars indicate statistically significant differences.

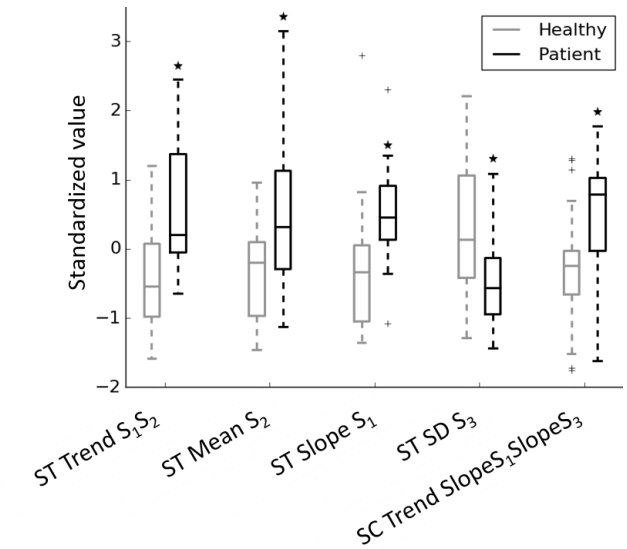


Figure 4-5: Boxplots of the five most important features of the response feature set for healthy subjects and patients.
*Features are represented as standardized values. *, p<.05 vs. healthy subjects. Feature names contain 3 parts, separated by a space: (1) the physiological signal for which the feature was computed, ie, HR, SC, or ST; (2) the feature (see Table 4-2); and (3) the stress task(s) for which the feature was computed: S1 = stress task 1 (i.e., Stroop Color-Word test), S2 = stress task 2 (i.e., math test), S3 = stress task 3 (i.e., stress talk).*

The trend of the ST means from the first to second stress task, i.e. Stroop test to math test, was close to zero for patients and significantly lower, i.e. more negative, for healthy subjects ($p=.007$, $d=1.06$). Since the trend is the difference of S_2 and S_1, this indicates that healthy subjects have a lower ST in the second stress task compared to the first, while this difference is less distinct for patients. The mean ST from the second stress task, i.e. the math test, was significantly higher for patients compared to healthy subjects ($p=.02$, $d=0.90$). The slope of the ST during the first stress task, i.e. the Stroop test, was

significantly higher for patients compared to healthy subjects (p=.02, d=0.87). Since both slopes are negative (based on the not normalized values), this indicates a stronger ST decrease for healthy subjects. The standard deviation of the ST from the third stress task, i.e. the stress talk, was significantly lower for patients compared to healthy subjects (p=.04, d=0.80). Finally, the trend of the SC slopes from the first to the third stress tasks, i.e. Stroop test to stress talk, was significantly higher for patients compared to healthy subjects (p=.05, d=0.73). This indicates a stronger increase in SC slopes (i.e. a stronger SC response) for patients.

4.4. Discussion

We investigated the acute physiological response to and recovery from a stress task for early detection of stress-related complaints. Subjects with stress-related complaints could be distinguished from healthy subjects with an accuracy of 78%, sensitivity of 75% and specificity of 80%. Whether these results generalize to a larger population, patients with clinical diagnoses such as burnout, chronic fatigue syndrome, etc., or to other types of stressors requires further study.

Our analysis also points to several conclusions with respect to physiological sensing priorities. In previous reports, mainly cardiovascular or SC features have been used separately as physiological markers of stress-related diseases. Our analysis indicated that the best results can be obtained using SC related features. The classification performance using only features related to ST or HR was much lower. However, by combining the response features of HR, SC and ST the performance can be increased and insights from all physiological signals can be obtained. Since all three are standard measurements, readily available in many state-of-the art sensors, e.g. NeXus 10 – MK II, and in multiple wearables such as Empatica E4 (Empatica, Milan, Italy), it is advised to focus further research on the combination of these signals, rather than investigating them separately.

Furthermore, when only using the recovery related features, the accuracy was reduced by 15% as compared to using response related features. These findings are in disagreement with the suggestion of Linden *et al.* [149] that recovery features can unravel additional information to distinguish healthy

subjects and patients. Our findings indicate that these two groups differ more in their response to stress than in their recovery from a stress task. A possible explanation could lie in the timeframe of the analysis. In our research the immediate stress response and recovery were analysed in a timeframe of two minutes during and after the stress task. It is possible that differences become more apparent after a longer period. Further, we focused our research on persons with stress-related complaints for less than three months before consultation. It is possible that if the chronicity of the complaints increases, e.g. burnout patients with complaints for more than six months, also the difference in recovery phase becomes more pronounced. This hypothesis is supported in the meta-analysis of Miller *et al.* [37] who state that when chronic stress first begins, the HPA axis is activated, whereas prolonged chronic stress, which is the case for burnout patients, leads to diminished activity. In a follow-up study it could therefore be interesting not only to investigate healthy controls versus subjects with stress-related complaints, but also subjects with stress-related complaints versus overstrain versus burnout patients.

In our analysis, we also investigated which type of features, static or dynamic, is more important for classification purposes. We showed that both types are needed to reach the reported classification performances, with a higher number of dynamic features selected. In previous research towards identification of stress-related mental health problems, the focus has been on static features. In other research branches, such as the identification of physical condition, as opposed to mental, dynamic features have been already incorporated in the analysis [147]. We suggest that future research in the area of mental and physical health may benefit from including more dynamic features in the analysis. In our study, a linear approach was used to calculate the slopes, which, as can be seen in Figure 4-2, might not be the best representation. Therefore, in Lim *et al.* [156] an exponential approach was proposed. Additionally in Figure 4-2 it can be noticed that the onset of the physiological response already starts a few seconds before the start of the stress test. This could be due to anticipation of the test and could also be an interesting parameter for future research, e.g. anticipation effects could be more pronounced for patients as compared to healthy subjects.

Detailed investigation of the most important features for the model based on response related features, revealed that mainly feature slopes and trends are important (Figure 4-4). The five most important features showed significant

differences and medium to large effect sizes for the healthy subjects compared to the patients. A general observation of the results shows that patients often show a more rigid response to stress compared to healthy subjects (i.e. less variation between rest and stress). This could reflect one type of allostatic load, being the inadequate response of the allostatic systems as described by McEwen [148]. These results highlight the opportunities of using physiological stress responses as a means to discover new insights regarding the process of stress-related health disorders.

The current study was a methodological pilot study which was executed in a laboratory setting and with a limited number of patients (n=12). Results might be changed if more patients will be recruited in the future. Further, a possible application of this methodology could be large-scale population screenings for early detection of stress-related health problems. Therefore, to use this methodology in practice, it should be investigated whether similar results can be obtained in real-life conditions, outside the laboratory. To this end, wearables such as Empatica E4 (Empatica, Milan, Italy) could be used for ambulatory physiological measurement of HR, SC, and ST. Additional challenges will be related to signal quality [106]. In current study only persons with stress-related complaints were included. All patients confirmed their complaints started less than three months before consultation and all patients were still capable of fully functioning in their social and professional lives. However, since this information is based on self-report, it could be incorrect as patients might be unaware of problems in their functioning. Further, we suggest additional research to investigate whether the results generalize to larger populations and patients on different areas along the stress continuum (i.e. overstrain and burn-out). We aimed with this methodological pilot study to bring attention to new exploratory methodologies; further research is needed to validate and replicate the results.

4.5. Conclusion

We conclude that our pilot study demonstrated the potential of physiological signals during the response to a stress task to discriminate healthy subjects from persons having stress-related complaints. Our analysis also showed that a multi-parameter classification model based on response-related features can

outperform models based on single parameters (HR and ST) and models based on recovery-related features only. Investigation of the separate features can provide more insights and enhance our understanding of the physiological differences between healthy subjects and persons at risk of stress-related health problems. Although further research is needed to investigate if these conclusions generalize to a larger population and to multiple clinical diagnoses, these results highlight the potential of using physiological signals and an exploratory approach to gain more insight into the difference between healthy subjects and patients. Further longitudinal research using wearable technology to investigate the development of the three stages on the stress continuum, could provide a powerful technique for better understanding the development of stress-related disorders. Such research could unravel early detection points for early diagnosis and prevention.

Chapter 5: The SWEET study: A large-scale, multi-sensor trial for stress detection in the work environment

This chapter addresses the third objective of this thesis: The large-scale investigation of the physiological stress response in ambulatory conditions, including demographics and context information towards digital phenotypes for personalized and continuous stress detection. More specifically, this chapter focuses on the set-up of a large-scale study for "Stress detection in the Work EnvironmEnT" (SWEET). We present a trial, including 1,002 subjects, containing subject's demographics and baseline psychological information and five consecutive days of free-living physiological and contextual measurements. The dataset protocol was set up by Elena Smets with promotor Chris Van Hoof, the data collection was performed by Elena Smets and Jan Cornelis at 11 companies across Flanders and the Netherlands. We present the data collection protocol and results regarding compliance and data quality. In the next chapters we will analyse these data further towards a personalized stress detection model. The content of this chapter is submitted to npj Digital Medicine.

5.1. Problem statement

In recent years, the growing availability of wearable sensors has led to increased research towards the continuous, ambulatory monitoring of stress. However, detecting stress in daily life poses multiple challenges: First, the presence of physical activity can mask the effect of stress on the physiological signals [157]. Second, ambulatory measurements are more susceptible to (motion) artifacts, which imposes the need for accurate artifact-handling

techniques and signal-quality indicators [106]. Third, the lack of an objective stress reference. While cortisol has been reported as the main stress hormone, there is no unobtrusive cortisol measurement technique and it is often omitted in ambulant recordings. Instead, Ecological Momentary Assessments (EMAs), repetitive stress self-reports closely timed throughout the day, are the most commonly used [158].

So far, mainly small-scale studies, often underpowered, of 20-50 subjects have been conducted [53] [17] [107]. Multiple findings suggest that physiological responses to stress tend to be person-dependent [157, 128]. Therefore, in order to grasp variability among subjects and develop models that are generalizable on a large scale, large datasets are needed. Further, in the majority of ambulatory trials, subject's demographics, psychological baseline profiles (e.g. self-reported anxiety and depression levels) [53] [17], and context information [107] are not taken into account although these could improve personalization and classification performance, and provide actionable insights [14, 17].

We present the SWEET study (Stress in the Work EnvironmEnT): a comprehensive dataset to monitor stress responses in a free-living environment. From 1,002 subjects, we collected baseline psychological information (e.g. self-reported anxiety and depression levels) through an intake questionnaire, five consecutive days of free-living physiological data through wearables and smartphone-based contextual measurements (e.g. location), self-reported stress through EMAs and data from an application-based stress test.

5.2. Materials and methods

An experiment was conducted to investigate the physiological stress response in a free-living environment. The Medical Ethical Committee of KU Leuven approved the protocol and analysis methods of the experiment (protocol ID: S57916). In this section, the protocol and the sensing modalities are described. Further, the implementation of two quality indicators for SC and ECG is presented.

5.2.1. Data collection

Protocol

The trial was conducted with 1,002 subjects (484 male, 451 female, 67 NA: did not fill in questionnaire correctly), aged 39.4 ± 9.8, recruited in 11 technology-oriented, banking and public sector companies. Subjects were included if they were active employees at the time of the study, no other inclusion or exclusion criteria were applied. The experiment was conducted over a time span of two years and lasted five days per subject. The timeline of the study is depicted in Figure 5-1.

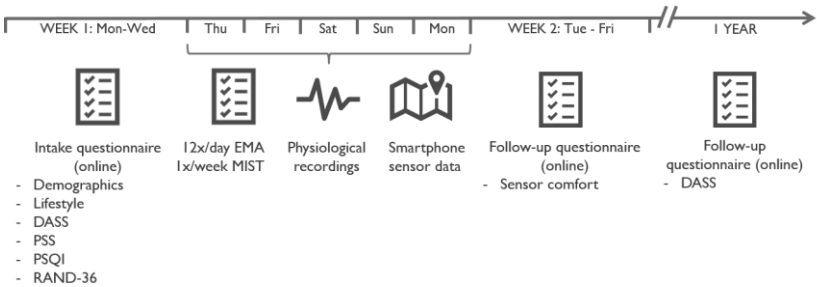


Figure 5-1: SWEET study protocol

Subjects were recruited through internal communication within the included companies. Detailed information about the experiment was given through a website and several info sessions. Persons willing to participate could subscribe on the website, where they could choose the week during which they wanted to participate to the experiment and leave their e-mail address. On Monday the week before the start of the experiment, subjects received an automatically generated e-mail with a reminder that the experiment would take place during the next week. On Monday in the week of the experiment subjects received an e-mail with the information to fill out intake questionnaire on the website (see *Sensing modalities - Questionnaires*). This needed to be done by Wednesday evening. If the information was still missing by Tuesday evening an e-mail with a reminder was sent automatically. On Thursday morning the subjects collected the sensors. For every subject a sensor package was prepared, containing the sensors for physiological measurements, a printed

user manual and a USB-stick with the user manual and a movie on how to apply the sensors.

The experiment took place from Thursday morning until Monday evening. The weekend was included to investigate the differences between a weekend and a working day. A smartphone application was used to trigger participants to fill out stress-related questionnaires and to collect contextual data (i.e. location, smartphone usage, audio-features, etc...) whenever participants gave permission. During the experiment the goal was to measure stress levels during daily life, therefore no interventions that could influence the subject's stress levels took place.

To assess individual physiological stress responses to a known common stressor, a short stress test was included in the smartphone application. The subject had to do this stress test during the first day of the experiment (Thursday) at a moment that fitted best (i.e. when he had time and was in a quiet environment). The stress test used was the Montreal Imaging Stress Task (MIST), which is based on the well-known Trier Social Stress Test and is explained in detail in [159].The MIST contains a series of computerized mental arithmetic challenges, along with social evaluative threat components that are built into the program (i.e. the application) [159]. The test consists of a five minute rest period (relaxing music and images), a five minute control period (simple mathematic tasks, no time restrictions or social control), five minute stress task (mathematic tasks with time restrictions and social control) and again five minute rest period (relaxing music and images). The flow of the MIST is depicted in Figure 5-2.

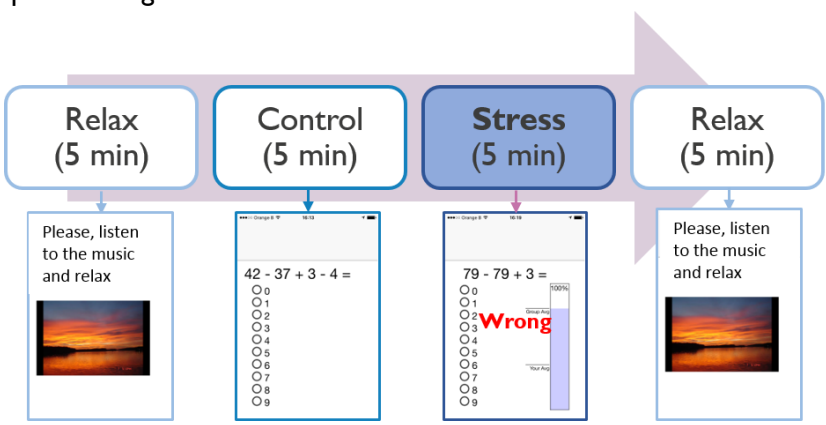


Figure 5-2: Montreal Imaging Stress Task (MIST)

On Monday evening the subsequent week participants had to return the sensor package. After the experiment, participants had to complete a questionnaire about sensor comfort on the website and a year later a reminder was sent to retake part of the intake questionnaire.

Sensing modalities

Three types of sensing modalities were used: questionnaires, as a gold standard for self-reported stress levels and for subject demographics and psychological background information; physiological recordings, to detect the physiological response to stress; and smartphone-based recordings, to measure context information and behavior which could be linked to stress, subject demographics and psychological background information.

Questionnaires – Three types of questionnaires were used: an intake questionnaire, annotation during the experiment and a follow-up questionnaire. All information and questionnaires were available in three languages, Dutch, French and English, to encourage multi-nationality among subjects. The intake and follow-up questionnaires were presented via the website, the annotation during the experiment via a smartphone application.
Before the start of the experiment an intake questionnaire was completed on the website. The first part inquired personal information such as age, gender, health problems, work situation, etc. Thereafter, four validated psychological questionnaires were used to assess baseline stress, depression, anxiety, sleep and general health levels. The first questionnaire was the Perceived Stress Scale (PSS) [13] which is a 10-item questionnaire, rated on a 5-point Likert scale for information about a person's perceived stress level over the last month. The second questionnaire was the Pittsburgh Sleep Quality Index (PSQI) [160]. This is a 10-item questionnaire for information about a person's sleep quality. This is of interest since research has shown that job stress can lead to poor sleep quality [161]. As third questionnaire the 21-item Depression Anxiety Stress Scale (DASS) [162] was used to measure the three related emotional states of depression, anxiety and stress. Finally the 36-item RAND-36 [163] questionnaire was used as health-related quality of life questionnaire. An overview of all questions with average population responses is presented in Appendix A.

During the five days of experiment, Ecological Momentary Assessments (EMAs) on a mobile application were used to assess self-reported stress. An overview of the EMAs is presented in Figure 5-3. Previous research has shown correlations between stress and sleep efficiency [164] and between stress and digestive diseases (e.g. irritable bowel syndrome) [165]. Therefore, each morning the sleep quality of the previous night was inquired (e.g. at what time did you go to bed? How long did it take to fall asleep?) and each evening gastro-intestinal symptoms were inquired based on the Leuven Postprandial Distress Scale [166]. Throughout the day the application sent twelve alarms for stress annotation. The alarms were sent at random times, but at least thirty minutes apart. When the subject did not fill out the annotation, one reminder was sent ten minutes later. If after forty minutes the annotation was still ignored, the questionnaire closed and missing values were assigned. The annotation existed of four short questions. The first was the Self-Assessment Manikin (SAM) [167] which is a visual scale to assess pleasure, arousal and dominance, i.e. affective emotions related to stress. The pleasure level could be used to differentiate "good" stress from "bad" stress, i.e. eustress vs. distress, where eustress reflects the transition of the body to a lower allostatic load (i.e. "the price the body pays for being forced to adapt to unfavorable psychosocial or physical situations" [168]) and distress to a higher allostatic load [168]. The next screen contained a drop-down menu to indicate the maximum stress level over the last hour on a 5-point scale (i.e. not at all, slightly, moderately, very and extremely stressed). Since eating and drinking behavior and physical activity can influence physiology [157] [65], the third and fourth questions were used to indicate food and beverage consumption (i.e. caffeine, alcohol, soft drinks, breakfast, lunch, dinner, snack or none) and activity levels (i.e. lying down, sitting, standing, walking, running, biking , driving the car or something else), for which subjects could select multiple answers.

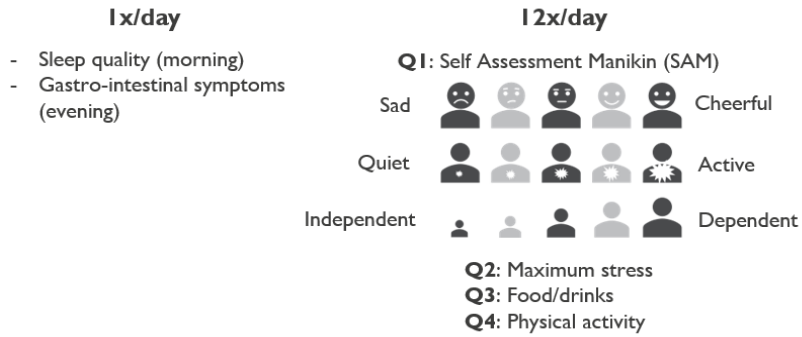


Figure 5-3: Overview of Ecological Momentary Assessments (EMAs).
Left: once per day sleep quality and gastro-intestinal symptoms, right: 12 times per day SAM, stress, food and activity annotation.

Two follow-up questionnaires were used. The first was presented to the subject right after the experiment and inquired about the sensor comfort. The second questionnaire was a copy of the DASS, presented a year after participation on the study. This questionnaire can be used to assess chronic stress and clinical outcomes such as depression.

Physiological recordings – Two sensors were used for physiological recordings (Figure 5-4). The first sensor is a chest patch to measure the electrocardiogram (ECG) and acceleration (ACC). It contained a sensor node designed to monitor ECG at 256 Hz and ACC at 32 Hz continuously for seven successive days. Subjects wore the patch the entire day and night. The sensor is not waterproof, so subjects received a waterproof cover to be able to take a shower without removing the patch. While practicing sports, subjects were advised to remove the sensor to avoid sweat from damaging the device. The second sensor is the imec's Chillband, a wrist worn device, designed for the measurement of skin conductance (SC), skin temperature (ST) and ACC. The SC was sampled at 256 Hz, ST at 1 Hz and ACC at 32 Hz. Subjects wore the sensor the entire day, but could take it off during the night. Subjects were asked to remove the band while taking a shower or during vigorous activities. The battery life of both sensors exceeded the duration of the experiment. Although both devices featured wireless connectivity, data were recorded on internal SD cards and uploaded to a central data platform at the end of the experiment.

Figure 5-4: Physiological recordings.
Left: chest patch for ECG and ACC, right: Chillband for SC, ST and ACC measurement

Smartphone recordings – The smartphone application used for the self-reporting (see *Questionnaires*), was also used for the measurement of context information. Subjects could choose to either use their own smartphone or to borrow one. This information is marked in the dataset, since it can influence phone use. Context information was only collected if subjects gave explicit additional consent. If not the app was only used for EMA annotation. The first context signal that was recorded is activity. Literature has shown that physical activity can decrease stress levels [169]. Since people often carry their smartphone with them, it can be investigated if activity measured from the acceleration sensors inside the smartphone is correlated with the detected/reported stress levels. Second, audio features were recorded. Many research has already indicated that social support has a large influence on stress and health related effects caused by stress [170]. Audio features can be used to detect conversations and social interaction. To ensure privacy, only audio-related features were stored (e.g. noise level), not the actual conversation. Third, location could be recorded. Based on the location it is possible to detect if a person is at work, at home, in the car, etc. The goal is to find the correlation between detected/reported stress levels and locations. To investigate this correlation, the coordinates of the smartphone location were saved every time the person answered a questionnaire and if battery life allowed, every 15 minutes. The actual coordinates of the location were anonymized by applying a random translation and rotation to the coordinates, distances between coordinates remained unchanged. Locations were clustered as unique stay locations, i.e. average location in more than 60 min within a radius of 1km and commuting. Fourth, if both the company and participant

allowed, also the timestamps of incoming and outgoing e-mails were stored. These can be used as a possible measure for workload. Finally, the app could monitor smartphone use, such as screen on or off, environment light and if the phone was locked or unlocked.

5.2.2. Quality indicators

Raw sensor data and subject self-assessments were synchronized using UTC timestamps. Data were stored as HDF5 files containing all synchronized physiological information[3]. Quality indicators were applied subsequently. Assessing the quality of the ECG and SC signals is necessary since these signals are prone to artifacts due to motion or incorrect sensor attachment.

The ECG quality indicator[4] is based on Orphanidou *et al.* [171] , which has shown a sensitivity for artifact detection of 94% and a specificity of 97%, and consists out of three rules and a template matching, verified on 10-second segments of ECG data: first, the extracted HR should be within 40 and 180 bpm. Second, the maximum gap between successive R-peaks cannot exceed 3 s. Third, the ratio of the maximum beat-to-beat interval to the minimum beat-to-beat interval within the segment should be less than 2.2. If all rules are satisfied, an adaptive QRS template matching is performed. The 10-second segment is either classified as of good or of bad quality.

In the SC quality indicator [126] for each 5-second window the ratio of lost versus overall signal is calculated. The signal is deemed lost if its value is below 0.001μS. If this ratio is above 0.9, the signal is classified as of bad quality. Next, the algorithm searches for anomalies. For each second the maximum increase of a signal value is set to 20% and the maximum decrease to 10%, as suggested by Boucsein *et al.* [172]. If SC values within the segment do not satisfy these conditions the segment is classified as of bad quality.

