Wearable-based Affect Recognition

DISSERTATION

zur Erlangung des Grades eines Doktors der Naturwissenschaften

vorgelegt von

M.Sc. Philip Schmidt

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät

der Universität Siegen

Siegen 2019

Betreuer und erster Gutachter

Prof. Dr. Kristof Van Laerhoven

Universität Siegen

Zweiter Gutachter

Prof. Dr.-Ing. Klaus David

Universität Kassel

Tag der mündlichen Prüfung

7 November, 2019

Wearable-based Affect Recognition: Advances in wearable-based sensor technology like smartphones and watches allow to monitor users in a minimally intrusive way. At the point of writing, wearables are, for instance, used to count steps or estimate burned calories. Recently, a first generation of smartwatches entered the consumer market offering data driven insights into affective states.
Over the course of this thesis physiological and motion data recorded using wearables have been employed to detect the affective state (e.g., stress or amusement) of users. The contributions made are threefold: First, a comprehensive literature review of the state-of-the art in wearable-based affect recognition was conducted. Second, concluding from this review a lack of publicly available multimodal datasets was identified. This gap was closed by recording, benchmarking, and publishing a lab study dataset for WEarable Stress and Affect Detection (WESAD). Third, a field study was conducted recording physiological and motion data as well as affective labels from 11 healthy subjects. Prior to the field study guidelines for smartphone-based labelling apps were formulated and they were evaluated using the field study data. Furthermore, data and labels acquired during the field study were used to train both feature-based and latest end-to-end trainable machine learning classifiers, detecting affective states on different scales. Both types of classifiers performed on par (averaged F_1 score across scales: $\sim$ 45%). Hence, potential pitfalls for wearable-based affect recognition were discussed in detail and implications for further research were provided.

Emotionserkennung basierend auf tragbarer Sensorik: Der technologische Fortschritt im Bereich tragbarer Sensorik ermöglicht die minimalinvasive Erfassung von Nutzerdaten wie beispielsweise zurückgelegte Schritte oder verbrauchte Kalorien. Zudem ist die neueste Generation Smartwatches in der Lage, basierend auf physiologischen Daten affektive Zustände der Nutzer zu schätzen.
Im Rahmen dieser Arbeit wurden physiologische und Bewegungsdaten mithilfe tragbarer Sensorik aufgenommen. Diese wurden zur Erkennung affektiver Zustände (z.B. Stress) verwendet. Die geleisteten Beiträge sind die folgenden: Der aktuelle Stand der Forschung im Bereich der Emotionserkennung basierend auf tragbarer Sensorik wurde umfassend dargestellt. Dabei zeigte sich ein Mangel an öffentlich verfügbaren und multimodalen Datensätzen. Diese Lücke wurde durch die Aufzeichnung und Publikation eines Laborstudiendatensatzes (WESAD) geschlossen. Des Weiteren wurden im Rahmen einer Feldstudie physiologische, Bewegungs- sowie affektive Daten von 11 Versuchspersonen aufgezeichnet. Im Vorfeld dieser Feldstudie wurden Empfehlungen für Smartphone-basierte Fragebogen-Apps entwickelt und diese wurden anhand der gesammelten Einblicke überprüft. Zudem wurden die Klassifizierungsraten von feature-basierten Klassifikatoren sowie aktuellsten Convolutional Neural Networks untersucht. Dabei zeigte sich, dass die Erkennung affektiver Zustände im Feld auf unterschiedlichen Skalen nur eingeschränkt möglich ist (über Skalen gemittelter F_1 Score: $\sim$ 45%). Daher wurden Fallstricke für Emotionserkennung basierend auf tragbarer Sensorik diskutiert und Auswirkungen für die weitere Forschung dargelegt.

Contents

1 Introduction

1.1 Motivation and Scope

Affective computing is an emerging field, inspired by the vision to improve human machine interaction by building empathic machines. In general, affective computing research can be divided into two directions. The first direction is the synthesis of affective states within avatars or robots. This topic has gained a lot of attention in recent Science-Fiction movies like "A.I." (Spielberg [2001]), "I, robot" (Sietz [2004]), and "Ex Machina" (Garland [2015]). However, there is still a long way to go before humanoid robots, like the ones portrayed in these movies, can be realized. Nevertheless, a first generation of semi-humanoid robots has been developed, recognizing and reacting to human emotions, see for instance *Pepper* by SoftBankRobotics. The second direction of affective computing is affect recognition, aiming to detect affective states, like emotions or stress, based on observables. The motivation for this research direction originates from different areas of application:
First, considering a **cybernetic point of view**, human decision making is strongly linked to the affective state. Hence, in order to build a holistic user model the affective state of the user is key. As a result, in human machine interaction the machine could adapt its behaviour to the current state of the user. This could find application in industry 4.0 applications, where a machine, for instance, reduces its throughput once it "recognizes" that the workload is too high for the operator. Considering an affect-aware (autonomous) vehicle the affective state of the driver/passengers could be used to provide services, ranging from re-routing options (e.g., when the car detects that the driver/passengers is/are stressed) to "coffee break stops" (e.g., when driver/passenger fatigue is detected). The latter case has, for instance, been addressed by a driver drowsiness detection system developed by Bosch [2019]. Second, considering affect recognition from the viewpoint of the **"quantified self" movement**, the affective state is an interesting property. Assessing affective states in a continuous and data driven way, could increase users' awareness of their affective

states. In addition, such a system could help to correlate certain affective states with locations or events and thus help users to avoid a subset of stressful situations. The smartphone app Daylio [2019], where users track activities and moods manually, promises this kind of insight. Third, considering **psychological care**, automated affect recognition could aid diagnostics and treatment. Grünerbl et al. [2015], for instance, presented an approach to detect state changes in bipolar disorder patients. Such a system could automatically schedule an appointment with a psychiatrist once a certain state, e.g., manic-episode is detected. Fourth, from a **health care** point of view, continuous and automated stress detection is a particularly interesting application of affect recognition. This is due to the severe side effects of long-term stress, which range from headaches and troubled sleeping to an increased risk of cardiovascular diseases, see McEwen and Stellar [1993], Chrousos and Gold [1992], Rosmond and Björntorp [1998]. One approach for assessing stress in mobile environments was, for instance, presented by Hovsepian et al. [2015].

Depending on the setting, a number of affect recognition systems relying on different input modalities are available: In an automotive context, for instance, Eyeris [2019] and Vayyar [2019] utilize video data and vital signs to detect the driver's state. Furthermore, companies like Beyondverbal [2019] or Vokaturi [2019] offer audio-based emotion recognition. Considering the findings of Tzirakis et al. [2017] and Mirsamadi et al. [2017], emotions are detected reliably using audio and/or video data. In addition, stress detection based on audio samples is feasible, as presented by Lu et al. [2012]. The high performance of these systems is strongly linked to recent advances in the computer vision and audio analysis domain, where the advent of convolutional and long short-term memory neural networks led to breakthroughs. Depending on the application, audio and/or video might be valid modalities (e.g., in callcenter applications or for human machine interaction (HMI)). However, these modalities exhibit two crucial limitations: First, recording audio and/or video data continuously is intrusive in terms of privacy. Second, from a technical point of view, continuous recording is difficult. Consequently, these modalities are only available in specific settings and circumstances (e.g., HMI or vehicles). However, the aim of affect recognition is to detect affective states continuously and in unconstrained environments, e.g., in the everyday life of the subjects. As a result, audio and video data are inappropriate modalities for most long-term affect recognition and monitoring scenarios.