Previous research defined the ST range at the wrist between 20-40°C [173]. Therefore, ST values outside this range are classified as of bad quality.

[3] The data processing pipeline was implemented by Imen Chakroun

[4] The ECG quality indicator was implemented by Bishal Lamichhane

5.3. Results

The aim of current chapter is to investigate dataset quality and compliance. We used three sensing modalities: questionnaires, smartphone-based recordings and physiological recordings.
An overview of the intake questionnaire is available in Appendix A. Further, EMA was used to capture self-reported stress. A total of 23,429 stress reports were collected for 920 subjects, on average 25 (range: 1 – 54) per subject (Figure 5-5). This corresponds to an average compliance rate of 42% (range: 2%-90%). Further, self-reported stress levels are highly imbalanced: 53.4% no stress, 32.3% light stress, 11.4% moderate stress, 2.6% high stress and 0.3% extremely high stress.

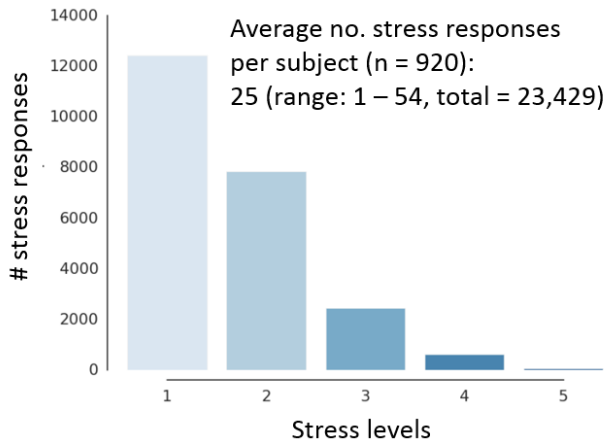


Figure 5-5: Self-reported stress levels of 1,002 subjects

From Thursday until Sunday subjects reported stress on average 6 (SD=2.68) to 6.20 (SD=2.86) times respectively, while on Monday compliance dropped to 4.7 times (SD = 2.14, Wilcoxon ranksum $p < 0.001$) as can be seen in Figure 5-6.

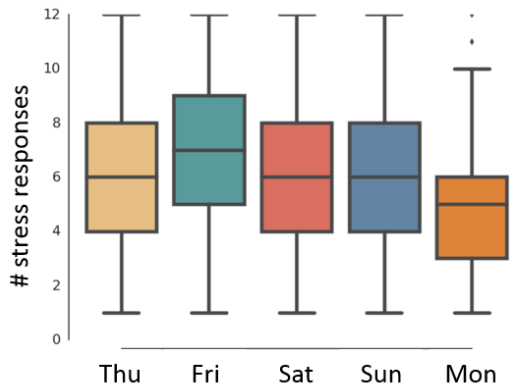


Figure 5-6: Annotation compliance throughout the five day experiment

When subjects approved, we used the smartphone application to collect contextual data. 720 subjects gave permission to monitor their location, for 612 subjects, due to battery lifetime restrictions, the location was only monitored while the app was open for answering a questionnaire, resulting in an average of 35 (range: 1-72) data points per subject (covering 0.02-1.2% of the time). One data point reflects one minute of information. The average coverage is calculated by dividing the number of data points (represented in timeframes of minutes) by the total time of the study in minutes. For 312 of these subjects location was monitored every 15 minutes, with an average of 452 (range: 1-3292) data points per subject (covering 0.02-52.8% of the time). Location information was more often available during the day than during night hours. On average, subjects spent time in 5 unique locations. Further, 501 subjects gave permission to monitor audio features. Of these only for 240 subjects actual data was recorded, due to additional phone privacy settings, with an average of 51 (range: 1-503) data points per subject (covering 0.02-8.1% of the time). No actual conversations were recorded, rather informative features such as amplitude and variance of the sound signal and likelihood of voice activity. Call and SMS logs were available for 183 and 201 subjects respectively, with an average of 14 (range: 1-78) call logs and 23 (range: 1-154) SMS logs. Information on ambient light was available for 490 subjects, temperature for 4 subjects, air pressure for 50 subjects and screen mode (on/off) for 569 subjects. An overview of these data are presented in Table 5-1.

Table 5-1: Overview of smartphone-based sensor data

	Nr of subjects with at least one datapoint recorded	Average nr of data points per subject	Average coverage across 5 days of trial	Unique locations per subject
Location recorded when app was open	612	Mean = 35, range: 1-72	Mean = 0.6%, range: 0.02-1.2%	5
Location recorded continuously	312	Mean = 452, range: 1-3292	Mean = 7.2%, range: 0.02-52.8%	5
Audio features	240	Mean = 51, range: 1-503	Mean = 0.8%, range: 0.02-8.1%	NA
SMS Logs	201	Mean = 23, range: 1-154	NA	NA
Call Logs	183	Mean = 14, range: 1-78	NA	NA
Ambient light	490	Mean = 557, range: 1-3540	Mean = 8.9%, range: 0.02-56.7%	NA
Air pressure	50	Mean = 1180, range: 1-5427	Mean = 18.9%, range: 0.02-87.0%	NA
Temperature	4	Mean = 106, range: 3-307	Mean = 1.7%, range: 0.05-4.9%	NA
Screen mode (on/off)	569	Mean = 1220, range: 1-6152	Mean = 19.6%, range: 0.02-98.6%	NA

To capture their physiology, subjects wore a chest patch and a wristband (Chillband). 845 subjects wore both sensors, 61 only wore the chest patch, 60 only wore the Chillband, 36 subjects did not wear the sensors and only recorded smartphone data. The chest patch was worn continuously and data of 4356 days and 2979 nights were collected, across all subjects. For the Chillband, data of 4366 days and 1744 nights was collected. When wearing the sensors the chest patch had on average 86.3 ± 8.2% good quality data, the Chillband 96.4 ± 2.2% (Table 5-2). In Figure 5-7 two examples of an ECG and SC signal with and without quality indicator applied are shown. It can be seen that abnormal signals (outliers) are efficiently removed. However, due to the

10s window segment, also some high quality signals are falsely removed, as indicated by the red box.

773 subjects gave feedback on sensor comfort: 18% estimated it likely they would wear the Chillband daily in the future, while only 8% would wear the chest patch daily. The main reasons for not wearing the Chillband were that the band was too big (69.4%) and not comfortable (45.6%); the chest patch was found not comfortable (67.3%, 14% of the subjects rated their level of irritation 8 or higher, scale of 1-10), too visible (39.1%) or too big (38.2%).

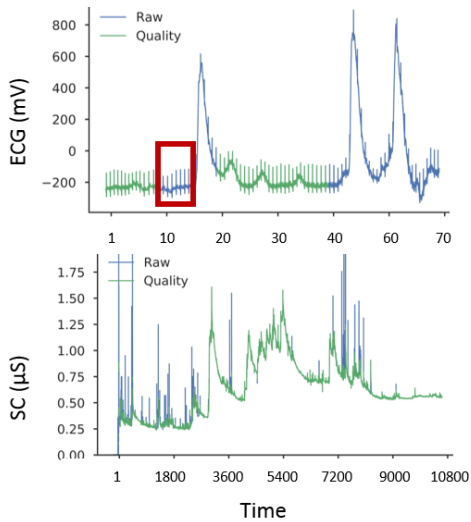


Figure 5-7: ECG and SC signals, raw (blue) and with quality indicator (green). The red box indicates the falsely removed high quality signals.

Table 5-2: Overview sensor quality Chillband and chest patch

Sensor	Chillband (n = 905)	Chest patch (n = 924)
Days - mean (range)	4.5 (1 – 5)	4.4 (1 – 6)
Hours - mean (range)	107 (1 – 130)	97 (1 – 184)
QI (0 - 100) - mean (range)	96.4 (82.6 – 99.9)	86.3 (31.5 – 98.8)

5.4. Discussion

The results have indicated several lessons learned for future use of EMA and physiological sensing in large-scale studies. We have seen that compliance to the EMAs was 42%, which is less than commonly reported rates between 65-85% [174]. Earlier research has shown that compliance during longitudinal trials drops when the novelty aspect disappears [175]. A possible solution to increase adherence is to provide feedback. However, to avoid influencing subjects' behavior, no feedback was provided during this study. Further, the dataset has also been shown to be highly imbalanced in terms of self-reported stress responses, as was also found in previous ambulatory stress research [107]. A possible solution to improve both adherence and retrieve more balanced datasets is the use of context-aware EMAs. These permit more sophisticated event-based sampling where algorithms are used to detect specific events (e.g. HR increase, running,...) and can provide the signaling cue [176]. This way the situation of interest (i.e. stress) can be monitored without unnecessarily burdening the subject. Although this technique already exists more than a decade, it is still rarely implemented in ambulatory studies (e.g. using location-based notifications for smokers [177] and alcoholics [178]). Another approach is to eliminate traditional questionnaires and use the subject's digital footprint (i.e. smartphone use, facebook profiles, Twitter, etc.) which may provide a convenient and reliable way to measure psychological traits at a low cost [179]. Such automated assessment could prove to be more accurate and less prone to cheating and misrepresentation than traditional questionnaires.

Further, it can be seen that both physiological and contextual data streams contain missing data. A first cause are privacy considerations, about 70% of subjects gave permission to record their location and 50% gave permission to record audio information. Researchers need to take the increasing awareness and legislation regarding data privacy and data protection (e.g. the General Data Protection Regulation of the EU) [180] into strong consideration when developing trials and mobile health (mHealth) solutions. Other causes of missing data are sensor-related such as sensor/smartphone breakdown, incorrect sensor use or the termination of sensor use. We investigated the data quality when wearing the sensor, which was on average 86.3% for the chest patch and 96.4% for the Chillband. Based on Figure 5-7 it could be seen

that outliers were efficiently removed. However, also some high quality ECG signals were falsely removed due to the 10s window segment. Improvements on the quality indicators could be made by using shorter time windows and applying a continuous quality indicator scale rather than a discrete (i.e. good vs bad). A new approach for an ECG quality indicator is suggested by Moeyersons et al. [181][5].

We also asked subjects to rate the sensor comfort. About 20% of subjects estimated it likely they would wear the wristband on a daily basis in the future, which is in line with a report of Forbes mentioning in 2016 one in six consumers owned and used wearable technology [182]. It has also been shown that worldwide most users are young (48 % between 18-34) [182] and about a third of owners of smart wearables stops using them within six months after purchase [183]. This calls for action of technology developers and researchers, our survey showed that most subjects would not wear the sensors because they are too big or not comfortable, less than 10% indicated they found the sensors not useful. Although these numbers could be biased due to the fact that subjects participating in this study are probably more interested in wearable technology compared to the general population, they do signal that focus of research should be on design and comfort of wearables, next to performance (which is currently the primary focus).

5.5. Conclusion

To assess stress we collected a dataset of 1,002 subjects during five consecutive days, including a wide variety of subject background information, physiological data in ambulatory settings and smartphone-based self-reports and contextual information. Investigation of the data underlines the importance of feedback and motivation to increase compliance to the EMAs. Overall, we presented a large-scale study with high quality physiological and self-reported stress data which can be used to develop models for stress detection and to investigate the link between behavior and health indicators.

[5] This publication is currently under consideration at Computer Methods and Programs in Biomedicine

Chapter 6: Linking behavior, health indicators and physiology: towards digital phenotyping

In Chapter 5 we presented the SWEET study protocol and dataset quality. In current chapter we use the demographics, psychological background information, physiological and contextual information to infer behavior patterns. Current analysis was performed by Elena Smets with promotor Chris Van Hoof in collaboration with Emmanuel Rios Velazquez and Giuseppina Schiavone. Figures 6-1 to 6-4 were produced by Emmanuel Rios Velazquez, Figure 6-5 was produced by Giuseppina Schiavone, Figure 6-6 was produced by Elena Smets. We present multiple interactions between behavior and health, confirming findings in earlier studies. Significant differences are found between physiological signals during self-reported stress levels, confirming for the first time on a large scale the potential of physiological stress detection in daily life. We highlight the need for personalized models to detect stress, based on the development of digital phenotypes. The content of this chapter is submitted to npj Digital Medicine.

6.1. Problem statement

A phenotype is "The set of observable characteristics of an individual resulting from the interaction of its genotype with the environment" [184]. In 2015, Jain *et al.* [185] introduced the 'digital phenotype'. Through social media, wearable technologies and mobile devices, there is a wide variety of health-related data that can extend our assessment of human illness beyond traditional examinations. These data can fundamentally change our understanding of diseases and provide new insights and hypotheses. It can be used for disease interception, treatment and chronic disease management. For example,

Health-Map is a public website, created in 2006 by a team of data scientists, epidemiologists and researchers, which utilizes informal online information for disease outbreak monitoring (e.g. the flu) and real-time assessment of public health threats [186].
Digital phenotypes could be useful to detect stress and even more to prevent and intercept co-morbidities linked to stress on a large scale. These include an increased risk at cardio-vascular disease, gastro-intestinal disorders such as irritable bowel syndrome, obesity and a variety of neurological disorders including Alzheimer's disease [187].
Although the term 'digital phenotype' is fairly new, the concept has been introduced already for a longer time. The use of wearables to detect stress is also a form of digital phenotyping. However, current research has mainly focused on using only one type of digital technology, e.g. using wearables for physiological stress detection [107], or using pressure-sensitive key-boards to discriminate between relax and stress in computer users [188]. Although the importance of personalization in stress detection has already been highlighted [157, 128], so far very few studies have actually incorporated subject background or context information to improve stress detection performance and to gain more insight in digital phenotypes for stress detection.
In the SWEET study we collected a wide variety of digital and non-digital information, defining a subject's personal traits, behavior and physiology. In current chapter we present the connection between behavior and health indicators, between stress and physiology, and we present an approach towards digital phenotypes for psychophysiological stress detection.

6.2. Materials and methods

The data collection protocol of the SWEET study has been presented in Chapter 5. In this section we further elaborate on physiological feature computation and analysis methods.

6.2.1. Feature computation

In Chapter 3, we investigated 22 physiological features in a laboratory setting. Based on the results of Chapter 3 and based on new findings in literature,

including research specifically focusing on ambulatory stress detection, 18 physiological features [17] [101] [189] [97] [103] [95] [53] [130] [63] [129] were included in our study: 6 features for ECG, including mean HR and time and frequency domain HRV features, 8 SC features, including tonic and phasic features, and 4 ST features. For accelerometer-based activity we included the standard deviation of the accelerometer magnitude (ACC SD) [190]. To compute HR and HRV, R peaks were detected from the ECG signal using the beat detector for ambulatory cardiac monitoring developed by Romero *et al.* [191] with a sensitivity of 99.86% for an ECG dataset with high levels of activity. A complete list of all features is available in Table 6-1. Features were calculated in a window of 5 minutes with 4 minutes overlap. This is the minimum window required to calculate HRV features such as the root mean squared difference of successive RR intervals (RMSSD) [192] due to the inherent regulation periodicity [193]. The 4 minutes overlap was set to obtain a resolution of smoothed processed data of one sample per minute.

Table 6-1: List of features computed for each sensing modality (SWEET study)

Nr.	Feature	Abbreviation	Reference
1	Mean heart rate (HR)	ECG mean HR	[101] [189] [97]
2	Standard deviation of RR intervals	ECG SDNN	[97] [103] [17]
3	Root mean square of successive RR differences	ECG RMSSD	[103] [17]
4	Low frequency signal (power in the 0.04-0.015 Hz band)	ECG LF	[101] [189] [103] [17] [95]
5	High frequency signal (power in the 0.15-0.4 Hz band)	ECG HF	[101] [189] [103] [17] [95]
6	Ratio of low and high frequency	ECG LFHF	[101] [189] [97] [103] [17] [95]
7	SC level – average SC	SC mean	[97] [103] [130]
8	Phasic SC – signal power of the phasic SC signal (0.16-2.1 Hz)	SC phasic	[63]

9	SC response rate – number of SC responses in window divided by the total length of the window (i.e. responses per second)	SC RR	[103]
10	SC second difference - signal power in second difference from the SC signal	SC diff2	[129]
11	SC response - number of SC responses	SC R	[97] [103] [130] [63] [53]
12	SC magnitude - the sum of the magnitudes of SC responses	SC mag	[97] [103] [130] [63] [53]
13	SC duration - the sum of the duration of SC responses in seconds	SC dur	[97] [103] [63] [53]
14	SC area - the sum of the area of SC responses in seconds. The area is defined using the triangular method (1/2*SC mag*SC dur)	SC area	[53]
15	Mean ST	ST mean	[95]
16	Median ST	ST median	/
17	Standard deviation ST	ST SD	[95]
18	Slope of the ST – slope of a straight line fitted through the data	ST slope	[97]
19	Standard deviation of the magnitude of accelerometer signal – a measure for movement intensity	ACC SD	[190]

6.2.2. Analysis methods

Statistical tests were performed using the nonparametric Wilcoxon ranksum test for continuous variables. To assess differences of continuous variables across multiple demographic groups we used the Kruskal-Wallis test. The X^2 test was used for comparisons of categorical variables. Two-sided p-values of

<0.05 were considered statistically significant. All statistical tests requiring multiple comparisons were corrected based on the Benjamini-Hochberg procedure.

Associations between longitudinal data (e.g. questionnaires presented 12 times per day, or continuous wearable data) were assessed using linear mixed effects models, using the lme4 R package [194], with self-reported pleasure or continuous wearable feature data as fixed effects and the subjects as random effect. A gaussian family was used to model continuous variables (e.g. ACC SD), while a Poisson family was used to model stress responses. An ANOVA test was used to assess whether model parameters differed significantly from zero by comparing the change in model performance (Akaike's information criterion52) when a fixed effect (e.g. pleasure) was excluded from the model. Correlations between stationary data (e.g. questionnaires with single responses) were calculated using the Spearman correlation coefficient (r).

Location data were anonymized based on a random translation and rotation. Locations were clustered as unique stay locations, i.e. average location in more than 60 min within a radius of 1km and commuting.

A machine learning model was developed to predict self-reported stress levels based on physiological responses. Only good quality physiological data (good QI in ≥80% of data points in the 5 minutes window) were used and features were normalized (z-normalization) per subject. Redundant features were removed based on correlations (max r = 0.7), resulting in a reduced feature set. Since self-reported stress responses (based on the maximum stress during the last hour, i.e. Q2 inFigure 5-3) were highly imbalanced (Figure 5-5), the three highest stress levels were merged, representing 14.3% of the data, so that three, instead of five, levels of stress (S1 = no stress, S2 = light stress, S3 = high stress) were considered.

Based on these data, associations between physiological features and self-reported stress levels were investigated. For each stress level the average of the normalized features across the entire population was calculated. Additionally, the average during the night (N) (00-06 am) was included as baseline. For each feature, the differences between averages of different states were computed: N-S1, N-S2, N-S3, S1-S2, S2-S3 and S1-S3. A Wilcoxon-test was performed to investigate significant differences and corrected for multiple comparisons.

Additionally, machine learning models were calculated. Subjects reporting only one stress level (e.g. only 'no stress') were discarded. Since self-reported stress levels reflect the situation of the last hour, the stress value reported was registered for the 60 data points pertaining to that entire hour. We included only data for windows of at least 10 minutes of good quality and low physical activity (ACC SD ≤ 0.04, based on [190] and adapted according to subject's self-reported activity levels). A false discovery rate supervised feature selection was applied on the training set, according to the Benjamini-Hochberg procedure (python scikit-learn, alpha=0.05). We trained Random Forest (RF) models with 100 trees (based on out-of-bag samples) and balanced class weight, in a leave-one-subject-out approach (python scikit-learn, RandomForestClassifier). This means a model was trained based on the data of all subjects but one and tested on the data of that subject. This procedure was repeated until all subjects were tested exactly once. RF models were chosen based on their relatively high performance in Chapter 3 and based on an initial comparison on current dataset in which RF models outperformed SVMs and LR. We used the F1-score, a weighted average between precision and recall, to evaluate the model's performance on the left-out-subject. As comparison, we also calculated the F1-score for all subjects if the RF model classified all samples as the majority class, i.e. S1.

We further evaluated subject's physiological response, demographics and psychological information based on individual model performance. Subjects were categorized in groups of low performance, with F1-score < 0.33 (performance as good as random), medium performance, with 0.33 < F1-score < 0.66 and high performance, with F1-score > 0.66. For each group we evaluated three characteristics: first, we evaluated the imbalance of the self-reported stress levels, as a higher imbalance (e.g. mainly reporting S1), could lead to a higher classification performance. Second, we investigated the average dynamic range of each group, where the dynamic range represents the average difference per physiological feature of each group between low (S1) and high (S3) self-reported stress levels. A higher dynamic range could be beneficial for model performance, as the feature can better differentiate between low and high stress. Third, we investigated subject's demographics and psychological information based on the intake questionnaire. A Wilcoxon ranksum test was performed to investigate significant differences across low

and high performance groups, we corrected for multiple comparisons. All data analyses were performed using Python (version 2.7).

6.3. Results and discussion

6.3.1. Associations between physiology, context and behavior

We aim to use this comprehensive dataset to investigate correlations between physiology, context and behavior in order to improve our understanding of stress in daily life.

Through EMAs we daily asked questions related to stress, activity, food and beverage consumption, sleep quality and gastro-intestinal symptoms. Based on self-reported wake-up and bed times, circadian rhythms are evident, with lower mean HR and higher mean SC and ST during the night as compared with during the day. The average values for the population are presented in Table 6-2, with significant differences for all physiological signals (Wilcoxon ranksum $p < 0.001$).

Table 6-2: Day and night time physiology.

	Day (06am-23:59pm)	Night (00:06am)
Mean HR	74.6±12.7	63.0±10.2
Mean SC	1.7±2.7	2.8±3.4
Mean ST	31.4±2.09	33.1±2.6

During weekdays (i.e. Thursday, Friday and Monday) consumption of caffeinated beverages or breakfast corresponded to higher stress levels (caffeine: 1.84±0.81, breakfast: 1.87±0.81, average: 1.77±0.83, Wilcoxon ranksum $p < 0.001$), while dinner or alcohol consumption, corresponded to lower stress levels (dinner: 1.51±0.71, alcohol: 1.30±0.64, average: 1.77±0.83, Wilcoxon ranksum $p < 0.001$). During the weekend (i.e. Saturday and Sunday), the consumption of alcohol was associated with lower stress levels (alcohol: 1.34±0.66, average: 1.45±0.71, Wilcoxon ranksum $p = 0.001$), other reported consumptions did not show significant differences. A possible confounder here

is time of the day since breakfast is consumed most in the morning (82% of reports between 6-10h), and alcohol and dinner most in the evening (alcohol: 61% of reports between 18-22h, dinner: 65% of reports between 18-22h), caffeine was reported equally throughout the day, but less during the evening (32% of reports between 6-10h, 37% between 10-14h, 23% between 14-18h and 7% between 18-22h).

Further, linear mixed effects models were computed to investigate associations between repeated measures. A significant negative association between self-reported stress and self-reported pleasure (based on the SAM, see Figure 5-3) was observed, with higher levels of self-reported stress corresponding to decreasing levels of pleasure (Table 6-3). It can be speculated that rating of high self-reported stress is likely associated to the feeling of distress (negative stress) rather than eustress (positive stress). The standard deviation of the magnitude of acceleration (ACC SD), was associated with intensity of movement as ACC SD was higher during self-reported high-intensity activities (low-intensity, i.e. lying, sitting and standing: 0.0175±0.0089, high-intensity, i.e. walking, running, biking, driving car and other activities: 0.0189±0.0096; kruskal-wallis $p < 0.001$) and HR and SC features increased with ACC SD (Table 6-3) while ST decreased with ACC SD (Table 6-3). This illustrates the challenge of differentiating physiological changes caused by physical activity from those caused by stress. Therefore, to account for the cofounding effect of physical activity on physiology and stress, we excluded segments of high activity in the subsequent analysis.

Finally, increasing activity levels (ACC SD), decreased the quality of physiological signals (Table 6-3), an issue inherent to the free-living nature of the study.

A representative instance of five days of measurements of physiological data, acceleration, self-reports and smart-phone sensor data is shown in Figure 6-1.

Table 6-3: Effects of repeated measures.
Results of the linear mixed effects models for repeated measures. For each model, the fixed effect coefficient is presented with standard error (B±SE) and inferential statistics on the significance of the effect, which were calculated by testing the change in model performance (based on Akaike's information criterion) when a given predictor (e.g. pleasure) was excluded from the model using an ANOVA test.

Formula	B±SE	Inferential statistics		
		Test statistic and df	p-value	significance
Stress ~ Pleasure + (1 | subject)	-0.22±0.01	$X^2(2,3)=1353.5$	<0.001	*
SC mean ~ ACC SD + (1 | subject)	0.93±0.02	$X^2(3,4)=2448.6$	<0.001	*
HR mean ~ ACC SD + (1 | subject)	12.04±0.02	$X^2(3,4)=488051$	<0.001	*
ST mean ~ ACC SD + (1 | subject)	-5.61±0.02	$X^2(3,4)=87378$	<0.001	*
SC Quality ~ ACC SD + (1 | subject)	-0.23±0.0008	$X^2(3,4)=90652$	<0.001	*
ECG Quality ~ ACC SD + (1 | subject)	-0.33±0.0009	$X^2(3,4)=115642$	<0.001	*

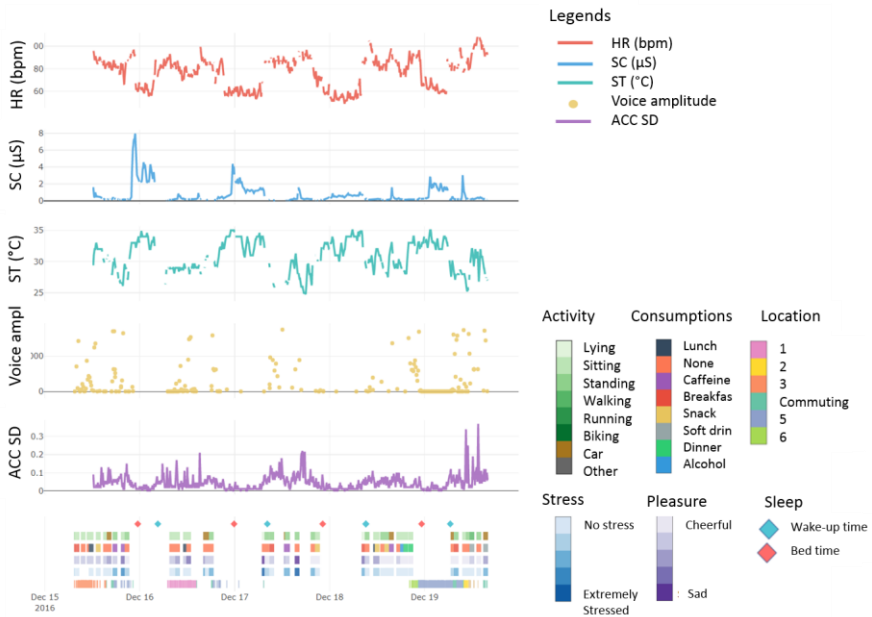


Figure 6-1: Physiology and context timeline of one subject.

Timeline over 5 days of measurements depicting daily profiles of one feature per physiological signal, activity and context data. For ECG, SC and ST signals, only good quality data (QI > 0.8) are shown. For visual inspection, these physiological signals as well as ACC SD were smoothed. Self-reported annotations (stress, pleasure, activity, consumptions and wake up/bed times) as well as location are indicated with vertical lines when available. Location data are indicated as unique stay locations or commuting locations. An online version of this figure can be downloaded from: https://drive.google.com/open?id=1-qhskpGn1pVefyFI4MUadTu6ubn7XC_e.

6.3.2. Associations between questionnaire-based lifestyle and health indicators

In the past mostly non-digital, questionnaire-based information has been used to assess lifestyle and health. In our study we collected such information in the intake questionnaire. Here we investigate if findings based on these questionnaires confirm existing literature linking lifestyle to health.

Data on lifestyle, general health and health indicators (validated psychological questionnaires, i.e. PSQI, DASS, PSS and RAND-36) was available for 932 subjects. Associations between several aspects of the subject's lifestyle (sports, diet and habits) and health indicators (RAND-36, PSS and PSQI) are investigated.

RAND-36 taps eight health concepts, ranging from physical functioning to emotional well-being and social functioning perceptions [163]. We found a positive correlation between energy levels and emotional wellbeing (r=0.66), with increased levels of both health indicators for subjects who more often practice sports (>5 weekly sports hours; emotional well-being: 78.0±14.0; energy: 69.2±16.3) compared to subjects who do not exercise often (0-1 weekly sports hours; emotional well-being: 70.0±14.5; energy: 56.9±17.8; kruskal-wallis p = 0.004, Figure 6-2A). This positive relation between sports and emotional wellbeing has long been recognized and is here confirmed [195]. Furthermore, self-perceived stress (based on PSS) was negatively correlated (r = -0.75) with emotional wellbeing. PSS was higher for subjects who reported medication intake (15.2±6.4), compared to subjects with no medication (14.0±5.9, ranksums p = 0.013, Figure 6-2B).

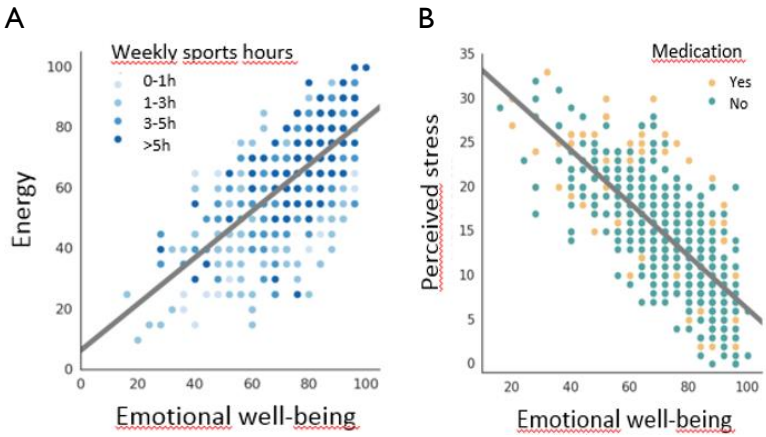


Figure 6-2: Associations between questionnaire-based lifestyle and health indicators – emotional wellbeing.
A) Correlation between emotional wellbeing and energy reinforced by practicing sports. B) Correlation between perceived stress and emotional well-being, reinforced by medication intake.

Subjects who smoke tended to report higher levels of self-perceived stress (PSS), and tended to rate their lifestyle as being less healthy (RAND36 General Health: 35.0±7.1) than those who do not smoke and feel less stressed (RAND36 General Health: 68.4±16.3; kruskal-wallis p = 0.07, Figure 6-3A). These results are in line with earlier reports indicating that a one-unit increase in PSS results in a 5% increased odds of smoking [196], and that people who smoke rate their lifestyle as less healthy than people who do not [197].
We found no correlation between BMI and general health (r = -0.12); however, we observed decreasing general health values with increasing number of take-out food times through the week (No take-out: 70.4±15.9 vs ≥5 times: 62.5±18.1; kruskal-wallis p = 0.008; Figure 6-3B).

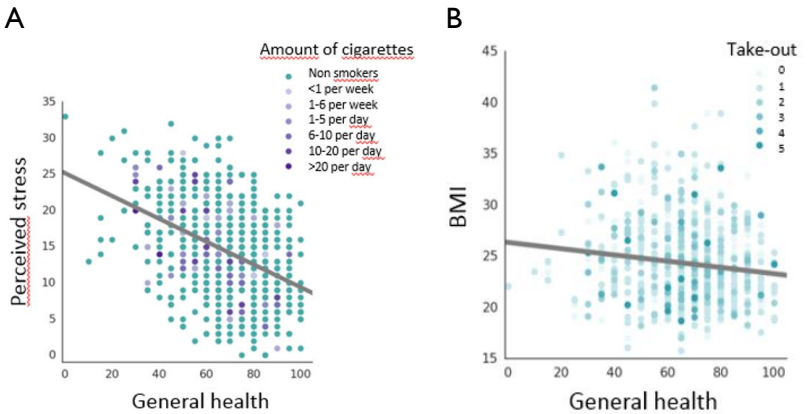


Figure 6-3: Associations between questionnaire-based lifestyle and health indicators – general health. A) Negative correlation between perceived stress and general health, reinforced by smoking. B) No significant correlation between BMI and general health, but worse general health levels for people who eat more take-out food.

There was no correlation between caffeinated beverages consumption and PSQI scores (PSQI scores higher than 5 indicate worse sleep quality; Figure 6-4A), although literature indicates that people with lower sleep quality on average consume more caffeine [198]. Women tend to have inferior sleep quality than men (5.2±2.8 vs 4.7±2.4, respectively; ranksums p = 0.023), as is reported previously [199]. We also found a negative correlation between energy levels and PSQI, indicating that subjects with inferior sleep quality have lower energy (r=-0.47). There was no significant difference on PSQI with alcohol intake (ranksums p = 0.13; Figure 6-4B). Literature suggests that for non-dependent alcohol users (e.g. light/occasional, habitual weekend use), alcohol consumption just before bedtime can reduce sleep quality [200]. However, the impact of afternoon or early evening alcohol consumption on sleep quality is not yet well understood [200].

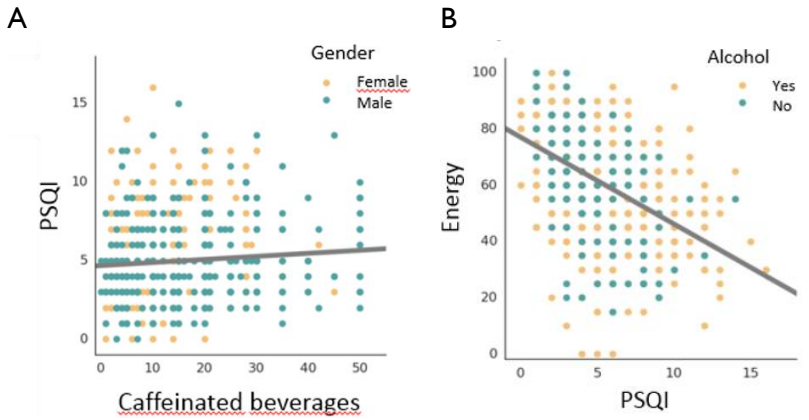


Figure 6-4: Associations between questionnaire-based lifestyle and health indicators – PSQI. A) No significant correlation between PSQI and caffeine consumption, women showed worse sleep quality. B) Negative correlation between energy and PSQI, not linked to alcohol consumption.

All together these findings represent a large-scale verification supporting previous work and confirm the value of data sampled with validated questionnaires: lifestyle and health indicators are strongly linked, underlining the need for behaviour change interventions that support preventive health.

6.3.3. Associations between physiological signals and self-reported stress levels

We aim to use physiological patterns to develop models for psychophysiological stress detection. Therefore, we show here a landscape of the associations between physiological features and self-reported stress levels, in a heathy population. An overview of all physiological features calculated is presented in Table 6-1. Mean differences of independent physiological features ($r < 0.7$), normalized per subject, across self-reported daytime stress levels (S1, S2 and S3) and nighttime (00-06 am, N), included as a baseline rest condition, are shown in Figure 6-5. Population variations (averages and 95% CI) of all physiological features across self-reported stress levels and nighttime are presented in Table 6-4.

For all the time instances, S1 to S3 and N, only periods in which the activity level was lower than the empirical threshold (ACC SD < 0.04) and good quality (Quality > 0.8) data were considered to exclude artifacts and physiology

variations due to physical activity. All features, except ECG LF and ECG HF were significantly different during nighttime (N) compared to during daytime self-reported stress levels (S1, S2, S3) (Figure 6-5). ECG LFHF, mean HR, SC area and ST SD were lower at night. The slope and median ST and SC phasic were higher at night.

Additionally, mean HR was significantly lower in S1 as compared to S3. Mean HR has a strong negative correlation with RMSSD (mean HR, RMSSD: r = -0.99), which is significantly higher in S1 as compared to S3 (see Table 6-4). These results confirm findings in laboratory studies reporting an increase in HR and decrease in HRV with increasing stress levels [201] [202] [203]. The frequency domain HRV features, i.e. the LF signal, HF signal and the ratio of LF and HF signals, did not change significantly during S1 as compared to S2 and S3. In literature, the HF component is thought to represent the cardiac parasympathetic nerve activity, which is active during rest conditions, and the LF component to represent the sympathetic system, which is active during stress conditions [204]. The LF and LFHF components are therefore expected to be higher during stress conditions and the HF component lower [204]. However, varying results have been reported in literature and in general RMSSD has been reported to be more reliable than LFHF [205] [204], in particular because of the mechanical effects of respiration on HF power and the influence of the prevailing heart rate on LF power [204]. Furthermore, SC area was lower in S1 compared to S3, as reported previously in [206]. SC phasic was lower in S1 compared to S2 and S3, as expected based on previous laboratory research [63] [207], indicating that higher stress levels are associated with higher power of the phasic SC component. Finally, the ST median and ST SD were higher in S1 compared to S2 and S3, which indicates that ST amplitude and variation decrease with stress [94]. For most of the features no significant differences were found between S2 and S3. This could either indicate that in general subjects have difficulties in making distinctions between light and high stress levels or that physiological features cannot distinguish between these levels at a population level.

Overall, the physiological signals measured in daily life showed significant differences between night and different stress levels, in line with previous findings of laboratory studies. These results confirm on a large scale the potential of physiological signals for detecting stress in daily life.

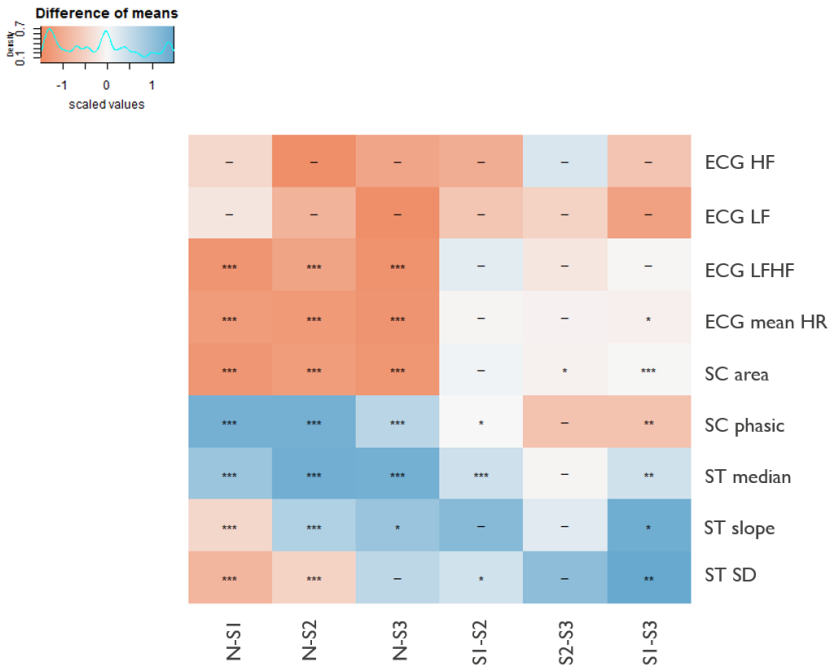


Figure 6-5: Associations between physiological features and self-reported stress levels.
*Each row represents a physiological feature, columns represent the difference of normalized features during the night (N) (00-06 am) and stress levels (S1, S2 and S3). Colors indicate positive (blue) or negative (red) differences. For example, SC phasic is significantly higher blue) during the night as compared to during all reported stress levels, and significantly lower (red) during S1 as compared to S3. Symbols: *=p<0.05, **=p<0.005,***=p<0.0005.*

Table 6-4: Physiological features across self-reported stress levels. Population mean and 95% confidence interval (CI) of all physiological features during the night (00-06 am) and different stress levels (S1, S2 and S3).

	Night (**mean**, 95% CI)	S1 (**mean**, 95% CI)	S2 (**mean**, 95% CI)	S3 (**mean**, 95% CI)
ECG mean HR	**62.1** (48.5 - 78.2)	**72.4** (55.7 - 88.8)	**73.0** (56.0 - 90.7)	**74.7** (56.3 - 94.8)
ECG SDNN	**77.1** (39.9 - 132)	**71.2** (39.4 - 117)	**72.0** (38.6 - 120)	**71.5** (38.7 - 118)
ECG RMSSD	**993** (772 - 1257)	**853** (684 - 1090)	**845** (666 - 1078)	**824** (638 - 1070)
ECG HF ($x10^{-3}$)	**0.78** (0.11 - 3.0)	**0.67** (0.11 - 2.2)	**0.71** (0.09 - 2.3)	**0.68** (0.09 - 2.2)
ECG LF ($x10^{-3}$)	**1.2** (0.27 - 3.2)	**1.1** (0.25 - 2.8)	**1.2** (0.23 - 2.8)	**1.2** (0.24 - 3.1)
ECG LFHF	**3.3** (0.65 - 9.6)	**3.6** (1.0 - 9.5)	**3.6** (1.0 - 9.4)	**3.5** (0.85 - 9.6)
SC mean	**1.9** (0.018 - 8.0)	**1.4** (0.05 - 6.5)	**1.4** (0.044 - 6.0)	**1.6** (0.027 - 7.9)
SC phasic	**12.4** (0 - 78.6)	**8.4** (0 - 57.3)	**8.7** (0 - 62.2)	**9.6** (0 - 75.0)
SC RR ($x10^{-2}$)	**2.3** (0 - 7.8)	**2.7** (0.11 - 7.9)	**2.8** (0.097 - 9.2)	**3.1** (0.026-12.1)
SC diff2 ($x10^{-9}$)	**20.5** (0.0010 - 103)	**34.1** (0.022 -257)	**33.3** (0.011 -238)	**51.8** (0.010 -293)
SC R	**5.8** (0 - 34.5)	**14.3** (0.31 - 51.2)	**13.8** (0.098-60.4)	**15.8** (0.044-74.4)
SC mag	**64.1** (0 - 274)	**141** (1.3 - 714)	**129** (0.32 - 729)	**145** (0.044-1000)
SC dur	**545** (0 - 3315)	**1377** (32.9 -4961)	**1362** (10.2 -5522)	**1533** (3.2 - 7068)
SC area	**0.57** (0 - 2.9)	**2.7** (0 - 14.8)	**2.6** (0 - 18.1)	**2.8** (0 - 20.0)
ST mean	**31.9** (22.5 - 34.8)	**31.5** (28.7 - 33.5)	**31.8** (28.4 - 33.6)	**31.1** (28.0 - 33.7)
ST median	**31.9** (22.5 - 34.8)	**31.5** (28.7 - 33.5)	**31.8** (28.4 - 33.6)	**31.1** (28.0 - 33.7)
ST SD	**0.13** (0.01 - 0.39)	**0.14** (0.03 - 0.28)	**0.13** (0.01 - 0.29)	**0.14** (0 - 0.41)
ST slope ($x10^{-3}$)	**0.37** (-0.49 - 3.1)	**0.38** (-0.24 - 1.4)	**0.38** (-0.34 - 1.8)	**0.41** (-0.82 - 2.7)

6.3.4. Digital phenotypes in physiological stress detection

We used a data-driven approach to uncover digital phenotypes of subjects' daily life stress responses. We developed random forest models using a leave-one-subject-out cross-validation to link physiological features to self-reported stress. We used the classifiers' performances to identify and characterize digital phenotypes representing subjects with similar psychological baseline, physiological responses to stress and health indicators.

Only good quality (Quality > 0.8) and low activity (ACC SD < 0.04) data were included for 568 subjects, with complete data (i.e. simultaneous continuous recording from wearables and EMAs). The remaining subjects had missing data in one of the two sensors or lacked mobile EMA data, and were not included in this analysis. The classification performance, as calculated using the average F1-score across all subjects, was 0.43 (95% CI: 0.05-0.86), which is slightly better than the F1-score of 0.36, obtained when all samples are classified as the majority class (i.e. S1). Subjects were categorized in groups of low performance (n=216), with F1-score < 0.33 (performance as good as random), medium performance (n=249), with 0.33 < F1-score < 0.66 and high performance (n=103), with F1-score > 0.66. We compared three aspects of each group: self-reported stress imbalance, physiological dynamic range and demographics and psychological background information.

Subjects in the high performance group had on average a more imbalanced dataset (86% no stress, 12% light stress and 2% high stress), compared to the low performance group (26% no stress, 45% light stress and 29% high stress). This imbalance could provide an explanation for the difference between low and high performance.

However, we also found that for 15 out of 18 features, the high performance group has a higher dynamic range (i.e. a larger average difference per physiological feature between low and high stress) as compared to the low performance group. In Appendix B we show that this effect is significantly different as compared to dividing subjects randomly in three groups. Examples for mean HR, phasic SC and median ST, are shown in Figure 6-6A, B and C respectively; a complete summary for all features is provided in Appendix C.

To account for possible confounders we further investigated subjects' demographics and psychological information, based on the intake questionnaire, in the three groups. There was no difference in gender in all

three groups (X2 low-high performance: p=0.62, X2 low-medium performance: p=0.41, X^2 medium-high performance: p=0.92). On average subjects in the high performance group reported a healthier lifestyle and lower baseline depression, anxiety and stress levels than subjects in the low performance group (Figure 6-6D, E and F). They report to eat less take-out (low performance group: 1.1±1.3 times per week, high performance group: 0.8±0.9 times per week, kruskal-wallis p=.04), to practice more sports (low performance group: 26% does not practice sports, high performance group: 18% does not practice sports, X^2=0.01), they have higher sleep quality based on the Pittsburgh Sleep Quality Index (PSQI scores higher than 5 indicate worse sleep quality; low performance group: 5.3±2.5, high performance group: 4.1±2.3, kruskal-wallis p<.001) and score lower on depression scale (Depression Anxiety Stress Scale (DASS) – depression scale; low performance group: 3.5±3.4, high performance group: 1.4±2.1, kruskal-wallis p<.001), anxiety scale (DASS – anxiety scale; low performance group: 2.6±2.9, high performance group: 1.0±1.7, kruskal-wallis p<.001) and stress scales (DASS – stress scale; low performance group: 6.5±3.9, high performance group: 3.1±3.2; Perceived Stress Scale (PSS); low performance group: 17.1±5.6, high performance group: 10.5±5.5, kruskal-wallis p<.001) as compared to subjects in the low performance group. Subjects in the high performance group are also significantly older (low performance group: 38.6±10.0, high performance group: 41.7±10.0, kruskal-wallis p=.007).

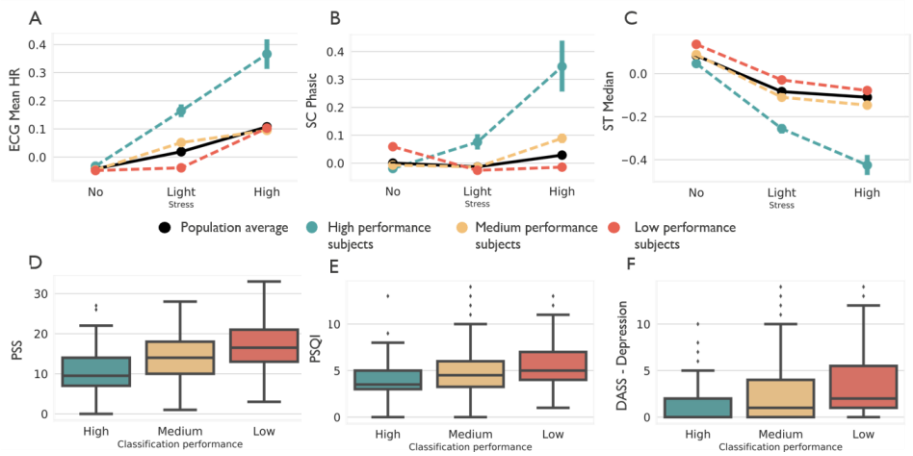


Figure 6-6: Comparison of subjects with low, medium and high classification performance.
In A, B and C average features ECG mean HR, SC phasic and ST median are shown respectively for low (red), medium (yellow) and high performance (green) groups and compared with the entire population average in phases of no, light and high stress. In D, E, and F baseline psychological information of subjects in low, medium and high performance groups are compared. The high performance group has a larger dynamic range and a lower average score on the PSS, PSQI and DASS-Depression scale than the low performance group.