Recent advances in wearable-based sensor technology and computing facilitate new opportunities in the domain of ubiquitous computing and human monitoring. Up-to-date smartphones or watches are used by many to count steps, assess sleep quality, or monitor physiological parameters, like cardiac activity. One advantage of these devices, smartwatches in particular, is that they facilitate long-term monitoring of physiological parameters (e.g., cardiac activity) while being only minimally intrusive. Some affective states, like stress, exhibit a distinct influence on certain physiological parameters and lately a first generation of smartwatches has

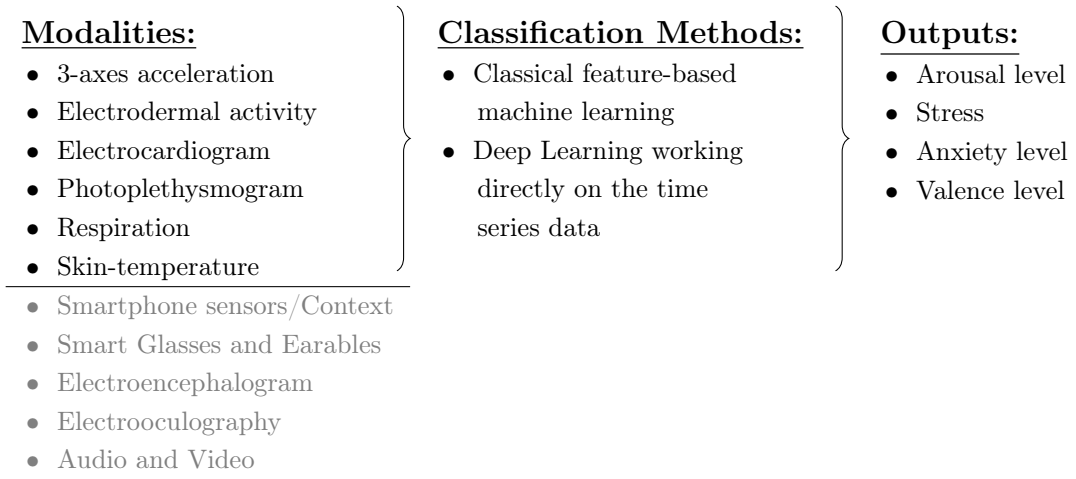


Figure 1.1: Schematic representation of the scope of the presented thesis. Modalities not considered are indicated in gray.

been launched, promising insights into personal stress levels, see Garmin [2019] or Apple [2019]. Inspired by these devices and the link between affective states and physiology, the performance of wearable-based affect recognition systems using physiological data, like cardiac and electrodermal activity, and motion information is explored in this thesis. However, this work not only aims to detect stress, but also aspires to investigate the properties of systems detecting additional affective states (like amusement or other points in valence-arousal space). For this purpose both classical (feature-based) machine learning and deep learning methods are applied. Considering the minimally intrusive nature and the broad acceptance of smartwatches, the focus of this thesis lies on data collected using this type of device. Smart fabrics and garments are on the verge of being the next big thing in the wearable domain. Hence, modalities potentially integrated in smart clothes recording physiological data from the torso of subjects are also explored. The above detailed considerations boil down to two main inclusion criteria: First, the wearables considered in this work have to be worn *directly on the body*, which facilitates the recording of physiological and motion data. Second, they should be only *minimally intrusive* and *facilitate long-term monitoring*. Figure 1.1 displays a schematic overview of the scope of this thesis outlining the input modalities, the employed machine learning systems, and the targeted affective states. Due to the inclusion criteria formulated above the following modalities are *not* in scope of this work:

- *Smartphones:* Smartphones are very popular among users. They are equipped with numerous sensor modalities, which can be used to generate contextual information (e.g., location logging). However, smartphones are not necessarily worn *directly* on the subject's body. This is best illustrated considering the

following example: As smartphone screens become larger and larger they tend not to fit into trouser pockets any more. Consequently, phones are often placed somewhere (e.g., on the table) and not kept *directly* on the body. Due to this, the data (e.g., activity, or location) do not always represent the actual activity. Hence, smartphone-based sensory data (e.g., location) are excluded from the considered inputs.