6.4. Conclusion

To assess stress we collected a dataset of 1,002 subjects during five consecutive days, including a wide variety of subject background information, physiological data in ambulatory settings and smartphone-based self-reports and contextual information. We found significant differences between physiological features for ECG, SC and ST between different stress levels and nighttime baseline, confirming laboratory findings and indicating the potential of psychophysiological stress detection in daily life.
Additionally, we compared digital phenotypes based on wearable and self-reported data emerging from a data-driven analysis. We found that physiological responses to stress strongly differ among subjects, distinguishing groups with small and large dynamic ranges of the physiological features. These groups are also characterized by different psychological baselines and demographics, where the group with a more blunted physiological stress-reactivity (small dynamic range) tend to report a less healthy lifestyle and

higher depression, anxiety and stress scores than the more responsive group (large dynamic range). These findings suggest that self-reported poor health and high depression scores are negatively correlated to physiological reactivity. Similar findings have been reported previously in laboratory research [208], but to date no studies have investigated this relationship in real-life ambulatory physiological recordings.

These results provide a baseline for large-scale ambulatory population monitoring to uncover blunted physiological responses to stress and provide personalized disease interception. Furthermore, these findings have important implications related to stress modeling strategies, indicating that stress detection models should be tailored to phenotypes by including multi-sensor data sources, as subjects with different health statuses, display different physiological responses to stress. This study exemplifies how large-scale, data-driven analytics can be used to derive digital phenotypes and generate new insights into stress detection and disease interception in general. Continuous stress detection will form the basis to enable highly personalized, just-in-time interventions to enable preventive health.

Chapter 7: The MIST as calibration towards personalized stress detection

In previous chapters we have shown that the physiological stress response is person-dependent. In the current chapter we introduce a methodology to use subject-specific information, based on the physiological response to the Montreal Imaging Stress Task (MIST), to improve ambulant classification performance. The methodology is developed by Elena Smets with promotor Chris Van Hoof. We validate, for the first time, the application of the MIST in an uncontrolled, daily-life environment. Next, we show that models developed based on a personalized normalization outperform a generalized normalization, confirming findings in earlier studies and highlighting the person-dependent nature of the physiological stress response. Finally, we compared three ambulant modeling approaches in which the subject-specific information based on the MIST was used in a different way. Although none of the approaches improved classification performance significantly, we showed a relation between model performance and feature importances found during the MIST and during daily-life settings, highlighting the potential of this methodology, in which subject-specific information based on the MIST is transferred for learning ambulant models. The content of this chapter is submitted to IEEE Transactions on Affective Computing.

7.1. Problem statement

In Chapter 3 and Chapter 6 we have shown that the physiological stress response is person dependent. In Chapter 3 we developed personalized models, which were trained and tested individually per subject. In Chapter 6 we identified digital phenotypes characterized by self-reported poor health indicators and high depression, anxiety and stress scores that are associated to blunted physiological responses to stress.

Several studies have already tried to improve physiological stress detection by including personalized techniques in the data processing. A first, commonly

used, approach, is to normalize the physiological signals based on the subject's baseline recordings to remove subject-specific components [107, 53]. Second, Giakoumis *et al.* [136] suggested a subject-dependent feature calculation. An adaptive filtering is applied using "rest signatures", which are developed based on the subject's baseline physiology and represent a personalized deviation from standard rest templates. The proposed subject-dependent features reach a stress detection accuracy of 95%, which is a significant improvement as compared to the accuracy of 85 % using subject-independent features [136]. A third suggestion is to use personalized models, which are trained and tested individually per subject, as presented in Chapter 3. This technique improved classification accuracy when using Bayesian Networks.

These methods have provided significant improvement for stress detection in laboratory settings. In free-living conditions however it is not trivial to identify the baseline needed for normalization or subject-dependent feature calculations. Furthermore, in free-living conditions an objective stress reference is lacking and the stressor causing the stress response is often unknown. Therefore, not only baseline physiology is missing, but also a signature stress response to a known stressor.

To tackle these issues, we explore a methodology to identify this baseline and signature stress response through a standardized stress test on a smartphone application, which can be executed in uncontrolled settings, outside the laboratory. We investigate whether subject-specific information retrieved from such an application-based stress test can improve ambulatory stress detection.

In the SWEET study, we collected data of 1,002 subjects of which 660 subjects executed the Montreal Imaging Stress Task (MIST, see Chapter 5.2.1 Data Collection), a social stress test where subjects solve arithmetic tasks under time pressure [159]. Additionally for these subjects, we collected five days of ambulatory physiological measurements, complemented with self-reported stress levels.

To our knowledge, this is the first time such a stress test is conducted in free-living environments, a fully uncontrolled setting. Therefore, the first goal was to compare participant performance metrics with reported metrics in laboratory studies to validate the stress test. Second, we compared classification performances of three approaches towards personalized models for ambulatory physiological stress detection:

a) A **reference approach** training and testing the models based on ambulatory data
b) A truly **personalized approach** training, for each subject, a model on the MIST data and testing it on the subject's ambulatory data
c) A **hybrid approach**, clustering subjects based on the feature importance of their personal MIST model and training and testing ambulant models for each cluster separately.

We hypothesize that a more personalized approach ((b) and (c)) will outperform the reference approach.

7.2. Materials and Methods

7.2.1. Data collection

The analysis is based on data collected in the SWEET study. A detailed explanation of the protocol can be found in *Chapter 5.2.1. Data Collection*. In short, we collected five days of physiological data, combined with self-reported stress levels. On the first day of the experiment, subjects had to complete the MIST: an application-based stress test including a five minute rest period (relaxing music and images), a five minute control period (simple mathematic tasks, no time restrictions or social control), five minute stress task (mathematic tasks with time restrictions and social control) and again five minute rest period (relaxing music and images).
The MIST was initially developed by Dedovic *et al.* [159] to induce moderate stress levels in a functional imaging setting to study the effects of stress on physiology and brain activation. It is derived from the Trier Social Stress Test, and comprises multiple social threat elements included in a program with computerized arithmetic challenges. In current study the MIST was chosen because it is a validated computerized stress test, which allows us to use it in an ambulatory environment.
The MIST was built in such a way that during the stress task the subject could not give three successive correct answers, regardless of his or her ability to solve mathematic tasks. When the subject gave two correct answers, the time to give the next answer was reduced and/or the difficulty of the arithmetic was

increased [159]. When the subject gave an incorrect answer, the screen displayed 'wrong' (Figure 7-1b), when the time was up the screen displayed 'time's up' (Figure 7-1c). Additionally, a social component was added: subjects were told they could increase their chances to win a travel voucher or dinner if they scored high on the test. On the right side of the screen, subjects saw their average score (~50%) as compared to the fake average score of the other participants (~90%) (Figure 7-1a).

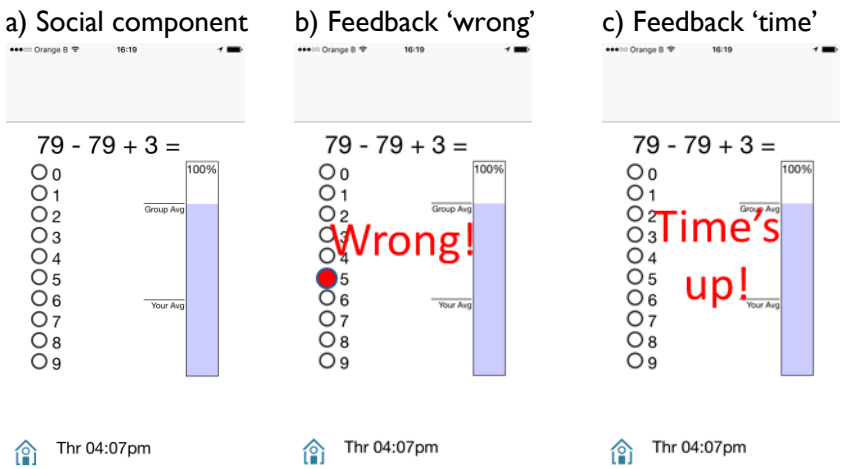


Figure 7-1: MIST stress task. Three stress components were introduced: a) a social component, b) feedback 'wrong' and c) feedback 'time's up'.

In total, 1,002 subjects participated on the SWEET study of which 342 subjects did not start the MIST. Additionally, 306 subjects did not fully complete the MIST, e.g. they only completed the rest and control part, but not the stress. These subjects were excluded from the analysis, leaving 354 subjectfs with a fully completed MIST.

For the physiological signals quality indicators and features were calculated as presented in *Chapter 5.2.2. Quality indicators* and *Chapter 6.2.1. Feature computation*. For current analysis features were calculated in a time window of 1 minute without overlap. Only high quality features (QI > 0.8) were retained for analysis. Subjects with less than two high quality data points, each representing one minute of data, in each five minute segment of the MIST (i.e. rest, control, stress and rest) were removed for analysis, resulting in 199 subjects that were included with high quality data and a fully completed MIST.

The remainder of the analysis is subdivided into two parts: MIST and ambulant data analysis.
For the MIST analysis, first, we investigated subject performance metrics, i.e. the number of correct responses on MIST and time to respond in control and stress tasks. The goal was to validate the stressful effect of the tasks, in uncontrolled conditions (i.e. outside the laboratory). Second, we used Random Forests (RFs) to train and test two physiological models for the MIST: a model based on personalized normalization and a model based on generalized normalization. The goal was to verify if a personalized normalization increases performance as is suggested in laboratory research [107, 53]. Third, we trained, for each subject, a RF model based on the MIST data and clustered subjects with similar feature importance. We hypothesize that these models and clusters can be used to improve the performance of the ambulatory stress detection.
For the ambulant analysis, first, we computed a reference classification performance by training and testing RF models based on the ambulatory data. Second, we compared the classification performance of the reference model with the classification performance when applying the models trained on the MIST data, on the ambulant data. Third, we subdivided the subjects in the same clusters as identified based on the subject's MIST model's feature importance. For each cluster the ambulant data was then used to train and test RF models for physiological stress detection. We compared the classification performance of the reference model with the classification performance when training and testing models per cluster.
A schematic overview of the analysis is shown in Figure 7-2.

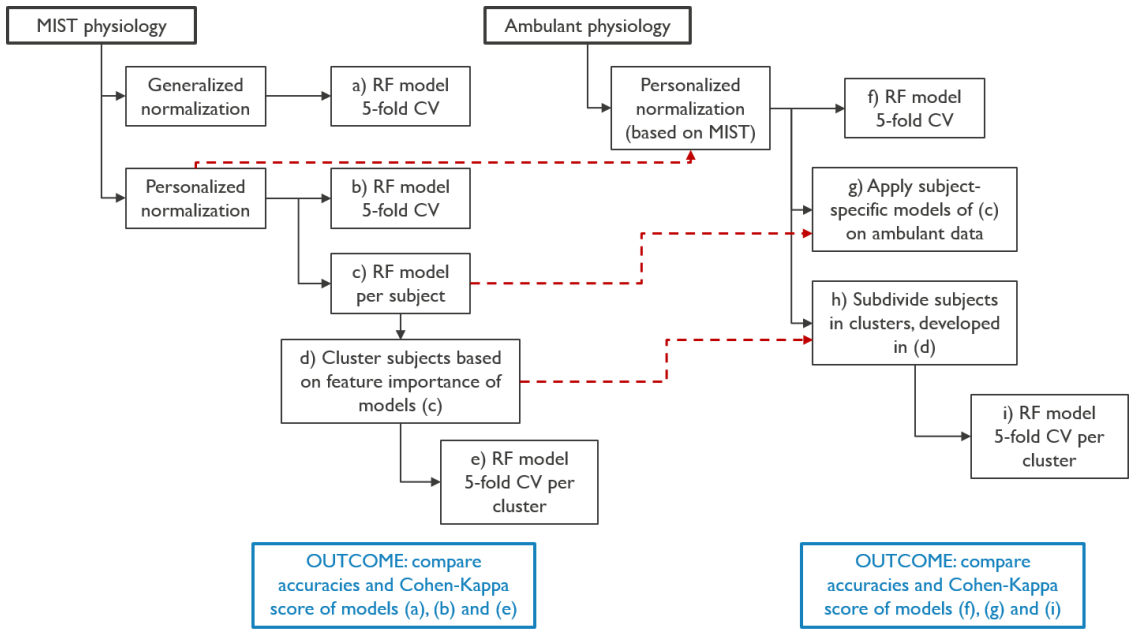


Figure 7-2: Schematic of analysis.
Four Random Forest (RF) models based on the MIST were developed (a-c, e) and one clustering based on subject's model feature importance (d). A reference RF model was developed based on the ambulant data (f), subject-specific MIST models were applied on the ambulant data (g) and subjects were subdivided in clusters based on (d) and RF models were developed per cluster (i). Three locations of Information transfer from MIST to ambulant data are indicated with red arrows. Outcomes are the comparison of model performance (accuracy and Cohen-Kappa score) between models (a), (b) and (e) and between (f), (g) and (i).

7.2.2. MIST analysis

MIST validation in free-living environments

The goal was to validate the use of the MIST in fully uncontrolled, free-living environments. Therefore, we compared performance metrics of the MIST in control and stress test, where we expect a reduced performance during the stress test as a consequence of the social threat components and time pressure [159].
Two performance metrics were measured during the control and stress task of the MIST: time to respond to the question and number of correct answers versus number of incorrect or out of time (only for stress task) answers. We performed a Wilcoxon ranksum test, with significance level at 0.05, to compare the performance metrics in control and stress task. We hypothesize the time to answer the question is reduced in the stress task and the number of incorrect answers is increased.

Model development

We considered a binary classification model in which the control and stress phase of the MIST were combined to represent activation of the ANS, here denoted as stress, the two rest phases at the start and end of the MIST represent rest.
Quality indicators and features were calculated as presented in *Chapter 5.2.2. Quality indicators* and *Chapter 6.2.1. Feature computation*, in a time window of one minute without overlap. Only high quality features (QI > 0.8) were retained for analysis.
Two normalization techniques were compared: personalized vs. generalized normalization. In the personalized approach, for each subject, the mean and standard deviation (SD) for each feature during the entire MIST were used to normalize the data (z-normalization). For the generalized approach, for each feature, the mean and SD of the entire MIST dataset, including all subjects, were used to normalize the data. Since the personalized approach removes subject-specific components (e.g. baseline HR), we hypothesize the classification performance of the models based on personalized normalization will outperform the performance based on the generalized normalization.

A feature reduction step was introduced on the entire dataset, removing correlated features (r=0.7). Then, the data was split in training and test set, using a five-fold cross-validation on the subjects. This means that 80% of subjects was considered as training and 20% as test set. This approach was repeated five times, so that each subject was exactly once in the test set. A false discovery rate supervised feature selection was applied on the training set, according to the Benjamini-Hochberg procedure (python scikit-learn, alpha=0.05). RF models of 100 trees, based on out-of-bag error, and a balanced class weight were trained (python scikit-learn, RandomForestClassifier).
As classification performance metrics the accuracy and Cohen-Kappa (CK) score were computed. The accuracy measures the percentage of correctly classified samples. The CK score helps in the interpretation of the prediction performance of a classifier when the dataset is not perfectly balanced, it is a measure of the relative improvement of the model as compared to a random classification [209].

Clustering

Next to the five-fold cross-validation, also personalized models were developed and subjects were clustered based on their feature importances (Figure 7-2, models (c) and (d)).
For the personalized model development the same approach was used as described above, using personalized normalized features. However, instead of splitting the subjects based on a five-fold cross-validation, RF models were trained based on the data of each subject separately, resulting in 199 models, one for each subject.
For each model the feature importances were computed (python scikit-learn, RandomForestClassifier. feature_importances_, higher values represent higher importance), resulting in a matrix with subjects in the rows and feature importances for all features in the columns. An example for some features of two subjects is shown in Table 7-1. The total sum of all feature importances per subject adds up to one.

Table 7-1: Feature importance matrix.

	Mean SC	SC RR	Mean ST	ST slope	...	Mean HR	ECG RMSSD
User0091	0.10	0.01	0.10	0.005	...	0.14	0.21
User0092	0.09	0.06	0.004	0.002	...	0.23	0.20

Based on this matrix an unsupervised KMeans clustering was used to cluster groups of subjects with similar feature importances, indicating that similar features are responsive to stress. The number of clusters ranged between 2-7, the silhouette score was used to identify the most optimal number of clusters. The silhouette score is calculated as the mean inter-cluster and mean intra-cluster distance for each sample [118]. The best value is 1 and the worst value is -1. Values near zero indicate overlapping clusters. Negative values indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar [118].
To evaluate if the clustering approach could improve model performance, we computed RF models using a five-fold subject cross-validation for each cluster. We compared classification performances (accuracy and CK) between clusters and with the average performance of the models developed on the entire MIST dataset (Figure 7-2 (a) and (b)).

7.2.3. Ambulant analysis

The pre-processing of the data for the ambulant analysis is similar to the MIST. Signal quality indicators and features were computed according to *Chapter 5.2.2. Quality indicators* and *Chapter 6.2.1. Feature computation*, in a time window of one minute without overlap. Only high quality features (QI > 0.8) and low activity data (ACC SD < 0.04) were retained for analysis.
As stress reference the self-reported EMAs were used, in which subjects indicated on a 5-point Likert scale their stress levels, ranging from 1 'not at all stressed' to 5 'extremely stressed'. A binary classification problem was considered of rest (including stress level 1) versus stress (including stress levels 2-5).

The data were normalized based on the subject mean and SD of the MIST features, introducing a personalized normalization based on the MIST. Correlated features were removed (r=0.7).

Three model performances (i.e. accuracy and CK) were compared: a reference model (Figure 7-2 (f)), a MIST-based personalized model (Figure 7-2 (g)) and a hybrid model based on the MIST clustering (Figure 7-2 (i)).
For the reference model a five-fold subject cross-validation was used. This model is the reference model since a classic approach is used where the model is trained and tested solely based on the ambulant data. A false discovery rate supervised feature selection was applied on the training set, according to the Benjamini-Hochberg procedure (python scikit-learn, alpha=0.05). RF models of 100 trees, defined based on out-of-bag error, and a balanced class weight were trained (python scikit-learn, RandomForestClassifier). Model performances were evaluated based on the average accuracy and CK of the five-fold cross-validation.
For the MIST-based personalized model, the models that were trained for each subject based on the MIST (Figure 7-2 (c)) were tested on the subject's ambulant data. Model performances were evaluated on the ambulant data based on the average accuracy and CK of all subjects.
For the hybrid approach, first, subjects were subdivided in clusters developed based on the MIST feature importance (Figure 7-2 (d)). For each cluster models were developed based on a five-fold cross-validation, similar to the reference model. Model performances were evaluated based on the average accuracy and CK of each cluster separately and of all clusters combined.

7.3. Results

First, we aimed to validate the MIST as a stress test in an uncontrolled, free-living environment. The response times in the stress condition were faster as compared to the control condition (mean stress = 3.77±2.54s, mean control = 6.64±7.50s; Wilcoxon ranksum $p < 0.001$; Figure 7-3).

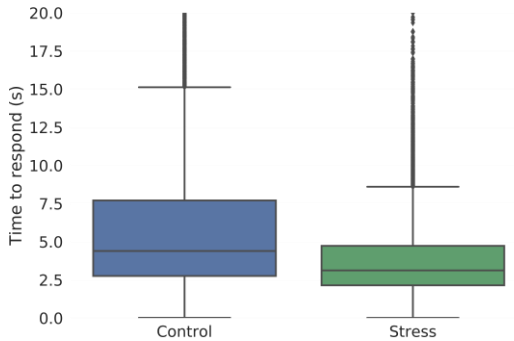


Figure 7-3: Boxplots of the response times to the MIST arithmetic tasks in control and stress condition.

Additionally, the percentage of incorrect answers was higher in the stress condition as compared to the control condition (Table 7-2).

Table 7-2: Comparison of percentage of correct, incorrect and out of time answers in control and stress condition.

	Control	Stress
Percentage correct	93.5	58.3
Percentage wrong	6.5	14.7
Percentage out of time	/	27.0

Second, we developed RF models for a binary classification, i.e. rest versus control and stress combined. We compared a personalized normalization approach versus a generalized normalization. In Figure 7-4 an example feature (mean SC) is shown when applying no normalization (a), a generalized normalization (b) and a personalized normalization (c). It can be seen that when applying no or a generalized normalization, many outliers are present, due to the subject-specific baseline physiological profiles, i.e. some subjects naturally sweat more than others. Differences between rest and stress become more pronounced when applying a personalized normalization as compared to a generalized normalization (wilcoxon ranksum mean SC rest vs. stress; generalized: p = 0.49, personalized: p = 0.002).

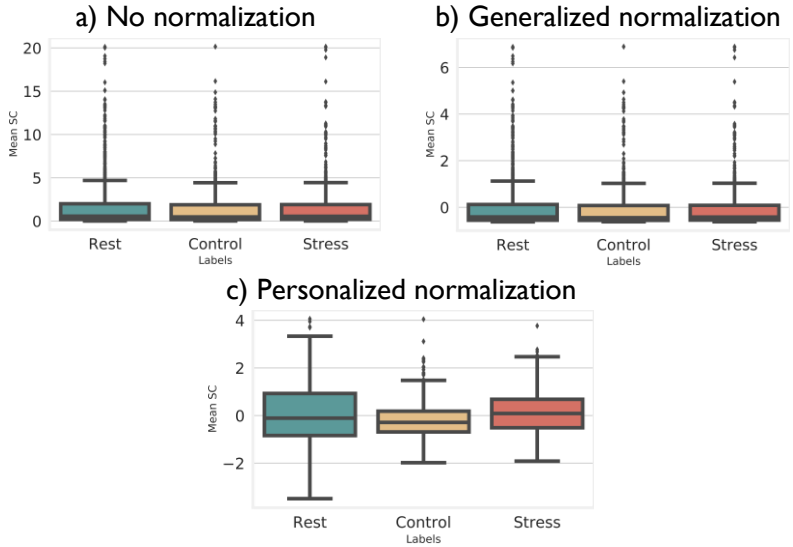


Figure 7-4: Comparison of mean SC in MIST rest, control and stress conditions when applying a) no normalization, b) a generalized normalization and c) a personalized normalization.

As a result, the average classification performance of the RF models, validated through a five-fold cross-validation, is better for models based on a personalized normalization approach as compared to a generalized normalization (Table 7-3). Therefore, for the following analyses a personalized normalization approach was used.

Table 7-3: Classification performance of RF models based on a generalized and a personalized normalization

	Accuracy	Cohen-Kappa
Generalized normalization	0.58	0.16
Personalized normalization	0.70	0.40

Third, we computed personalized models, which were trained on the entire MIST of each subject separately, resulting in 199 models, one for each subject. The feature importances of these models were then investigated. In Figure 7-5 the boxplots of the feature importances for all subject-specific models are shown. It can be seen that on average mean SC has the highest importance. Further, the large standard deviations (whiskers of the boxplots in Figure 7-5),

indicate that each feature has a different importance depending on the subject. This implies that per subject different features are responsive to stress.

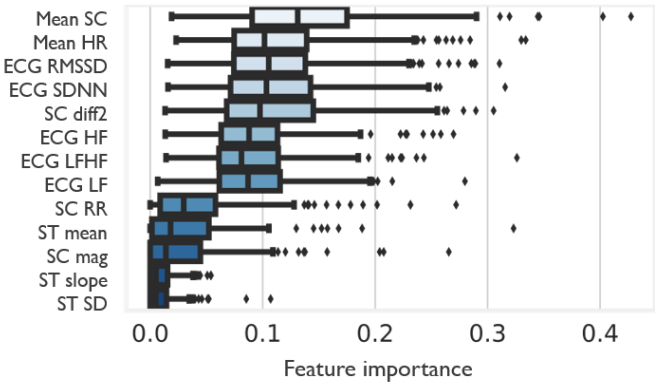


Figure 7-5: Boxplots of the feature importances of the subject-specific models.

Fourth, we computed clusters based on the feature importances of the subject-specific models. The goal was to cluster subjects for who similar features are important, implying they have a similar response to stress. We found an optimum of 2 clusters, with a silhouette score of 0.18. In Figure 7-6 the separation of the two clusters are visualized based on the first two principal components of the feature importances matrix. The centroids of each feature importance in the two clusters (A and B) are represented in a heatmap in Figure 7-7. Cluster A represents 152 subjects, cluster B represents 47 subjects. It can be seen that for cluster A mainly SC related features are important (i.e. Mean SC and SC diff2), for cluster B mainly ECG related features are important (i.e. Mean HR and ECG RMSSD).

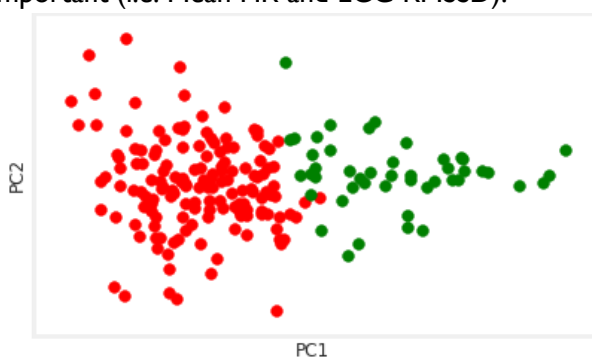


Figure 7-6: Visualization of the feature importance clusters based on the first and second principal components (PC).
Cluster A (red) and cluster B (green).

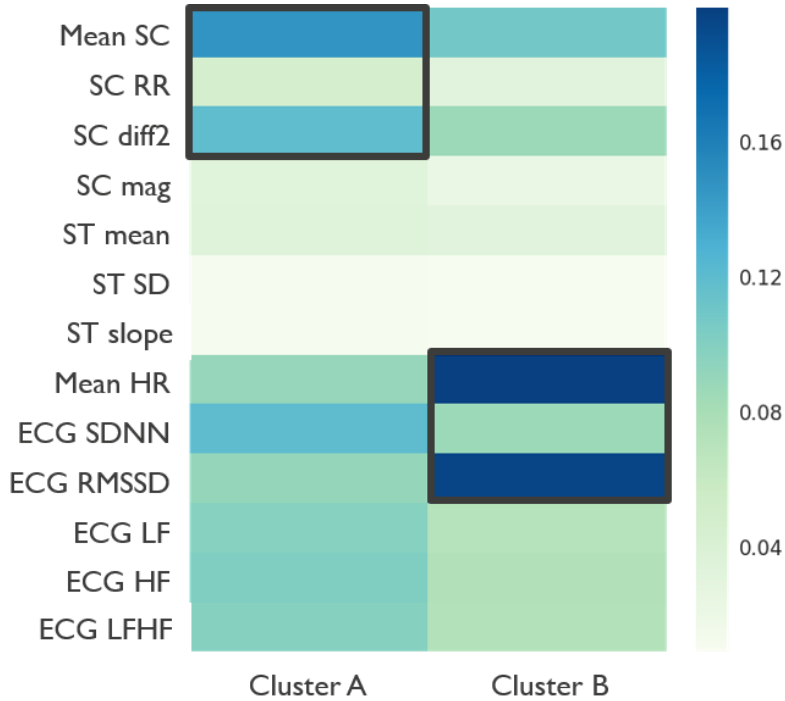


Figure 7-7: Heatmap of the feature importances in cluster A and B. Black squares indicate the most important features per cluster.

To investigate if the split in clusters improves classification performance, we trained and tested RF models using a five-fold cross-validation in each cluster separately and compared the average performance with that of the models developed on the entire dataset using personalized normalization (see Table 7-3). In Table 7-4 it can be seen that the average performance of cluster A and B, weighted with the number of subjects in each cluster, does not differ much from the average performance when training and testing on the entire MIST dataset (accuracy = 0.7, CK = 0.4, as presented in Table 7-3). However, a higher performance is observed for cluster B as compared to cluster A.

Table 7-4: Classification performance of RF models trained and tested in each cluster separately.

	Accuracy	Cohen-Kappa
Cluster A (SC-related features) (n=152)	0.66	0.33
Cluster B (ECG-related features) (n=47)	0.80	0.60
Weighted average (n=199)	0.69	0.38

Finally, the aim of this study was to use the MIST to increase personalization on the ambulant stress detection. Therefore, the classification performances of three models on the ambulant data were compared: a *reference model*, using a five-fold cross-validation to train and test a model only using the ambulant data, a *personalized model*, using the subject-specific model that was trained on the MIST and testing it on the ambulant data of that subject, and a *hybrid model* in which first subjects were split according to the clusters that were developed based on the subject's MIST feature importances, and second, for each cluster, RF models were trained and tested only using the ambulant data.
The results are presented in Table 7-5. It can be seen that the personalized model has a lower performance as compared to the reference model. The weighted average performance of the hybrid model does not differ from the reference model. However, as for the MIST, it can be seen that the classification performance of the models trained and tested in cluster B is on average higher than the performance of the models trained and tested in cluster A.

Table 7-5: Classification performance of RF models on ambulant data.

	Accuracy	Cohen-Kappa
Reference model	0.54	0.07
Personalized model	0.50	0.01
Hybrid model weighted average	0.54	0.07
- Cluster A (n=152)	0.53	0.06
- Cluster B (n=47)	0.55	0.11

7.4. Discussion

We investigated whether subject-specific information retrieved from an application-based stress test, which was executed in an uncontrolled, daily-life environment, could improve ambulatory stress detection.

Since, to our knowledge, this is the first time the MIST, or any related stress test, has been conducted in a daily-life environment, we first validated the test based on subjects' performance metrics. We showed a significant reduction in response time for the stress phase as compared to the control phase. Additionally, we reported an average of 58.3% correct answers in the stress phase as compared to 93.5% in the control phase. These findings suggest the stress test did have a significant stressful effect on the subjects. We compared these results with findings in laboratory studies. The study of Dedovic *et al.* [159] in which the MIST was presented for the first time, reported an average rate of incorrect answers during the stress phase of 20-45%, which is even lower than the rate reported in our study. In later studies failure rates of around 55% during the stress phase have been reported [210], which are similar to our findings. We can thus conclude that the stressful effect is similar in controlled and uncontrolled conditions. However, we did have to remove a significant percentage of subjects (65%) from the analysis due to no (34%) or incomplete (31%) execution of the MIST, whereas in a controlled, laboratory setting the experimenter can ensure a correct execution of the test. This is a pitfall of a stress test executed in uncontrolled, daily-life settings. In future studies additional incentives could be used to motivate the subject to execute the stress test. Alternatively, the application could be developed in such a way that it can only be used for ambulant stress detection after executing the stress test, forcing the subject to execute the stress test before receiving feedback.

Second, we confirmed the findings of laboratory studies [107, 53], stating that a personalized normalization improves classification performance. We showed that stress detection accuracy improved from 0.58 to 0.7 and CK improved from 0.16 to 0.40 when using a personalized normalization as opposed to a generalized normalization. Therefore we suggest future research focuses on stress detection models, including a personalized normalization. However, an important risk of this approach is that, per subject, not only the signal of interest is increased, but also the noise levels. For personalized normalization, we used the subject's mean and SD to standardize each feature. If the signal

was very noisy, the SD will mainly reflect noise instead of the signal of interest. A personalized normalization will in this case increase the importance of the noise component. Therefore, it is highly important to first apply quality indicators or a noise filtering technique, before applying a personalized normalization.

A limitation of the MIST modeling is the way we split rest and stress states. In our analysis we used a binary classification, i.e. rest versus control and stress combined. It could be argued that a three-class classification of rest versus control versus stress would be more representative. Alternatively the control task could also be combined with the rest tasks instead of the stress task. We chose to combine the control task with the stress task, since they both represent an activation of the ANS. Future research should investigate the results based on the other two approaches.

Third, we trained physiological stress detection models based on the MIST data for each subject separately and clustered subjects in two groups according to their feature importances. We found cluster A (152 subjects) for which mainly SC-related features were important and cluster B (47 subjects) for which mainly ECG-related features were important. We trained and tested stress detection models for each cluster and found that the average performance of cluster B outperformed cluster A. This could imply that ECG-related features are more suitable for stress detection, or that these subjects display a more pronounced stress response. Few studies have compared classification performances of different sensing modalities. Zhai and Barreto [97] compared models based on SC, BVP, ST and pupil dilation in a laboratory setting. They found pupil dilation the most important feature and did not see large differences between SC, BVP or ST related features. Sun et al. [105] compared ECG and SC sensing modalities for activity-aware stress detection and found that ECG-related features were more susceptible to physical activity and decreased classification performance [105]. Finally, Mozos *et al.* [122] showed in a laboratory experiment that in 16% of AdaBoost classifiers, SC is one the five most important features, as compared to 9% for BVP-related features. These studies report diverse findings and more research towards physiological sensing priorities is needed.

Further, we compared the weighted average performance of these two clusters with the performance of the models developed based on the entire data pool (i.e. without clustering). We expected that, by training and testing

models based on more similar data (i.e. clusters with similar feature importances), the classification performance would improve. However, the average performance did not change. A possible explanation could be that the clusters were not well-defined, as the silhouette score is 0.18, indicating only a moderate separation. This can also be seen in Figure 7-6 in which there are many overlapping points between cluster A (green) and cluster B (red). In this case, subjects in the same cluster, do not necessarily all have more similar feature importances and the training of physiological stress detection models is not improved. An alternative could be to use a different clustering technique such as PAM (partitioning around medoids), Fuzzy c-means, etc. [211], as opposed to the K-means clustering technique, which tends to have a high susceptibility for local optima [211].

Finally, we used the MIST to identify person-specific information to improve classification performance for ambulant stress detection models. We used the subject's mean and SD per feature of the MIST to normalize the ambulant data. Next, we compared the performance of a reference model with the performances of a personalized and hybrid model. We found the lowest performance for the personalized model. This could indicate that the magnitude of the stress response during a standardized stress test is not equal to daily-life settings. In that case the model for stress detection based on the MIST will over- or underestimate the physiological stress response during daily-life settings. Further, we found the weighted average performance of the hybrid model did not increase as compared to the reference model. However, as for the MIST, cluster B, with high MIST feature importances for ECG-related features, outperformed cluster A, with high MIST feature importances for SC-related features (especially the CK metric almost doubled). These findings imply that subjects with high classification performance during a short stress test, also have higher classification performance in daily-life settings. These findings could also imply that feature importance during a stress test and during daily-life settings are related, meaning that for the same subject, similar features are responsive to stress during the MIST and in ambulatory settings. So, although the results based on the fully personalized approach have shown that the magnitude of the stress response during the MIST does not correlate with daily-life settings, the hybrid approach has shown that in terms of feature importances, the MIST could still provide valuable information towards personalized physiological stress detection in daily-life settings. Concretely,

the results suggest the MIST could be used as 'calibration' to a) identify the subject's mean and SD for personalized normalization and b) to identify the cluster to which the subject belongs for model training and testing. Depending on the cluster to which a subject belongs we could already make predictions on whether the model will perform good or bad. Additional research is needed to leverage this information towards improving model performance.
In general, we found relatively low classification performances for all three models in the ambulant analysis. Several modifications in the analysis pipeline should be investigated to improve model performance. First, other modeling techniques besides RF should be investigated, such as ANNs which have shown to be powerful classification techniques for stress detection [63]. Second, in current analysis, per stress level annotation, we labelled the entire previous hour with the same self-reported stress level. The reason is that the question we asked in the application was 'What was your maximum stress level during the last hour?'. However, during this hour still many variations in physiology are possible, we should therefore investigate a technique to identify the point of interest in the previous hour, or for example take the mean of the physiological signals during that hour.

7.5. Conclusion

Our study investigated a methodology in which an application-based stress test (i.e. MIST) was used as 'calibration' towards person-specific model development in daily-life settings. We compared a reference, personalized and hybrid approach to use personal information retrieved from the MIST to improve ambulant classification performance. Although none of the approaches improved the classification performance, we showed a relation between model performance and feature importances found during the MIST and during daily-life settings, highlighting the potential of this methodology, in which person-specific information based on the MIST is transferred for learning ambulant models. Further research should refine the clusters which are currently developed solely based on the MIST physiology, by adding information on demographics and context, by including the subject's digital phenotype.

Chapter 8: Conclusions and future prospects

In the 21st century, stress and mental health have become major concerns worldwide. Yet, a continuous, quantitative measurement technique, allowing just-in-time interventions to reduce stress, is lacking.
Therefore, research has focused on exploiting the sympathetic nervous system's (SNS) fight-or-flight response, by investigating physiological signals for monitoring stress. These have shown to be reliable indicators of stress in well-controlled laboratory conditions, but large-scale ambulatory validation is missing.
The goal of this research was to identify physiological sensing priorities and machine learning techniques for physiological stress detection and next to deploy these on a large population in real-life, ambulatory conditions.

Three main objectives have been identified:

a) *The identification of the most suitable **markers for physiological stress detection** in a controlled laboratory environment on healthy subjects, to translate this knowledge to the ambulant environment.*
b) *The **differentiation between healthy subjects and patients** based on their physiological stress response, towards disease prevention and interception.*
c) *The **large-scale** investigation of the physiological stress response in **ambulatory conditions**, including demographics and context information towards digital phenotypes for personalized and continuous stress detection.*

Below, we discuss the conclusion for each objective separately.

8.1. Markers for physiological stress detection

We investigated both physiological sensing priorities and machine learning techniques for stress detection.

8.1.1. Physiological sensing priorities

In the literature study of Chapter 2, we identified several physiological signals that are responsive to stress, including HR, HRV, SC, EMG, BP, ST and pupil dilation. However, findings in literature related to sensing priorities are not consistent. The main issue is the difference in experimental design, sensor quality and analysis methods among different studies.

In Chapter 3, we developed generalized and personalized physiological stress detection models in controlled laboratory conditions on healthy subjects. For the generalized models, SC and HR related features have shown to be more important than ST and respiration related features. This was confirmed for the personalized models. However, further investigation of the personalized models showed that feature importance is subject-dependent and one should be careful drawing general conclusions related to sensing priorities. For the personalized models, there were only 5 features (2 HRV and 3 respiration related features) that were never selected, meaning that for none of the subjects these features were important. The other 17 remaining features, including ST and respiration related features, were therefore important for at least one subject, meaning sensing priorities are subject-dependent, and model performance of some subjects could drop if for example ST related features were to be removed from the dataset. So, overall SC and HR related features are most important, but for some subjects also ST and respiration related features should be considered.

In Chapter 4, we developed a model to classify healthy subjects and patients based on their physiological stress response to a stress task. We found an equally high classification performance based on the response related features, including all physiological signals, and the SC related features only. This confirms again the high importance of the SC signal. Surprisingly, when we investigated the feature importances of the response model further, we noticed that four out of five most important features were ST related, indicating the relevance of ST features, next to SC.

In Chapter 6, we showed that, in daily-life conditions, all three physiological signals, HR(V), SC and ST, show significant differences between self-reported stress levels. These results confirm on a large scale the potential of all three signals for physiological stress detection in daily life.

In Chapter 7, we identified clusters of feature importances based on personalized models for stress detection during the MIST. We found two clusters, one with highest importances for HR(V) related features and one with highest importances for SC related features. We found the highest classification performances, both during MIST and in ambulant data, for the cluster with SC related features.

Overall, these findings indicate that SC would be the most important physiological signal for physiological stress detection, followed by HR(V), and lastly ST. However, further investigation of personalized models, both in controlled and ambulatory conditions, has shown that **physiological sensing priorities differ across subjects**. Therefore, in future research, a multi-sensor approach is suggested.

Above findings have important implications for hardware development. Although all three investigated physiological signals, SC, HR(V) and ST, are readily available in research focused wearables such as Empatica E4 and the imec Chillband+, these are not (yet) incorporated in widespread commercial wearables such as the Fitbit, Apple watch or Samsung Gear, which only feature HR. Although our research clearly points to the high importance of SC signals, further validation towards a multi-sensor approach is needed to convince large manufacturers to incorporate all three signals in their wearables. To this end, a first step is to investigate model performances for ambulant stress detection (based on the SWEET study) for each physiological signal separately to calculate the relative gain of using the combination of all signals as opposed to only using HR (which is currently available in most wearables). Additionally, current research only investigated HR(V), SC and ST since these signals are most easy to measure in ambulant conditions, we suggest to focus additional research on adding other physiological signals such as EMG and BP, since these have proven to be relevant markers for physiological stress detection in laboratory settings. Recent advances in ambulant BP measurement based on the combination of ECG and photoplethysmography (PPG) measurements through pulse transit time analysis, should allow ubiquitous BP measurements in the future [212]. Finally, to be truly unobtrusive and widespread adopted, physiological sensing should move beyond wearables, towards non-contact sensing. Recent research has shown the potential of contactless capacitively-

coupled ECG, which allows measurement of HR and HRV through clothing by incorporating the sensors in e.g. mattresses, car seats or office chairs [213].

8.1.2. Machine learning techniques

In the literature study of Chapter 2, we discussed the most widely used machine learning techniques for physiological stress detection, including logistic regression (LR), support vector machines (SVMs), decision trees (DTs), random forests (RFs), artificial neural networks (ANNs) and Bayesian networks (BNs). Findings in literature related to machine learning technique priorities are not consistent, due to differences in experimental design, sensor quality and analysis methods among different studies.
In Chapter 3, we compared the classification performance of six techniques, including LR, SVM, DT, RF, static BN, and dynamic BN. We compared performances of generalized and personalized models for each of these six techniques. We found that for generalized models SVMs perform best and for personalized models dynamic BNs perform best. We concluded that the choice of model depends on the context of the application and whether insight in the model structure is needed.
In Chapter 4, we used a LR model to differentiate healthy subjects and patients based on their physiological response to a stress task. This model allowed insights in the feature importances and physiological sensing priorities.
In Chapter 6 and 7 we used RF models which gave the best result for physiological stress detection in an ambulant environment. We identified digital phenotypes which correspond to different model performances.

Overall, our research results have not identified one dominant machine learning technique for physiological stress detection. In our view, the **selection of a technique is strongly dependent on the context of the application**. For example, when the goal is develop a model to gain insight into a subject's stress response, white-box models such as LR and BNs are preferred. However, when the goal is to develop a fast algorithm for real-time stress detection with high accuracy, black-box models such as SVM could be a better choice.
The results of Chapter 3, 6 and 7 do consistently show that the physiological stress response tends to be subject-dependent. We have shown that

personalized models can outperform generalized models and that digital phenotypes correspond to different model performances. We have identified clusters of subjects with similar feature importances, which show different model performances both during MIST and in ambulant data. These findings call for a more **personalized approach in stress detection modeling techniques**, which to date have mainly focused on generalized methods. A possible approach could be to use population models, such as mixed effects models, which are extensions of linear regression models for data that are collected and summarized in groups [214], in this case digital phenotypes. We also suggest to investigate ANNs, including physiological, demographical and contextual information, as ANNs can capture more complex relationships among these variables.

Finally, we suggest further research towards temporal models. In Chapter 3, we have demonstrated the potential of dynamic BNs for stress detection. Previous research has confirmed the advantages of temporal models in the context of stress detection, since a current stress state tends to be strongly dependent on the previous state [107] [215].

8.2. The differentiation between healthy subjects and patients

In Chapter 4, we have shown the potential of the physiological stress response to a controlled stress task, to differentiate healthy subjects and patients with stress-related complaints, i.e. the first stage on the stress continuum before overstrain and burnout. Although this was only a preliminary study with limited number of subjects, it showed that, using ubiquitous sensing modalities, early-stage stress-related complaints could be detected.

In Chapter 6, we have used a data-driven approach to identify digital phenotypes characterized by self-reported poor health indicators and high depression, anxiety and stress scores that are associated to blunted physiological responses to stress. These results confirm the differentiation between healthy subjects and persons further along the stress continuum based on their physiological stress response.

These results are encouraging to develop a new approach **towards disease prevention and interception**, i.e. detecting the disease before there are any symptoms. This involves a paradigm shift from a 'diagnose and treat' approach to a 'predict and pre-empt' (i.e. intervening early in the disease) model [216]. In many mental health diseases early intervention is key, yet tools for disease interception are lacking. Our results show the potential of using physiological signals for the interception of stress-related diseases (e.g. burnout). Further research is needed for population-wide long-term monitoring, i.e. more than six months, of healthy subjects until disease onset, i.e. until they develop a burnout. This will provide insight into the exact physiological dynamics during disease progression, which could allow to define a cut-off point at which the subject is 'in danger' for developing a burnout and should find appropriate treatment (i.e. disease interception). Such studies are now made possible using ubiquitous sensing modalities such as wearables and smartphones.

Additionally, these insights could be used **towards relapse prevention**, a major problem in mental health diseases. For example, more than 50% of depression patients have a relapse after their first episode and approximately 80% have a relapse after their second episode [217]. Early detection of stress-induced relapse episodes, using continuous monitoring systems, could provide early warning signs and early intervention of clinicians.

8.3. Towards physiological stress detection in an ambulant environment

In Chapter 3, we have presented a binary physiological stress detection model in a controlled environment, with the highest accuracy for personalized dynamic BNs of 84.6%, which is in line with results obtained in literature [102] [103].

In Chapter 5, we have presented the SWEET study. The is the world's largest ambulatory stress detection study, including 1,002 subjects who were continuously monitored during 5 days. We presented a protocol including physiological sensing, baseline psychological information, self-reported stress and contextual sensing based on smartphone information. Initial results have

revealed important insights related to user compliance, user privacy, data quality and the need for user sensor comfort.
In Chapter 6, we developed a three-class RF classification model for ambulatory physiological stress detection, reaching an F1-score of 0.43, which is slightly better than a random classification.
In Chapter 7, we developed a binary RF classification model for stress detection during the MIST, reaching an average F1-score of 0.7, which is slightly lower, but still comparable with, findings in laboratory settings [102] [103]. Further, we developed a binary classification model for ambulatory physiological stress detection, reaching an average F1-score of 0.54, which is slightly better than a random classification.
Both in Chapter 6 and 7, we identified clusters of digital phenotypes, based on physiological and/or psychological baseline information which are characterized by low or high classification performances.

Overall, these findings indicate that **stress can accurately be detected in controlled environments**. However, **stress detection in daily life conditions remains challenging**. A main difference between controlled and ambulatory settings, lies in the gold standard for stress. In a controlled environment the gold standard is based on the timestamps at which tasks are executed (i.e. when a stressful tasks is executed, the gold standard is labeled as 'stress', when a relaxing task is executed it is labeled as 'rest'). However, in a daily-living setting the gold standard is based on self-reports. This could introduce bias in the gold standard and could lead to lower classification accuracies. This also raises the question whether we are in fact detecting stress or rather an activation of the ANS. From a data analytical perspective, one could argue we detect stress since the models are trained based on a, self-reported, stress reference. From a psychophysiological perspective, it is not clear whether our physiological sensing models can differentiate between actual stress, which we have defined as the combination of high demands and low decision latitude [26], and arousal or an ANS activation. In the analysis of the MIST we merged control and stress phases as one ANS activation phase, mainly because physiological differences between control and stress were not very pronounced. This could point to the fact that what we actually measure is ANS activation rather than stress as such. To test this hypothesis the self-assessment manikin (SAM) of the SWEET study, which measured pleasure,

arousal and control levels, could be investigated in combination with self-reported stress levels and physiological responses.
Our ambulant stress detection classification performance results are on average lower than those reported in literature. However, it is difficult to compare results of different studies. For example, Hovsepian *et al*. found a classification accuracy of 72% of their binary physiological stress detection model in ambulatory settings, but their approach differs strongly from ours. They included self-reported stress levels of a previous time instant as a feature to predict the current stress level. Although such a temporal model is an interesting approach, one could argue that incorporating self-reported information in the model's prediction does not allow continuous, unobtrusive stress detection.
We suggest future research to focus on personalized and temporal models to improve classification performance.