- *Smart Glasses and Earables*: Academic research has shown that smart glasses and earables are able to measure physiological parameters like skin-temperature or heart rate, see for instance Yasufuku et al. [2016] or Budidha and Kyriacou [2014]. However, at the moment of writing, smart glasses and earables have not experienced a breakthrough on the consumer market. Consequently, these potential modalities are not considered here.
- *Electroencephalogram and Electrooculography*: Both EEG and EOG require the placement of electrodes on the scalp/face of the user. In addition, despite being popular among researchers, EEG and EOG are not available on the consumer market. Hence, these devices are not very frequently employed by users. Reflecting on their intrusive nature and small market share, EEG and EOG will not be considered.
- *Audio and Video:* In constrained settings, e.g., a living room or a car interior constant video and audio recordings of subjects might be possible. However, outside this constrained setting currently no ubiquitous audio/video recording infrastructure is available. Due to the strong privacy concerns of ubiquitous audio and video surveillance it seems unlikely that these modalities will be available. Hence, long-term monitoring based on audio and video data in unconstrained environments is not possible. As a result, audio and video are modalities not considered in the scope of this work.

1.2 Research Questions

The aim of this thesis is to investigate the performance of purely wearable-based affect recognition (AR) systems. The employed algorithmic approaches are based on multi-modal data, focusing on physiological indicators and motion patterns. Using these modalities, the aim is to recognize and distinguish affective states like stress, anxiety, or other points in valence-arousal space. For this purpose, data is collected using wearables and different AR systems are trained based on this data (or the derived features). The presented thesis addresses the following research questions:

RQ 1 What is the current state-of-the-art in wearable-based affect recognition?

Since the term affective computing has been coined by Picard [1995], a lot of research has been conducted. However, the community currently lacks a comprehensive literature review focusing on wearable-based AR. To target this issue a detailed analysis of the state-of-the-art is performed (see Chapter 2) and related work is examined carefully. The aim of this review is to provide an introduction into psychological constructs targeted in AR. In addition, the commonly employed classification chain is presented, focusing on the preprocessing, feature extraction, and classification steps. This review can be used by other researchers to obtain a full overview of the methodology employed in wearable-based AR. This review was published in:

> Schmidt, P., Reiss, A., Dürichen, R., Van Laerhoven, K., Wearable-Based Affect Recognition - A Review, In Sensors 2019, 19(19), 4079; DOI: 10.3390/s19194079, https://doi.org/10.3390/s19194079

RQ 2 How is benchmarking and direct comparison of different algorithmic approaches for wearable-based affect recognition feasible?

One key finding of the literature review is that wearable-based AR is a well investigated research topic. However, the comparison of different algorithmic approaches is difficult as the community is missing a common benchmarking dataset. Such a benchmarking dataset should meet the following criteria:

I. The data should be acquired using high quality wearables only. In order to perform a high resolution spectral analysis, the employed sensors should be sampled at high frequencies. Considering smartwatches and smart fabrics, the data should be acquired in a redundant fashion from different locations (e.g., chest and wrist).
II. For the purpose of creating a benchmarking dataset, acquiring data in a laboratory setting is absolutely satisfying. However, the subjects should

have some freedom with regards to their posture and movements. This is motivated by the fact that in real world applications, strong motion artefacts are to be expected. Hence, in order to be more realistic the dataset should not be completely free of motion artefacts.

III. Each study participant should experience multiple affective states. The different affective states should be elicited using appropriate stimuli and be reproducible. A binary stress/no-stress dataset would not be sufficient, due the strong physiological stress response.

IV. In order for the results to be of statistical relevance, the dataset should contain physiological data and self-reports from at least 10 subjects.

V. A benchmark, should be created employing a standard set of features and classifiers.

This work aspires to fill the identified gap by introducing WESAD - a dataset for WEarable Stress and Affect Detection. In Chapter 3, the study protocol and the employed modalities are detailed. In addition, WESAD is benchmarked using a number of classical (feature-based) machine learning classifiers. The WESAD dataset and benchmark was introduced in the following conference contribution:

> Schmidt, P., Reiss, A., Dürichen, R., Marberger, C., Van Laerhoven, K., Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, DOI: 10.1145/3242969.3242985, http://doi.acm.org/10.1145/3242969.3242985.

RQ 3 What is the performance of machine learning systems that detect multiple affective states in unconstrained environments?