8.4. Future work on the rich data content of the SWEET study

As the opportunities for data analyses based on the SWEET study are immense, not all possible investigations or insights fit within the scope of this thesis. We aimed to focus on personalized physiological stress detection in a daily-life setting. Here, we propose a non-exhaustive list of **suggestions for future research based on the SWEET study data**. First, it could be investigated how context information (e.g. location, audio, physical activity), combined with physiological data, could be used to improve the performance of stress detection models. A first analysis of stress detection, solely based on context information showed a user lift of 4.98%, meaning the model could predict stress 4.98% more accurately than always predicting the majority class [218]. Further, a comparative study among the classification performances of each physiological signal separately (i.e. HR/HRV, SC and ST) would be useful to provide more insight into which physiological features are indispensable for ambulatory stress detection. This could be valuable information for hardware developers related to the decision on which features to include in their wearables. Additionally, subjects reported their sleep quality throughout the experiment. The link between physiology and sleep quality for different

psychological profiles (e.g. high versus low depression) could be investigated. Further, subjects reported their gastro-intestinal symptoms using the Leuven Postprandial Distress Scale (LPDS). The link between physiological signals, stress, and gastro-intestinal complaints could be investigate. Last but not least, the SWEET study could be used as a healthy baseline to compare physiological responses of healthy subjects with those of patient populations and analyze the effectiveness of interventions on both healthy and patient populations.

8.5. Lessons learned towards large-scale, ambulatory data collection

The data collection throughout the three trials conducted in this thesis, have indicated several lessons learned towards ambulatory physiological data collection and analysis, which could be of interest far beyond the application domain of stress.

First, the **subject selection** procedure is crucial to collect useful data. In the SWEET study all employees who were fit to work could participate in the study. Advertisement of the study was conducted through the communication department of each company. We noticed that the buy-in of the responsible for prevention and safety at work and the HR department were key to reach a higher number of subjects. The advantage of having broad inclusion criteria, and only limited exclusion criteria, is that it allows a dataset with a high variety in population. However, it does not allow any control over the distribution of the dataset (e.g. the distribution of age, gender, etc.) and it is susceptible for self-selection bias. For example, it is possible that subjects with higher stress levels will less likely participate in the study because they feel they do not have the time for it. In the future, it would be useful to work with user panels in which selection of a more evenly distributed population is more feasible.

Further, **subject motivation** is key towards compliance and to reduce drop-outs. In the SWEET study, subjects did not receive any reward for participating nor any feedback on their results. The advantage of not giving a reward is that mainly subjects with a strong intrinsic motivation are attracted to participate, whereas providing money could motivate subjects to participate for the wrong reasons, eventually reducing motivation and resulting in poor data quality (not wearing the sensors) and low numbers of answered EMAs. However, we have

seen that compliance to the EMAs was 42%, which is less than commonly reported rates between 65-85% [174]. In the future, and especially when conducting longitudinal trials more effort is needed to keep subjects motivated on the long term. Providing subjects with meaningful feedback could be one solution. In addition subjects could be given a reward based on their compliance (e.g. based on the number of answered EMAs) [219]. Another solution could be to use context-aware EMAs, which use contextual information to identify when to send an EMA trigger instead of using a random time schedule. This could be less disturbing for the study participants, as the algorithm should identify whether the subject is capable to answer the EMA, and could further improve the quality of the dataset as specific points of interests could be questioned more frequently (e.g. moments of high stress).

We have used several **devices** throughout this thesis, both commercial (i.e. NeXus 10 – MK II) and noncommercial (i.e. imec Necklace, chest patch and Chillband). In the first two laboratory trials the main focus was on signal quality and bulky devices such as the NeXus 10 – MK II could be used. In ambulatory trials however, apart from the signal quality also the user comfort is key. Therefore, we used the chest patch and Chillband in the SWEET study which are smaller, more user friendly and have a longer battery lifetime (i.e. 7 days continuous monitoring). However, still the majority of the subjects in the study indicated they would not wear the sensors on a daily basis in the future because they are too big or not comfortable. This calls for action of hardware developers, who should put more focus on the aspect of user comfort, already in an early stage of development. Further analysis should indicate which physiological signals are indispensable for physiological stress detection and only those should be incorporated in newly developed hardware. In the future, non-contact sensing techniques could be an alternative for the use of wearables, allowing truly unobtrusive sensing.

Finally, the quality of the data collection could significantly be improved by implementing (near to) real-time, accurate and fine-grained **quality indicators**. In the SWEET study the quality assessment was performed offline, after the data collection. In the future, it would be useful to have (near to) real-time data collection, combined with (near to) real-time quality indicators. This could raise alarms for the investigator, allowing early intervention, when quality drops consistently, which could indicate the subject has a broken sensor or is using the sensor incorrectly.

8.6. A final note on precision medicine and personalized healthcare

To conclude, our findings provide a first step towards personalized stress detection, and more generally they are **a first step towards precision medicine and personalized healthcare**. By including context information, we cannot only improve modelling performance, but also enhance feedback strategies towards personalized care. Such tailored interventions, also called just-in-time adaptive interventions (JITAIs), use real-time data about subjects' health to deliver real-time interventions adapted to the subjects' specific needs [220]. The goal of JITAIs is to induce behavior change based on triggering personalized, appropriate support in real-life settings at the right time and the right place [221]. JITAIs have proximal and distal outcomes, where proximal outcomes are the short-term goals the intervention wants to achieve and which can be measured shortly after the intervention is presented (e.g. daily stress and anxiety levels), and distal outcomes are the ultimate goal of the interventions, usually the primarily clinical outcome (e.g. relapse prevention in schizophrenia) [222]. Trends of the proximal outcomes can be measured on a relatively short timeframe (i.e. hours, days, weeks), whereas to monitor trends in the distal outcomes longitudinal trials of months, even years are needed.
JITAIs are fairly new concepts and so far have not been tested in many disciplines. Most examples are related to diet (e.g. reduce energy intake by just-in-time audio cues [220]) or smoking cessation (e.g. a personalized, context-driven smartphone application for smoking cessation [223]). Another study investigated JITAIs for stress reduction, by providing subjects stress-management skills at times they reported high stress rather than randomly throughout the day or week [224]. Although this method still relied on self-reported stress, rather than on physiological, continuous stress detection, the results showed a significantly improved stress reduction outcome [224]. We suggest future research to investigate if JITAIs for stress reduction could be improved, by linking stress reduction interventions to contextual information (e.g. location, calendar, social interaction, weather, etc.). This way, ubiquitous sensors, computing and feedback could aid in precision medicine towards early detection and prevention of stress-related diseases and cause a paradigm shift from treatment to disease prevention and interception.

References

[1] L. Tan, M.-J. Wang, M. Modini, S. Joyce, A. Mykletun, H. Christensen and S. Harvey, "Preventing the development of depression at work: a systematic review and meta-analysis of universal interventions in the workplace," *BMC Medicine,* 2014.

[2] P. A. American, "Stress in America: The impact of discrimination. "Stress in America" Survey," 2016.

[3] EU-OSHA, "European Opinion Poll on Occupational Safety and Health," Publications Office of the European Union, Luxembourg, 2013.

[4] A. Bakker and E. V. W. Demerouti, "Using the job demands-resources model to predict burnout and performance," *Human Resource Management,* vol. 43, pp. 83-104, 2004.

[5] L. Hourani, T. Williams and A. Kress, "Stress, mental health, and job performance among active duty military personnel: Findings from the 2002 department of defense health-related behaviors study," *Military medicine,* vol. 171, pp. 849-856, 2006.

[6] S. Pflanz and A. Ogle, "Job stress, depression, work performance, and perception of supervisors in military personnel," *Military Medicine,* vol. 171, pp. 861-865, 2006.

[7] O.-L. Siu, "Job stress and job performance among employees in Hong Kong: The role of Chinese work values and organizational commitment," *International journal of psychology,* vol. 38, pp. 337-347, 2003.

[8] M. Kivimäki, M. Virtanen, M. Elovainio, A. Kouvonen, A. Väänänen and J. Vahtera, "Work stress in the etiology of coronary heart disease - a meta-analysis," *Scand J Work Environ Health,* vol. 32, no. 6, pp. 431-442, 2006.

[9] M. Milczarek, E. Schneider and E. González, "OSH in figures: Stress at work - facts and figures," Office for Official Publications of the European Communities, Luxembourg, 2009.

[10] EU-OSHA, "Calculation the costs of work-related stress and psychosocial risks - A literature review," Publications Office of the European Union, Luxembourg, 2014.

[11] NationaleArbeidsraad, "Voorkoming van stress - Collectieve Arbeidsovereenkomst nr 72," Brussel, 2004.

[12] "Council directive on the introduction of measures to encourage improvements in the safety and health of workers at work," 1989.

[13] E. Lee, "Review of the Psychosometric Evidence of the Perceived Stress Scale," *Asian Nursing Research,* vol. 6, no. 4, pp. 121-127, 2012.

[14] A. Alberdi, A. Aztiria and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of Biomedical Informatics,* vol. 59, pp. 49-75, 2016.

[15] W. Lovallo, Stress & Health: biological and psychological interactions, Los Angeles: SAGE, 2016.

[16] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Computer methods and programs in biomedicine,* vol. 108, pp. 1287-1301, 2012.

[17] A. Muaremi, B. Arnrich and G. Tröster, "Towards Measuring Stress with Smartphones and Wearable Devices During Workday and Sleep," *BioNanoScience,* vol. 3, no. 2, pp. 172-183, 2013.

[18] H. Selye, "A Syndrome produced by Diverse Nocuous Agents," *Nature,* vol. 138, p. 32, 1936.

[19] S. Szabo, Y. Tache and A. Somogyi, "The legacy of Hans Selye and the origins of stress research: A retrospective 75 years after his landmark brief "Letter" to the Editor of Nature," *Stress,* vol. 15, no. 5, pp. 472-478, 2012.

[20] M. Le Fevre, J. Matheny and G. Kolt, "Eustress, distress, and interpretation in occupational stress," *Journal of Managerial Psychology,* vol. 18, no. 7, pp. 726-744, 2003.

[21] K. Mcgregor, S. Maeght and R. T., "Understanding burnout and reducing its impact," in *The Sports Monograph: Critical perspectives on socio-cultural sport, coaching and Physical Education*, Preston, SSTO Publications, 2014, pp. 323-336.

[22] R. Yerkes and J. Dodson, "The relation of strength of stimulus to rapidity of habit formation," *Journal of Comparative Neurology of Psychology,* vol. 18, no. 5, pp. 459-482, 1908.

[23] H. Eysenck, "A dynamic theory of anxiety and hysteria," *British Journal of Psychiatry,* vol. 101, no. 1, pp. 28-51, 1955.

[24] M. Corbett, "From law to folklore: work stress and the Yerkes-Dodson Law," *Journal of Managerial Psychology,* vol. 30, no. 6, pp. 741-752, 2015.

[25] Y. Hanoch and O. Vitouch, "When less is more: Information, Emotional Arousal and the Ecological Reframing of the Yerkes-Dodson Law," *Theory & Psychology,* vol. 14, no. 4, pp. 427-452, 2004.

[26] R. Karesek, "Job demands, job decision latitude, and mental strain: Implications for job redesign," *Administrative Science Quarterly,* pp. 285-308, 1979.

[27] J. Johnson and E. Hall, "Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population.," *American Journal of Public Health,* vol. 78, pp. 1336-1342, 1988.

[28] J. Häusser, A. Mojzisch, M. Niesel and S. Schulz-Hardt, "Ten years on: A review of recent research on the Job Demand-Control (-Support) model and psychological well-being," *Work & Stress,* vol. 24, no. 1, pp. 1-35, 2010.

[29] J. Siegrist, "Adverse health effects of high-effort/low-reward conditions.," *Journal of Occupational Health Psychology,* vol. 1, pp. 27-41, 1996.

[30] J. de Jonge, H. Bosma, R. Peter and J. Siegrist, "Job strain, effort-reward imbalance and employee well-being: a large-scale corss-sectional study," *Social Science & Medicine,* vol. 50, pp. 1317-1327, 2000.

[31] E. Demerouti, A. Bakker, F. Nackreiner and W. Schaufeli, "The job demands-resources model of burnout," *Journal of Applied Psychology,* vol. 86, no. 3, pp. 499-512, 2001.

[32] A. Bakker and E. Demerouti, "The Job Demands-Resources model; state of the art," *Journal of Managerial Psychology,* vol. 22, no. 3, pp. 309-328, 2007.

[33] W. Schaufeli and T. Taris, "A Critical Review of the Job Demands-Resources Model: Implications for Improving Work and Health," *Bridging Occupational, Organizational and Public Health,* pp. 43-68, 2014.

[34] S. Cohen, D. Janicki-Deverts and G. Miller, "Psychological Stress and Disease," *JAMA,* vol. 298, no. 14, pp. 1685-1687, 2007.

[35] N. Campbell, J. Reece, L. Mitchell and M. Taylor, Biology: Concepts and Connections, San Fransisco: Pearson, 2005.

[36] M. Stephens and G. Wand, "Stress and the HPA Axis: Role of Glucocorticoids in Alcohol Dependence," *Alcohol Research: Current Reviews,* vol. 34, no. 4, pp. 468-483, 2012.

[37] G. Miller, E. Chen and E. Zhou, "If it Goes Up, Must It Come Down? Chronic Stress and the Hypothalamic-Pituitary-Adrenocortical Axis in Humans," *Psychological Bulletin,* vol. 133, no. 1, pp. 25-45, 2007.

[38] U. Lundberg, "Stress hormones in health and illness: The roles of work and gender," *Psychoneuroendocrinology,* vol. 30, pp. 1017-1021, 2005.

[39] R. Karasek, C. Brisson, N. Kawakami, I. Houtman, P. Bongers and B. Amick, "The Job Content Questionnaire (JCQ): an instrument for internationally comparative assessments of psychosocial job characteristics," *Journal of Occupational Health Psychology,* vol. 3, no. 4, pp. 322-355, 1998.

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

[41] W. Becker and J. Menges, "Biological implicit measures in HRM and OB: A question of how not if," *Human Resource Management Review,* pp. 219-228, 2013.

[42] B. Verkuil, J. Brosschot, M. Tollenaar, R. Lane and J. Thayer, "Prolonged non-metabolic heart rate variability reduction as a physiological marker of psychological stress in daily life," *Annals of Behavioral Medicine,* vol. 50, no. 5, pp. 704-714, 2016.

[43] J. Last, A Dictionary of Epidemiology, New York: Oxford University Press, 2001.

[44] T. McIntyre, S. McIntyre, C. Barr, P. Woodward, D. Francis, A. Durand, P. Mehta and T. Kamarack, "Longitudinal study of the feasibility of using ecological momentary assessment to study teacher stress: Objective and self-reported measures," *Journal of Occupational Health Psychology,* vol. 21, no. 4, pp. 403-414, 2016.

[45] Adams, P., Rabbi, M., Rahman, T., Matthews, M., Voida, A., Gay, G., Choudhury, T., Voida and S., "Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild," *Pervasive Health,* pp. 72-79, 2014.

[46] D. Lee, E. Kim and M. Choi, "Technical and clinical aspects of cortisol as a biochemical marker of chronic stress," *BMB Rep,* vol. 48, no. 4, pp. 209-216, 2015.

[47] K. Wright, R. Hickman and M. Laudenslager, "Hair Cortisol Analysis: A Promising Biomarker of HPA Activation in Older Adults," *Gerontologist,* vol. 55, pp. 140-145, 2015.

[48] N. El-Fahran, D. Rees and C. Evans, "Measuring cortisol in serum, urine and saliva - are our assays good enough?," *Annals of Clinical Biochemistry,* vol. 54, no. 3, pp. 308-322, 2017.

[49] L. Petrakova, B. Doering, S. Vits, H. Engler, W. Rief, M. Schedlowski and J. Grigoleit, "Psychosocial Stress Increases Salivary Alpha-Amylase Activity Independently from Plasma Noradrenaline Levels," *PLoS ONE,* vol. 10, no. 8, pp. 1-9, 2015.

[50] S. Dickerson and M. Kemeny, "Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research," *Psychological Bulletin,* vol. 130, no. 3, pp. 355-391, 2004.

[51] J. Zorn, R. Schür, M. Boks, R. Kahn, M. Joëls and C. Vinkers, "Cortisol stress reactivity across psychiatric disorders: A systematic review and meta-analysis," *Psychoneuroendocrinology,* vol. 77, pp. 25-36, 2017.

[52] Kusserow, M., Amft, O., Troster and G., "Monitoring stress arousal in the wild," *IEEE Pervasive Computing,* vol. 12, no. 2, pp. 28-37, 2013.

[53] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE transactions on intelligent transportation signals,* vol. 6, pp. 156-166, 2005.

[54] J. Hernandez, R. Morris and R. Picard, "Call center stress recognition with person-specific models," *Lecture Notes in Computer Science,* vol. 6974, pp. 125-134, 2011.

[55] P. Karthikeyan, M. Murugappan and S. Yaacob, "A Review on Stress Inducement Stimuli for Assessing Human Stress Using Physiological Signals," in *IEEE 7th International Colloquium on Signal Processing and its Applications*, Penang, 2011.

[56] J. Sriram, M. Shin, T. Choudhury and D. Kotz, "Activity-aware ecg-based patient authentication for remote health monitoring," in *Proceedings of the 2009 International Conference on Multimodel Interfaces*, New York, 2009.

[57] G. E. Prinsloo, W. E. Derman, M. I. Lambert and H. G. L. Rauch, "The effect of a single session of short duration biofeedback-induced deep breathing on measures of heart rate variability during laboratory-induced cognitive stress: a pilot study.," *Appl Psychophysiol Biofeedback,* vol. 38, pp. 81-90, 2013.

[58] P. Karthikeyan, M. Muragappan and S. Yaacob, "Analysis of Stroop Color Word Test-based human stress detection using electrocardiography and heart rate variability signals," *Arab J Sci Eng,* vol. 39, pp. 1835-1847, 2014.

[59] P. Gomes, M. Kaiseler, B. Lopes, S. Faria, C. Queirós and M. Coimbra, "Are standard heart rate variability measures associated with the self-perception of stress of firefighters in action?," in *35th Annual International Conference of the IEEE EMBS*, 2013.

[60] N. Michels, I. Sioen, E. Clays, M. De Buyzere, W. Ahrens, I. Huybrechts, B. Vanaelst and S. De Henauw, "Children's heart rate variability as stress indicator: Association with reported stress and cortisol," *Biological psychology,* vol. 94, no. 2, pp. 433-440, 2013.

[61] A. Rieger, R. Stoll, S. Kreuzfeld, K. Behrens and M. Weippert, "Heart rate and heart rate variability as indirect markers of surgeons' intraoperative stress," *Int Arch Occup Environ Health,* vol. 87, pp. 165-174, 2014.

[62] D. Widjaja, E. Vleminckx and S. Van Huffel, "Stress classification by separation of respiratory modulations in heart rate variability using orthogonal subspace projection," in *EMBC*, Osaka, 2013.

[63] R. Singh, S. Conjeti and R. Banerjee, "A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals," *Biomedical Signal Processing and Control,* vol. 8, pp. 740-754, 2013.

[64] A. Allen, L. Gullixson, S. Wolhart, S. Kost, D. Schroeder and J. Eisenach, "Dietary sodium influences the effect of mental stress on heart rate variability: a randomized trial in healthy adults," *Journal of hypertension,* vol. 32, pp. 374-382, 2014.

[65] U. Acharya, K. Joseph, N. Kannathal, C. Lim and J. Sure, "Heart rate variability: a review," *Med Bio Eng Comput,* vol. 44, pp. 1031-1051, 2006.

[66] K. Roeser, F. Obergfell, A. Meule, C. Vögele, A. Schlarb and A. Kübler, "Of larks and hearts - moringness/eveningness, heart rate variability and cardiovascular stress response at different times of day," *Physiology and behavior,* vol. 106, pp. 151-157, 2012.

[67] J. Choi and R. Gutierrez-Osuna, "Removal of respiratory influences from heart rate variability in stress monitoring," *IEEE Sensors journal,* vol. 11, pp. 2649-2656, 2011.

[68] J. Eisenach, J. Atkinson and R. Fealey, "Development of emotional sweating in preterms measured by skin conductance changes.," *Mayo Clin Proc,* vol. 80, pp. 657-666, 2005.

[69] M. van Dooren, J. de Vries and J. Janssen, "Emotional sweating across the body: comparing 16 different skin conductance measurement

locations.," *Physiology & behavior, 106, 298-304,* vol. 106, pp. 298-304, 2012.

[70] K. Wilke, A. Martin, L. Terstegen and S. Biel, "A short history of sweat gland biology.," *International journal of cosmetic science,* vol. 29, pp. 169-179, 2007.

[71] R. Vertugno, R. Liguori, P. Cortelli and P. Montagna, "Sympathetic skin response: basic mechanism and clinical applications.," *Clin Auton Res,* vol. 13, pp. 256-270, 2003.

[72] M. Svetlak, P. Bob, M. Cernik and M. Kukleta, "Electrodermal complexity during the Stroop colour word test.," *Autonomic neuroscience,* vol. 152, pp. 101-107, 2010.

[73] M. Villarejo, B. Zapirain and A. Zorrilla, "A stress sensor based on galvanic skin response controller by ZigBee.," *Sensors,* vol. 12, pp. 6075-6101, 2012.

[74] R. Hinrichs, V. Michopoulos, S. Winters, A. Rothbaum, B. Rothbaum, K. Ressler and T. Jovanovic, "Mobile assessment of heightened skin conductance in posttraumatic stress disorder," *Depression and Anxiety,* vol. 34, no. 6, pp. 502-507, 2017.

[75] D. Harrison, S. Boyce, P. Loughnan, P. Dargaville, H. Storm and L. Johnston, "Skin conductance as a measure of pain and stress in hospitalised infants.," *Early human development,* vol. 82, pp. 603-608, 2006.

[76] S. Jacobs, R. Friedman, J. Parker, G. Tofler, A. Jimenez, J. Muller, H. Benson and P. Stone, "Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research.," *Am Heart J,* vol. 128, pp. 1170-1177, 1994.

[77] C. Collet, C. Petit, A. Priez and A. Dittmar, "Stroop color-word test, arousal, electrodermal activity and performance in a critical driving situation.," *Biological psychology,* vol. 69, pp. 195-203, 2005.

[78] G. P. I. Schulter, "Bilateral electrodermal activity: relationships to state and trait characteristics of hemisphere asymmetry.," *International journal of psychophysiology, ,* vol. 31, pp. 1-12, 1998.

[79] P. Karthikeyan, M. Murugappan and S. Yaacob, "EMG signal based human stress level classification using wavelet packet transform.," *IRAM 2012 ,* p. 236–243, 2012.

[80] J. Wijsman, B. Grundlehner, J. Penders and H. Hermens, "Trapezius muscle EMG as predictor of mental stress.," *ACM Transactions on Embedded Computing Systems,* vol. 99, no. 12, pp. 1-20, 2013.

[81] L. Schleifer, T. Spalding, S. Kerick, J. Cram, R. Ley and B. Hatfield, "Mental stress and trapezius muscle activation under psychomotor challenge: a focus on EMG gaps during computer work.," *Psychophysiology,* vol. 45, pp. 356-365, 2008.

[82] U. Lundberg, M. Forsman, G. Zachau, M. Eklöf, G. Palmerud, B. Melin and R. Kadefors, "Effects of experimentally induced mental and physical stress on motor unit recruitment in the trapezius muscle.," *An international journal of work, health and organizations,* vol. 16, pp. 166-178, 2002.

[83] G. Krantz, M. Forsman and U. Lundberg, "Consistency in physiological stress responses and electromyographic activity during induced stress exposure in women and men.," *Integrative physiological and behavioral science,* vol. 39, pp. 105-118, 2004.

[84] U. Lundberg, "Psychophysiology of work: stress, gender, endocrine response and work-related upper extremity disorders.," *American journal of industrial medicine,* vol. 41, pp. 383-392, 2002.

[85] I. Veldhuizen, A. Gaillard and J. de Vries, "The influence of mental fatigue on facial EMG activity during a simulated workday.," *Biological Psychology,* vol. 63, pp. 59-78, 2003.

[86] E. Cerezo-Téllez, M. Lacomba, I. Fuentes-Gallardo, O. Mayoral del Moral, B. Rodrigo-Medina and C. Ortega, "Dry needling of the trapezius muscle in office workers with neck pain: a randomized clinical trial," *Journal of Manual & Manipulative Therapy,* vol. 24, no. 4, pp. 223-232, 2016.

[87] K. Holte and R. Westgaard, "Daytime Trapezius Muscle Activity and Shoulder-Neck Pain of Service Workers With Work Stress and Low Biomechanical Exposure," *American Journal of Industrial Medicine,* vol. 41, pp. 393-405, 2002.

[88] T. Sato, T. Kawada, T. Shishido, M. Sugimachi, J. Alexander and K. Sunagawa, "Novel therapeutic strategy against central baroreflex faillure: a bionic baroreflex system.," *Circulation,* vol. 100, pp. 299-304, 1999.

[89] Q. B. L. C. J. L. J. C. J. Zhao, J. Huang, J. Chen, T. Kelly, C.-S. Chen, D. Hu, J. Ma, T. Rice, J. He and D. Gu, "Reproducibility of blood pressure response to the cold pressor test.," *American journal of epidemiology,* vol. 17, pp. 91-98, 2012.

[90] D. Carroll, C. Ring, K. Hunt, G. Ford and S. Macintyre, "Blood pressure reactions to stress and the prediction of future blood

pressure effects of sex age and socioeconomic position.," *Psychosomatic medicine,* vol. 65, pp. 1058-1064, 2003.

[91] K. Matthews, J. Owens, M. Allen and C. Stoney, "Do cardiovascular reponses to laboratory stress relate to ambulatory blood pressure levels?: Yes, in some of the people, some of the time.," *Psychosomatic medicine,* vol. 54, pp. 686-697, 1992.

[92] C. Ottaviani, J. Brosschot, A. Lonigro, B. Medea, I. Van Diest and J. Thayer, "Hemodynamic Profiles of Functional and Dysfunctional Forms of Repetitive Thinking," *Ann. Behav. Med.,* vol. 51, no. 2, pp. 261-271, 2017.

[93] E. O'Brien, "Twenty-four-hour ambulatory blood pressure measurement in clinical practica and research: a critical review of a technique in need of implementation.," *Journal of internal medicine,* vol. 26, pp. 478-495, 2011.

[94] A. Kistler, C. Mariauzouls and K. von Berlepsch, "Fingertip temperature as an indicator for sympathetic response.," *International journal of psychophysiology,* vol. 29, pp. 35-41, 1998.

[95] P. Karthikeyan, M. Murugappan and S. Yaacob, "Multiple physiological signal-based human stress identification using non-linear classifiers," *Elektronika ir elektrotechnika,* vol. 19, no. 7, pp. 80-85, 2013.

[96] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing.," *Int J Human-Computer Studies,* vol. 59, pp. 185-198, 2003.

[97] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," *Proceedings of the 28th IEEE EMBS Annual International Conference,* pp. 1355-1358, 2006.

[98] G. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures.," *International journal of aviation psychology,* vol. 12, pp. 3-18, 2002.

[99] Katsis, C.D., Katertsidis, N., Ganiatsas, G., Fotiadis and D.I., "Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach," *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans,* vol. 38, no. 3, pp. 502-512, 2008.

[100] Wijsman, J., Vullers, R., Polito, S., Hermens and H., "Towards Ambulatory Mental Stress Measurement from Physiological Parameters," in *Humain Association Conference on Affective Computing and Intelligent Interaction*, Geneva, 2013.

[101] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang and Z. Zhao, "Detecting work-related stress with a wearable device," *Computers in Industry,* vol. 90, pp. 42-49, 2017.

[102] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto and O. Mozos, "Stress Detection Using Wearable Physiological Sensors," *IWINAC 2015: Artificial Computation in Biology and Medicine,* pp. 526-532, 2015.

[103] J. de Vries, S. Pauws and M. Biehl, "Insightful stress detection from physiology modalities using Learning Vector Quantization," *Neurocomputing,* vol. 151, pp. 873-882, 2015.

[104] A. de Santos Sierra, C. Sanchez Avila, J. Guerra Casanova and G. Bailador del Pozo, "A Stress-Detection System Based on Physiological Signals and Fuzzy Logic," *IEEE Transactions on Industrial Electronics,* vol. 58, no. 10, pp. 4857-4865, 2011.

[105] Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., Griss and M., "Activity-aware mental stress detection using physiological sensors," in *Mobile Computing applications and services*, 2012, pp. 211-230.

[106] R. Khusainov, D. Azzi, I. Achumba and S. Bersch, "Real-Time Human Ambulation, Activity, and Physiological Monitoring: Taxonomy of Issues, Techniques, Applications, Challenges and Limitations," *Sensors,* vol. 13, pp. 12852-12901, 2013.

[107] Hovsepian, K., al'Absi, M., Ertin, E., Kamarack, T., Nakajima, M., Kumar and S., "cStress: Towards a gold standard for continuous stress assessment in the mobile environment," *Proc ACM Int Conf Ubiquitous Comput,* pp. 493-504, 2015.

[108] A. Hayes and J. Matthes, "Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations," *Behavior Research Methods,* vol. 41, no. 3, pp. 924-936, 2009.

[109] E. Chumney and K. Simpson, Methods and Designs for Outcomes Research, Bethesda: ASHP Publications Production Center, 2006.

[110] StatisticsSolutions, "Homoscedasticity," 2013. [Online]. Available: http://www.statisticssolutions.com/homoscedasticity/. [Accessed 16 1 2018].

[111] S. Menard, Applied Logistic Regression Analysis, London: Sage Publications, 2002.

[112] S. Sayad, "Logistic regression," 2018. [Online]. Available: http://www.saedsayad.com/logistic_regression.htm. [Accessed 16 1 2018].

[113] B. Lantz, Machine Learning with R - Second Edition, Birmingham: Packt Publishing, 2015.

[114] C. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Data mining and knowledge discovery,* vol. 2, no. 2, pp. 121-167, 1998.

[115] S. Sayad, "Support Vector Machine - Classification (SVM)," 2018. [Online]. Available: http://www.saedsayad.com/support_vector_machine.htm. [Accessed 16 1 2018].

[116] S. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data mining and knowledge discovery,* vol. 2, no. 4, pp. 345-389, 1998.

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

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

[119] N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning,* vol. 29, no. 2, pp. 131-163, 1997.

[120] W. Liao, W. Zhang, Z. Zhu and Q. Ji, "A real-time human stress monitoring system using dynamic Bayesian Networks," in *Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[121] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang and Z. Zhao, "Detecting work-related stress with a wearable device," *Computers in Industry,* vol. 90, pp. 42-49, 2017.

[122] O. Mozos, V. Sandulescu, S. Andrews, D. B. N. Ellis, R. Dobrescu and J. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *International Journal of Neural Systems,* vol. 27, no. 2, pp. 1-17, 2017.

[123] D. Huysmans, E. Smets, W. De Raedt, C. Van Hoof, K. Bogaerts, I. Van Diest and D. Helic, "Unsupervised Learning for Mental Stress Detection - Exploration of Self-Organizing Maps," in *Proceedings of Biosignals 2018*, Madeira, 2018.

[124] G. Rigas, Y. Goletsis and D. Fotiadis, "Real-time driver's stress event detection," *IEEE Transactions on intelligent transportation systems,* vol. 13, no. 1, pp. 221-234, 2012.

[125] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al'Absi and S. Shah, "AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, Seattle, Washington, 2011.

[126] R. Kocielnik, N. Sidorova, F. Maggi, M. Ouwerkerk and J. Westerink, "Smart technologies for long-term stress monitoring at work," in *CBMS*, 2013.

[127] B. Lamichhane, U. Großekathöfer, G. Schiavone and P. Casale, "Towards Stress Detection in Real-Life Scenarios Using Wearable Sensors: Normalization Factor to Reduce Variability in Stress Physiology," *eHealth 360°,* pp. 259-270, 2016.

[128] E. Smets, P. Casale, U. Großekathöfer, B. Lamichhane, W. De Raedt, K. Bogaerts, I. Van Diest and C. Van Hoof, "Comparison of Machine Learning Techniques for Psychophysiological Stress Detection," *Pervasive Computing Paradigms for Mental Health,* vol. 604, pp. 13-22, 2016.

[129] J. Wijsman, B. Grundlehner and H. Hermens, "Wearable physiological sensors reflect mental stress state in office-like situations," in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[130] A. Sano and R. Picard, "Stress recognition using wearable sensors and mobile phones," in *Humain Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[131] C. Stoney, M. Davis and K. Matthews, "Sex differences in physiological responses to stress and in coronary heart disease: A causal link?," *Psychophysiology,* vol. 24, no. 2, pp. 127-131, 1987.

[132] W. Van der Elst, P. Van Boxtel, J. Van Breukelen and J. Jolles, "The Stroop Color-Word test: Influence of age, sex, and education; and normative data for a large sample accros the adult age range," *Assessment,* vol. 13, pp. 62-79, 2006.

[133] L. Brown, B. Grundlehner, J. v. d. Molengraft, J. Penders and B. Gyselinckx, "Body area network for monitoring autonomic nervous system responses," *Pervasive Computing Technologies for Healthcare,* 2009.

[134] K. Murphy, "The BayesNet Toolbox for Matlab," *Computing Science and Statistics,* vol. 33, pp. 331-350, 2001.

[135] E. Mezzacappa, R. Kelsey, E. Katkin and R. Sloan, "Vagal rebound and recovery from psychological stress," *Psychosomatic Medicine,* vol. 63, no. 4, pp. 650-657, 2001.

[136] D. Giakoumis, D. Tzovaras and G. Hassapis, "Subject-dependent biosignal features for increased accuracy in psychological stress detection," *Int J. Human-Computer Studies,* vol. 71, pp. 425-439, 2013.

[137] E. Smets, G. Schiavone, E. Velazquez, W. De Raedt, K. Bogaerts, I. Van Diest and C. Van Hoof, "Comparing task-induced psychophysiological responses between persons with stress-related complaints and healthy controls: A methodological pilot study," *Health Science Reports,* 2018.

[138] S. Stansfeld and B. Candy, "Psychosocial work environment and mental health - a meta-analytic review," *Scandinavian Journal of Work, Environment & Health,* vol. 32, no. 6, pp. 443-462, 2006.

[139] A. Rosengren, S. Hawken, S. Ôunpuu, K. Sliwa, M. Zubaid, W. Almahmeed, K. Ngu Blackett, C. Sitthi-amorn, H. Sato and S. Yusuf, "Association of psychosocial risk factors with risk of acute myocardial infarction in 11 119 cases and 13 648 controls from 52 countries (the INTERHEART study): case-control study," *The Lancet,* vol. 364, pp. 953-962, 2004.

[140] J. Dimsdale, "Psychological Stress and Cardiovascular Disease," *Journal of the American College of Cardiology,* vol. 51, no. 13, pp. 1237-1246, 2008.

[141] B. Terluin, J. Van der Klink and W. Schaufeli, "Stressgerelateerde klachten: spanningsklachten, overspanning en burnout," in *Psychische problemen en werk. Handboek voor een activerende begeleiding door huisarts en bedrijfsarts*, Houten, Bohn Stafleu Van Loghum, 2005, pp. 259-290.

[142] R. May, G. Seibert, M. Sanchez-Gonzalez and F. Fincham, "Physiology of school burnout in medical students: Hemodynamic and autonomic functioning," *Burnout Research,* pp. 63-68, 2016.

[143] C. Morgan, T. Cho, G. Hazlett, V. Coric and J. Morgan, "The impact of burnout on human physiology and on operational performance: A prospective study of soldiers enrolled in the combat diver qualification course," *Yale Journal of Biology and Medicine,* vol. 75, no. 4, pp. 199-205, 2002.

[144] W. De Vente, M. Olff, J. Van Amsterdam, J. Kamphuis and P. Emmelkamp, "Physiological differences between burnout patients and healthy controls: blood pressure, heart rate, and cortisol responses," *Occupational and environmental medicine,* vol. 60, pp. 54-61, 2003.

[145] P. Mommersteeg, The psychophysiology of burnout, Enschede: Febo Druk B.V., 2006.

[146] D. Mohr, M. Zhang and M. Schueller, "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning," *Annual Review of Clinical Psychology,* vol. 13, no. 8, pp. 23-47, 2017.

[147] E. Nishime, C. Cole, E. Blackstone, F. Pashkow and M. Lauer, "Heart rate recovery and treadmill exercise score as predictors of mortality in patients referred for exercise ECG," *Journal of the American Medical Association,* vol. 284, no. 11, pp. 1392-1398, 2000.

[148] B. McEwen, "Stress, adaptation, and disease - Allostasis and allostatic load," *Annals of the New York Academy of Science,* vol. 840, pp. 33-44, 1998.

[149] W. Linden, T. Earle, W. Gerin and N. Christenfeld, "Phsyiological stress reactivity and recovery: conceptual siblings separated at birth?," *Journal of Psychosomatic Research,* vol. 42, no. 2, pp. 117-135, 1997.

[150] I. Smits, M. Timmerman, D. Barelds and R. Meijer, "The Dutch Symptom Checklist-90-Revised: Is the Use of Subscales Justified?," *European Journal of Psychological Assessment,* vol. 31, no. 4, pp. 263-271, 2015.

[151] M. Holli, Assessment of psychiatric symptoms using the SCL-90, Helsinki: Helsinki University Printing House, 2003.

[152] J. Vandixhoorn and H. Duivenvoorden, "Efficacy of Nijmegen questionnaire in recognition of the hyperventilation syndrome," *Journal of Psychosomatic Research,* vol. 29, no. 2, pp. 199-206, 1985.

[153] D. Sheehan, Y. Lecrubier, K. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker and G. Dunbar, "The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10," *Journal of Clinical Psychiatry,* vol. 59, pp. 22-33, 1998.

[154] T. Overbeek, K. Schruers and E. Griez, "Mini International Neuropsychiatric Interview," University of Maastricht, Maastricht, 1999.

[155] J. Cohen, Statistical Power Ananlysis for the Behavioral Sciences, Hillsdale, NJ: Lawrence Erlbaum, 1988.

[156] C. Lim, C. Rennie, R. Barry, H. Bahramali, I. Lazzaro, B. Manor and E. Gordon, "Decomposing skin conductance into tonic and phasic components," *International Journal of Psychophysiology,* vol. 25, no. 2, pp. 97-109, 1997.

[157] Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., Griss and M., "Activity-aware mental stress detection using physiological sensors," *Mobile computing applications and services,* pp. 211-230, 2012.

[158] N. Bolger, A. Davis and E. Rafaeli, "Diary methods: Capturing lif as it is lived," *Annual Review of Psychology,* vol. 54, pp. 579-616, 2003.

[159] K. Dedovic, R. Renwick, N. Mahani, V. Engert, S. Lupien and J. Pruesnner, "The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain," *Journal of Psychiatry & Neuroscience,* vol. 30, no. 5, pp. 319-325, 2005.

[160] D. Buysse, C. Reynolds, T. Monk, C. Hoch, A. Yeager and D. Kupfer, "Quanitifcation of subjective sleep quality in healthy elderly men and women using the Pittsburgh Sleep Quality Index (PSQI)," *Sleep,* vol. 14, no. 4, pp. 331-338, 1991.

[161] H. Knudson, L. Ducharme and P. Roman, "Job stress and poor sleep quality: Data from an American sample of full-time workers," *Social Science & Medicine,* vol. 64, no. 10, pp. 1997-2007, 2007.

[162] J. Henry and J. Crawford, "The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample," *British Journal of Clinical Psychology,* vol. 44, no. 2, pp. 227-239, 2005.

[163] R. Hays and L. Morales, "The RAND-36 measure of health-related quality of life," *Annals of Medicine,* vol. 33, no. 5, pp. 350-357, 2001.

[164] H. Petersen, G. Kecklund, P. D'Onofrio, J. Nilsson and T. Akerstedt, "Stress vulnerability and the effects of moderate daily stress on sleep polysomnography and subjective sleepiness," *J. Sleep Res.,* vol. 22, pp. 50-57, 2013.

[165] S. Lee, I.-K. Sung, J. Kim, S.-Y. Lee, H. Park and C. Shim, "The Effect of Emotional Stress and Depression on the Prevalence of Digestive Diseases," *J. Neurgastroenterol. Motil.,* vol. 21, no. 2, pp. 273-282, 2015.

[166] F. Carbone, A. Vandenberghe, L. Holvoet, T. Vanuytsel, L. Van Oudenhove, M. Jones and J. Tack, "Validation of the Leuven

Postprandial Distress Scale, a questionnaire for symptom assessment in the functional dyspepsia/postprandial distress syndrome," *Aliment Pharmacol Ther,* vol. 44, no. 9, pp. 989-1001, 2016.

[167] J. Morris, "Observations: SAM: The self-assessment manikin - An efficient cross-cultural measurement of emotional response," *Journal of Advertising Research,* vol. 35, no. 6, pp. 63-68, 1995.

[168] R. Kupriyanov and R. Zhdanov, "The Eustress Concept: Problems and Outlooks," *World Journal of Medical Sciences,* vol. 11, no. 2, pp. 179-185, 2014.

[169] M. Hamer, R. Endrighi and L. Poole, "Physical Activity, Stress Reduction, and Mood: Insight into Immunological Mechanisms," *Psychoneuroimmunology,* pp. 89-102, 2012.

[170] M. Van der Doef and S. Maes, "The Job Demand-Control(-Support) model and psychological well-being: a review of 20 years of empirical research," *Work and Stress,* vol. 13, no. 2, pp. 87-114, 1999.

[171] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Vallance and L. Tarassenko, "Signal-quality indices for the electrocardiogram and photoplethismogram: Derivation and applications to wireless monitoring," *IEEE Journal of biomedical and health informatics,* vol. 19, no. 3, pp. 832-838, 2015.

[172] W. Boucsein, D. Fowles, S. Grimnes, G. Ben-Shakhar, W. Roth, M. Dawson and D. Filion, "Publication recommendations for electrodermal measurements," *Psychophysiology,* vol. 49, no. 8, pp. 1017-1034, 2012.

[173] L. Jones and S. Lederman, Human hand function, New York: Oxford University Press Inc., 2006.

[174] P. Silvia, T. Kwapil, K. Eddington and L. Brown, "Missed Beeps and Missing Data: Dispositional and Situational Predictors of Nonresponse in Experience Sampling Research," *Social Science Computer Review,* vol. 31, no. 4, pp. 471-487, 2013.

[175] R. Wang, E. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev and A. Campbell, "StudentLife: Assessing Mental Health, Academinc Performance and Behavioral Trends of College Students using Smartphones," in *Proceedings of the ACM Conference on Ubiquitous Computing*, 2014.

[176] J. Ho and S. Intille, "Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices," in *CHI*, Portland, 2005.

[177] F. Naughton, S. Hopewell, N. Lathia, R. Schalbroeck, C. Brown, C. Mascolo, A. McEwen and S. Sutton, "A Context-Sensing Mobile Phone App (Q Sense) for Smoking Cessation: A Mixed-Methods Study," *JMIR Mhealth Uhealth,* vol. 4, no. 3, pp. 1-13, 2016.

[178] D. Gustafson, F. McTavish, M. Chic, A. Atwood, R. Johnson, M. Boyle, M. Levy, H. Driscoll, S. Chisholm, L. Dillenburg, A. Isham and D. Shah, "A smartphone application to support recovery from alcoholism: a randomized clinical trial," *JAMA Psychiatry,* vol. 71, no. 5, pp. 566-572, 2014.

[179] R. Lambiotte and M. Kosinski, "Tracking the digital footprints of personality," *Proceedings of the IEEE,* vol. 102, no. 12, pp. 1934-1939, 2014.

[180] M. Rhoen, "Rear view mirror, crystal ball: Predictions for the future of data protection law based on the history of environmental protection law," *Computer Law & Security Review,* vol. 33, no. 5, pp. 603-617, 2017.

[181] J. Moeyersons, E. Smets, J. Morales, A. Villa, W. De Raedt, D. Testelmans, B. Buyse, C. Van Hoof, R. Willems, S. Van Huffel and C. Varon, "Artefact detection and quality assessment of ambulatory ECG signals (submitted)," *Computer Methods and Programs in Biomedicine,* 2018.

[182] B. Marr, "15 Noteworthy Facts About Wearables In 2016," 18 March 2016. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2016/03/18/15-mind-boggling-facts-about-wearables-in-2016/#102e6f272732. [Accessed 4 January 2018].

[183] EndeavourPartners, "Inside Wearables Part 2 - July 2014," 2017 April 27. [Online]. Available: https://blog.endeavour.partners/inside-wearables-part-2-july-2014-ef301d425cdd. [Accessed 4 January 2018].

[184] OxfordDictionaries, "Phenotype," [Online]. Available: https://en.oxforddictionaries.com/definition/phenotype. [Accessed 13 2 2018].

[185] S. Jain, B. Powers, J. Hawkins and J. Brownstein, "The digital phenotype," *Nature Biotechnology,* vol. 33, no. 5, pp. 462-463, 2015.

[186] J. Brownstein, C. Freifeld and L. Madoff, "Digital Disease Detection - Harnessing the Web for Public Health Surveillance," *N Engl J Med,* vol. 360, no. 21, pp. 2153-2157, 2009.

[187] V. Duric, S. Clayton, M. Leong and L.-L. Yuan, "Comorbidity Factors and Brain Mechanisms Linking Chronic Stress and Systemic Illnesses," *Neural Plasticity,* pp. 1-16, 2016.

[188] J. Hernandez, P. Paredes, A. Roseway and M. Czerwinski, "Under pressure: sensing stress of computer users," in *CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Toronto, 2014.

[189] Q. Xu, T. Nwe and C. Guan, "Cluster-Based Analysis for Personalized Stress Evaluation Using Physiological Signals," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, no. 1, pp. 275-281, 2015.

[190] Rahman, M., Bari, R., Ali, A.A., Sharmin, M., Raij, A., Hovsepian, K., Hossain, S.M., Ertin, E., Kennedy, A., Epstein, D.H., Preston, K.L., Jobs, M., Beck, J.G., Kedia, S., Ward, K.D., al'Absi, M., Kumar and S., "Are We There Yet? Feasibility of Continuous Stress Assessment via Wireless Physiological Sensors," *ACM BCB,* pp. 479-488, 2014.

[191] I. Romero, B. Grundlehner and J. Penders, "Robust beat detector for ambulatory cardiac monitoring," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, USA, 2009.

[192] A. Camm, M. Malik, J. Bigger, G. Breithardt, S. Cerutti, R. Cohen, P. Coumel, E. Fallen, H. Kennedy, R. Kleiger, F. Lombardi, A. Malliani, A. Moss, J. Rottman, G. Schmidt, P. Schwartz and D. Singer, "Heat rate variability - standard of measurement, physiological interpretation and clinical use," *European Heart Journal,* vol. 93, no. 5, pp. 1043-1065, 1996.

[193] C. Grant, D. van Rensburg, N. Strydom and M. Viljoen, "Importance of tachogram length and period of recording during noninvasive investigation of the autonomic nervous system," *Annals of Noninvasive Electrocardiology,* vol. 16, no. 2, pp. 131-139, 2011.

[194] D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.,* vol. 67, pp. 1-48, 2015.

[195] A. Steptoe and N. Butler, "Sports participation and emotional wellbeing in adolescents," *The Lancet,* vol. 347, no. 9018, pp. 1789-1792, 1996.

[196] B. Stubbs, N. Veronese, D. Vancampfort, A. Prina, P.-Y. Lin, P.-T. Tseng, E. Evangelou, M. Solmi, C. Kohler, A. Carvalho and A. Koyanagi, "Perceived stress and smoking across 41 countries: A global perspective across Europe, Africa, Asia and the Americas," *Scientific Reports,* vol. 7, no. 7579, pp. 1-8, 2017.

[197] C. Rius, E. Fernandez, A. Schiaffino, F. Borràs and F. Rodríguez-Artalejo, "Self perceived health and smoking in adolescents," *J. Epidemiol. Community Health,* vol. 58, pp. 698-699, 2004.

[198] E. Watson, A. Coates, M. Kohler and S. Banks, "Caffeine Consumption and Sleep Quality in Australian Adults," *Nutrients,* vol. 479, pp. 1-10, 2016.

[199] S. Smagula, K. Stone, A. Fabio and J. Cauley, "Risk factors for sleep disturbances in older adults: Evidence from prospective studies," *Sleep Medicine Reviews,* vol. 25, pp. 21-30, 2016.

[200] L. Irish, C. Kline, H. Gunn, D. Buysse and M. Hall, "The role of sleep hygiene in promoting public health: A review of empirical evidence," *Sleep Medicine Reviews,* vol. 22, pp. 23-36, 2015.

[201] Z. Li, H. Snieder, S. Su, X. Ding, J. Thayer, F. Treiber and X. Wang, "A longitudinal study in youth of heart rate variability at rest and in response to stress," *International Journal of Psychophysiology,* vol. 73, no. 3, pp. 212-217, 2009.

[202] J. Finke, G. Kalinowski, M. Larra and H. Schachinger, "The socially evaluated handgrip test: Introduction of a novel, time-efficient stress protocol," *Psychoneuroendocrinology,* vol. 87, pp. 141-146, 2018.

[203] C. Kirschbaum, K. Pirke and D. Hellhammer, "The Trier Social Stress Test - A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting," *Neuropsychobiology,* vol. 28, no. 1-2, pp. 76-81, 1993.

[204] G. Billman, "The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance," *Frontiers in Physiology,* pp. 4-26, 2013.

[205] J. McNames, T. Thong and B. Goldstein, "Reliability and Accuracy of Heart Rate Variability Metrics Versus ECG Segment Duration," in *Proceedings of the 25th annual international conference of the IEEE Engineering in Medicine and Biology Society*, Cancun, 2003.

[206] J. Healey, Ph.D. Thesis: Wearable and Automotive Systems for Affect Recognition from Physiology, Massachusetts Institute of Technology, 2000.

[207] A. Greco, G. Valenza and E. Scilingo, Advances in Electrodermal Activity Processing with Applications for Mental Health, Chamonix: Springer Interantional Publishing, 2016.

[208] D. Carroll, A. Ginty, A. Whittaker, W. Lovallo and S. de Rooij, "The behavioral, cognitive, and neural corollaries of blunted cardiovascular

and cortisol reactions to acute psychological stress," *Neuroscience and Biobehavioral Reviews,* vol. 77, pp. 74-86, 2017.

[209] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, no. 1, pp. 159-174, 1977.

[210] A. Dagher, B. Tannenbaum, T. Hayashi, J. Pruessner and D. McBride, "An acute psychosocial stress enhances the neural response to smoking cues," *Brain Research,* pp. 40-48, 2009.

[211] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. Patel, A. Tiwari, M. Er, W. Ding and C.-T. Lin, "A review of clustering techniques and developments," *Neurcomputing,* vol. 267, pp. 664-681, 2017.

[212] R. Mukkamala, J.-O. Hahn, O. Inan, L. Mestha, C.-S. Kim, H. Töreyin and S. Kyal, "Toward Ubiquitous Blood Pressure Monitoring via Pulse Transit Time: Theory and Practice," *IEEE Transactions on Biomedical Engineering,* vol. 62, no. 8, pp. 1879-1901, 2015.

[213] I. Castro, C. Varon, T. Torfs, S. Van Huffel, R. Puers and C. Van Hoof, "Evaluation of a Multichannel Non-Contact ECG System and Signal Quality Algorithms for Sleep Apnea Detection and Monitoring," *Sensors,* vol. 18, no. 577, pp. 1-20, 2018.

[214] J. Pinheiro, Mixed-Effect Models in S and S-plus, New York: Springer-Verlag, 2002.

[215] S. Tuarob, C. Tucker, S. Kumara, C. Giles, A. Pincus, D. Conroy and N. Ram, "How are you feeling?: A personalized methodology for predicting mental states from temporarily observable physical and behavioral information," *Journal of Biomedical Informatics,* vol. 68, pp. 1-19, 2017.

[216] V. Narayan, M. Mohwinckel, G. Pisano, M. Yang and H. Manji, "Beyond magic bullets: true innovation in health care," *Nature Reviews Drug Discovery,* vol. 12, pp. 85-86, 2013.

[217] S. Burcusa and W. Iacono, "Risk for Recurrence in Depression," *Clin. Psychol. Rev.,* vol. 27, no. 8, pp. 959-985, 2007.

[218] O. Janssens, E. Smets, G. Schiavone, E. Rios Velazquez, F. Ongenae, W. De Raedt, C. Van Hoof and S. Van Hoecke, "Context-aware stress detection," in *Biomedical and Healthy Informatics Conference (BHI)*, Las Vegas, 2018.

[219] A. Sano, S. Taylor, A. McHill, A. Phillips, L. Barger, E. Klerman and R. Picard, "Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using

Wearable Sensors and Mobile Phones: Observational Study," *Journal of Medical Internet Research,* vol. 20, no. 6, pp. 1-20, 2018.

[220] M. Farooq, M. McCrory and E. Sazonov, "Reduction of energy intake using just-in-time feedback from a wearables sensor system," *Obesity,* vol. 25, no. 4, pp. 676-681, 2017.

[221] A. Müller, A. Blandford and L. Yardley, "The conceptualization of a Just-In-Time Adaptive Intervention (JITAI) for the reduction of sedentary behavior in older adults," *Mhealth,* vol. 3, no. 37, 2017.

[222] I. Nahum-Shani, S. Smith, B. Spring, L. Collins, K. Witkiewitz, A. Tewari and S. Murphy, "Just-In-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support," *Annals of Behavioral Medicine,* pp. 1-17, 2013.

[223] C. Cerrada, E. Dzubur, K. Blackman, V. Mays, S. Shoptaw and J. Huh, "Development of a Just-in-Time Adaptive Intervention for Smoking Cessation Among Korean American Emerging Adults," *Int. J. Behav. Med.,* vol. 24, no. 5, pp. 665-672, 2017.

[224] J. Smyth and K. Heron, "Is providing mobile interventions "just-in-time" helpful? An experimental proof of concept study of just-in-time intervention for stress management," in *Wireless Health*, Bethesda, 2016.

[225] A. Stewart and J. Ware, Measuring Functioning and well-being: The medical outcomes study approach, Durham: Duke University Press, 1992.

Publication list

PUBLICATIONS IN INTERNATIONAL PEER-REVIEWED JOURNALS

E. Smets, G. Schiavone, E. Velazquez, W. De Raedt, K. Bogaerts, I. Van Diest and C. Van Hoof, "Comparing task-induced psychophysiological responses between persons with stress-related complaints and healthy controls: A methodological pilot study," Health Science Reports, 2018.

E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes, I. Van Diest and C. Van Hoof, "Large-Scale Wearable Data Reveal Digital Phenotypes for Stress Detection," npj Digital Medicine (submitted)

E. Smets, E. Rios Velazquez, G. Schiavone, W. De Raedt, A. Goris and C. Van Hoof, "The Montreal Imaging Stress Task as Calibration towards Personalized Ambulatory Physiological Stress Detection," IEEE Transactions on Affective Computing (submitted).

E. Smets, W. De Raedt, C. Van Hoof., "Into the Wild: The Challenges of Physiological Stress Detection in Laboratory and Ambulatory Settings," IEEE Journal of Biomedical and Health Informatics (submitted)

J. Moeyersons, **E. Smets**, J. Morales, A. Villa, W. De Raedt, D. Testelmans, B. Buyse, C. Van Hoof, R. Willems, S. Van Huffel and C. Varon, "Artefact detection and quality assessment of ambulatory ECG signals," Computer Methods and Programs in Biomedicine (submitted).

CONTRIBUTIONS AT INTERNATIONAL CONFERENCES

C. Van Hoof, G. Schiavone, **E. Smets**, P. Casale, W. De Raedt and R. De Fransisco Martin, "Wearables and IoT for persuasive technology and percipient systems," in Semicon Korea, Seoul, 2015.

E. Smets, P. Casale, U. Großekathöfer, B. Lamichhane, W. De Raedt, K. Bogaerts, I. Van Diest and C. Van Hoof, "Comparison of Machine Learning Techniques for Psychophysiological Stress Detection," Pervasive Computing Paradigms for Mental Health, vol. 604, pp. 13-22, 2016. (awarded 'Best Student Paper Award')

D. Huysmans, **E. Smets**, W. De Raedt, C. Van Hoof, K. Bogaerts, I. Van Diest and D. Helic, "Unsupervised Learning for Mental Stress Detection - Exploration of Self-Organizing Maps," in Proceedings of Biosignals 2018, Madeira, 2018.

O. Janssens, **E. Smets**, G. Schiavone, E. Rios Velazquez, F. Ongenae, W. De Raedt, C. Van Hoof and S. Van Hoecke, "Context-aware stress detection," in Biomedical and Healthy Informatics Conference (BHI), Las Vegas, 2018.

S. Anrijs, K. Bombeke, W. Durnez, K. Van Damme, B. Vanhaeleweyn, P. Conradie, **E. Smets**, J. Cornelis, W. De Raedt, K. Ponnet and L. De Marez, "MobileDNA: Relating physiological stress measurements to smartphone usage to assess the effect of a digital detox," in HCI International, Las Vegas, 2018.

PATENTS

E. Smets, G. Schiavone, E. Rios Velazquez, W. De Raedt and C. Van Hoof, "Personalized stress detection in daily living". Europe Patent EP 17209817.0, 21 12 2017.

Appendix A

Table A-1: Overview of intake questionnaire. For continuous variables the mean and standard deviation of the population (932 subjects who filled in the questionnaire) are presented, for categorical variables the percentage of the population for each class is presented. Subjects not answering the specific question are denoted by answer value 'NaN'

Variable	*Classes*
Age	Mean = 39.5, SD = 9.8
Length (cm)	Mean = 170.1, SD = 26.9
Weight (kg)	Mean = 73.8, SD = 14.5
Gender	Female (n=446, 44.5%) Male (n=481, 48.0%) NaN (n=75, 7.5%)
Origin	Africa (n=7, 0.7%) Asia (n=45, 4.5%) Europe (n=854, 85.2%) South America (n=13, 1.3%) Central America (n=3, 0.3%) North America (n=6, 0.6%) NaN (n=74, 7.4%)
Marital status	Single (n=183, 18.3%) Cohabiting (n= 229, 22.9%) Married (n=450, 44.9%) Divorced (n=57, 5.7%) Widowed (n=9, 0.9%) NaN (n=74, 7.4%)
Children	Yes (n=549, 54.8%) No (n= 380, 37.9%) NaN (n=73, 7.3%)
If "Yes" to "Children": how many	Mean = 2.1, SD = 0.8
Pregnant	Yes (n=9, 0.9%) No (n=919, 91.7%)

		NaN (n=74, 7.4%)
Healthy lifestyle (1 = unhealthy – 10 = healthy)		Mean = 6.0, SD = 1.8
Sports		Yes (n= 687, 68.6%) No (n=251, 25.0%) NaN (n=64, 6.4%)
	If "Yes" to "Sports": Hours of sports per week	0-1h (n=118, 11.8%) 1-3h (n=335, 33.4%) 3-5h (n=153, 15.3%) >5h (n=79, 7.9%) NaN (n=317, 31.6%)
	If "No" to "Sports": other hobbies	Yes (n=182, 18.2%) No (n=69, 6.9%) NaN (n=751, 74.9%)
Smoke		Yes (n=74, 7.4%) No (n=856, 85.4%) NaN (72, 7.2%)
	If "Yes" to "Smoke": how many cigarettes	Less than 1 per week (n=6, 0.6%) 1-6 per week (n=17, 1.7%) 1-5 per day (n=16, 1.6%) 6-10 per day (n=24, 2.4%) 10-20 per day (n=12, 1.2%) >20 per day (n=2, 0.2%) NaN (n=925, 92.3%)
Caffeinated beverages		Yes (n=796, 79.4%) No (n=132, 13.2%) NaN (n=74, 7.4%)
	If "Yes" to "caffeinated beverages": how many cups per week	Mean = 15.5, SD = 10.8
Alcohol		Yes (766, 76.4%) No (n=163, 16.3%) NaN (n=73, 7.3%)

If "Yes" to "Alcohol": how many glasses per week	Mean = 6.9, SD = 6.9
Fruit and vegetables	No (n=9, 0.9%) 1-6 portions per week (n=156, 15.6%) 1 portion per day (n=179, 17.9%) 2-3 portions per day (n=410, 40.9%) 4-5 portions per day (n=132, 13.2%) >5 portions per day (n=39, 3.9%) NaN (n=77, 7.7%)
Nr of take-out meals per week	Mean = 1.0, SD =1.0
Diet	Pescetarian (n=22, 2.2%) Vegetarian (n=24, 2.4%) Vegan (n=7, 0.7%) None of the above (n=867, 86.5%) NaN (n=82, 8.2%)
Medication	Yes (n=305, 30.5%) No (n=614, 61.3%) NaN (n=83, 8.3%)
If "Yes" to "Medication": Which medication	*Open text*
Current heart disease	Yes (n=26, 2.6%) No (n= 893, 89.1%) NaN (n=83, 8.3%)
If "Yes" to "Current heart disease": Which disease	*Open text*
Heart disease in the past	Yes (n=23, 2.3%) No (n= 894, 89.2%) NaN (n=85, 8.5%)

If "Yes" to "Heart disease in the past": Which disease	*Open text*
Chronic disease	Yes (n=113, 11.3%) No (n=793, 79.1%) NaN (n=96, 9.6%)
If "Yes" to "Chronic disease": Which disease	*Open text*
Education	Primary school (n=2, 0.2%) Secondary school (n=62, 6.2%) Bachelors (n=303, 30.2%) Masters/PhD (n=509, 50.8%) NaN (n=126, 12.6%)
Employee type	Full-time (n=606, 60.5%) Part-time (n=152, 15.2%) Shift worker (n=16, 1.6%) Interim (n=1, 0.1%) Consultant (n=13, 1.3%) PhD (n=38, 3.8%) Intern/student (n=20, 2.0%) NaN (n=156, 15.6%)
If "Employee type" is "part-time": Percent work	Mean = 77.5, SD = 13.8
People manager	Yes (n=159, 15.9%) No (n=713, 71.2%) NaN (n=130, 12.9%)
If "Yes" to "People manager": how many persons	1-5 (n= 27, 2.7%) 6-10 (33, 3.3%) 11-20 (n=31, 3.1%) 21-50 (n=22, 2.2%)

	51-100 (n=6, 0.6%) >100 (n=8, 0.8%) NaN (n=873, 87.1%)
PSQI	Mean = 4.9, SD = 2.6 Good sleep (<5) (n= 439, 43.8%) Poor sleep (>=5) (n= 402, 40.1%) NaN (n=161, 16.1%)
DASS - Depression	Mean = 2.6, SD = 3.1 Normal (0-4) (n= 692, 69.1%) Mild (5-6) (n= 80, 8.0%) Moderate (7-10) (n= 75, 7.5%) Severe (11-13) (n= 17, 1.7%) Extremely severe (>=14) (n=7, 0.7%) NaN (131, 13.1%)
DASS - Anxiety	Mean = 2.0, SD = 2.6 Normal (0-3) (n= 706, 70.5%) Mild (4-5) (n= 87, 8.7%) Moderate (6-7) (n= 35, 3.5%) Severe (8-9) (n=21, 2.1%) Extremely severe (>=10) (n=20, 2.0%) NaN (n=133, 13.3%)
DASS - Stress	Mean = 5.1, SD = 3.8 Normal (0-7) (n=661, 66.0%) Mild (8-9) (n=97, 9.7%) Moderate (10-12) (n=70, 7.0%) Severe (13-16) (n= 38, 3.8%) Extremely severe (>=17) (n= 7, 0.7%) NaN (n=129, 12.9%)
PSS	Mean = 14.4, SD = 6.1

	Very low (0-7) (n=121, 12.1%) Low (8-11) (n=171, 17.1%) Average (12-15) (n=224, 22.4%) High (16-20) (n=213, 21.3%) Very high (>=21) (n=146, 14.6%) NaN (n=127, 12.7%)
RAND-36 - physical functioning	Mean = 89.6, SD = 15.8 Results for baseline of the Medical Outcomes Study [225] (n=2471): 70.6 ± 27.4
RAND-36 - bodily pain	Mean = 85.7, SD = 16.1 Results for baseline of the Medical Outcomes Study [225] (n=2471): 70.8 ± 25.5
RAND-36 - role limitations due to physical health problems	Mean = 85.2, SD = 28.9 Results for baseline of the Medical Outcomes Study [225] (n=2471): 53.0 ± 40.8
RAND-36 - role limitations due to personal or emotional problems	Mean = 80.5, SD = 32.5 Results for baseline of the Medical Outcomes Study [225] (n=2471): 65.8 ± 40.7
RAND-36 - emotional well-being	Mean = 72.4, SD = 15.3 Results for baseline of the Medical Outcomes Study [225] (n=2471): 70.4 ± 22.0
RAND-36 - social functioning	Mean = 84.6, SD = 19.0

	Results for baseline of the Medical Outcomes Study [225] (n=2471): 78.8 ± 25.4
RAND-36 - energy/fatigue	Mean = 60.5, SD = 19.2 Results for baseline of the Medical Outcomes Study [225] (n=2471): 52.2 ± 22.4
RAND-36 - general health perceptions	Mean = 67.9, SD = 16.4 Results for baseline of the Medical Outcomes Study [225] (n=2471): 57.0 ± 21.11

Appendix B

Comparison of feature dynamic range in low and high classification performance groups.

Subjects were subdivided based on their classification performance (i.e. F1-score) into low, medium and high performance groups. In the main text the differences in dynamic range were shown (i.e. the average difference on physiological features between periods with low (S1) and high (S3) self-reported stress levels) between these performance groups, for ECG mean HR, SC phasic and ST median. In Fig. S2 these differences were shown for the remaining features. For 15 out of 18 features the dynamic range of the high performance group was larger than that of the low performance group, only for ECG SDNN, ECG LFHF and ST SD this was the opposite. Additionally, for 10 out of 18 features (ECG LF, ECG HF, SC mean, SC phasic, SC RR, SC diff2, SC R, SC dur, ST mean, ST median) the dynamic range of the high performance group was larger than that of the medium performance group and the dynamic range of the medium performance group was larger than that of the low performance group, meaning the dynamic range decreased monotonically from high to low performance group.

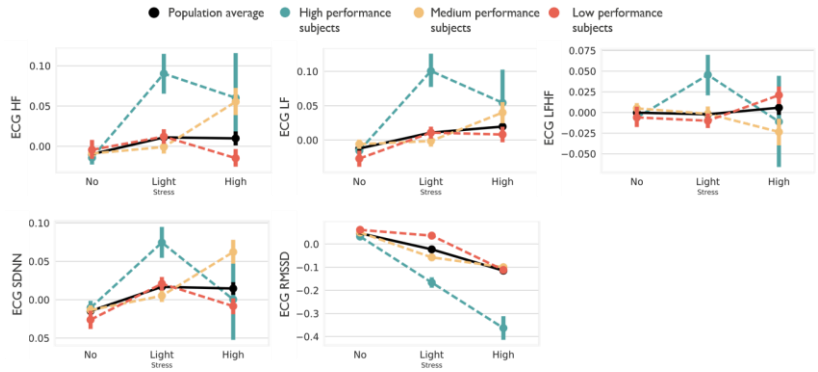
To assess if this trend could be due to chance, we subdivided subjects randomly, instead of based on their F1-score, and investigated the dynamic range for each group. For each feature we calculated the dynamic range per group (i.e. absolute value of (mean(high_stress) – mean(low_stress))). First, only two groups were compared and the number of features for which the dynamic range of one group was larger than for the other group was calculated. Then, all three groups were compared and the number of features for which there was a monotonic increase/decrease in dynamic range across the three groups was calculated. This analysis was repeated 100 times.

Based on a random division of subjects into two groups, on average 12 out of 18 features were found for which the dynamic range of one group was larger than the other group. Based on 100 repetitions, the 95% confidence interval of number of features for which the dynamic range of one group was larger than the other group was [11-16]. When dividing subjects based on their F1-score, 15 features were found for which the dynamic range of the high performance group was larger than of the low performance group. This falls

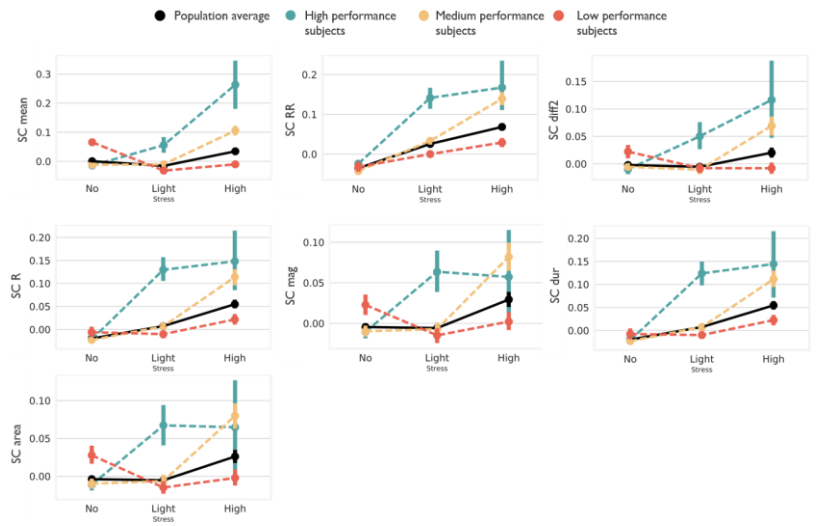
within the 95% confidence interval, suggesting that this result could be found by chance. However, we also found that on average 5 out of 18 features showed a monotonic increase/decrease in dynamic range with a 95% confidence interval of [2-9] features. When dividing subjects based on their F1-score, 10 features were found for which there was a monotonic decrease in dynamic range across the three groups (from high to low performance groups), indicating that this finding was unlikely to be due to chance.

Appendix C

A



B



C

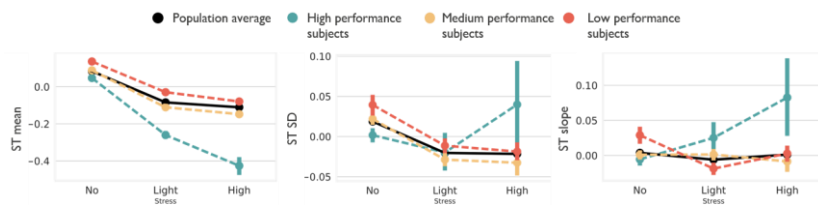


Figure A-1: Comparison of dynamic ranges per feature across low, medium and high classification performance groups.
In A, B and C average ECG, SC and ST related features are shown respectively for low (red), medium (yellow) and high performance (green) groups and compared with the entire population average in phases of no, light and high stress. The high performance group has a larger dynamic range, i.e. larger difference between physiology in no stress and high stress situations, for 16 out of 18 features as compared to the low performance group.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
MICAS DIVISION
Kasteelpark Arenberg 10
B-3001 HEVERLEE, BELGIUM
Elena.smets@imec.be