Wearable-based AR has the potential to facilitate long-term affect detection in everyday life. However, to-date only a few field studies have been conducted, aspiring to detect the affective state of the participants in unconstrained environments using wearable-based physiological and motion data. Most of these field studies, e.g., Healey et al. [2010] or Gjoreski et al. [2017], rely on classical feature-based machine learning methods. However, deep learning methods like convolutional neural networks (CNNs), can be used to learn features automatically. This has the potential to eliminate the need for feature engineering. CNNs have found application in the human activity recognition domain, see for instance Ordóñez and Roggen [2016] or Münzner et al. [2017]. Motivated by these approaches, the performance of automated feature extractors shall be investigated comparing them to classical feature-based machine learning methods.

In order to tackle this, a field study has been conducted, recording physiological time series data, context information, and labels (see Chapter 4). As there is no *standard procedure* for wearable-based AR field studies **RQ 3** is divided into the following two subtasks:

RQ 3a *What is an appropriate way to label affective states in everyday life reliably?*

In lab studies, stimuli can be designed carefully. In unconstrained environments, in contrast, affective states occur naturally. Hence, an appropriate way to label these needs to be developed. In related literature, smartphone apps scheduling questionnaires (so called ecological-momentary-assessments (EMAs)) were employed for this purpose. However, it seems that no guidelines for the development and application of smartphone-based EMA apps are available. Hence, the labelling tool used in the conducted field study and its design are presented in detail. Further, based on the insights gained during the real life data acquisition guidelines and lessons learned, for the development and application of EMA apps (see Section 4.3) are formulated. Parts of this work were presented in:

> Schmidt, P. , Reiss, A., Dürichen, R., Van Laerhoven, K., Labelling Affective States "in the Wild": Practical Guidelines and Lessons Learned. In Adjunct *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, DOI: 10.1145/3267305.3267551, http://doi.acm.org/10.1145/3267305.3267551.

RQ 3b *What is the performance of classifiers trained on labels generated with an ecological-momentary-assessment tool?*

Gjoreski et al. [2017] presented an approach detecting stress of five study participants using physiological and context data in the field. For this purpose, classical machine learning methods relying on features were employed. Furthermore, Taylor et al. [2017] developed a machine learning model, using multilayer perceptrons, to predict tomorrow's mood, health, and stress based on today's data. This model used context, survey information, and physiological indicators as input modalities. Inspired by these approaches, the current affective state of users based only on physiological and motion data shall be detected. Hence, a single-task and multi-class affect detection problem is posed, using the physiological and motion data as well as the labels acquired during the field study. In addition, the performance of CNNs is explored formulating the classifica-

tion of multiple affective states as multi-target and multi-class problem (see Section 4.4). The performance of both the single and the multi-task machine learning systems is surprisingly low. Hence, pitfalls and key challenges for wearable-based AR are discussed thoroughly. These results and discussion can also be found in the following conference paper:

> Schmidt, P., Dürichen, R., Reiss, A., Van Laerhoven, K., Plötz, T., Multi-target Affect Detection in the Wild: An Exploratory Study, In *Proceedings of the 23rd International Symposium on Wearable Computers*, ISWC '19, DOI: 10.1145/3341163.3347741, http://doi.acm.org/10.1145/3341163.3347741.

2

Interdisciplinary Background and Related Work

Wearable-based affect recognition (AR) aspires to detect the affective state of a person based on observables. Hence, from a theoretical point of view, AR can be seen as a signal processing and pattern recognition problem, see D'mello and Kory [2015]. However, due to the concepts (e.g., stress, emotions) targeted by AR and the used signals, wearable-based AR is a highly interdisciplinary research field with links to signal processing, machine learning, psychology, and behavioural neuroscience.

In this chapter, the interdisciplinary background and related work is summarized. First, in Section 2.1, the terminology frequently used in AR will be defined and psychological models for emotions and stress will be presented. Next, Section 2.2 details the influence of affective states on physiological parameters. In addition, sensory setups measuring these changes are listed. Protocols eliciting emotions in a laboratory setting and information on methods and questionnaires commonly employed in AR field studies are detailed in Section 2.3. Furthermore, publicly available datasets are listed and described. At the end of this chapter (see Section 2.4), the classical data processing pipeline of time series data is reviewed. Special attention is given to feature extraction based on physiological time series data and the state-of-the-art performance of wearable-based affect recognition systems is presented.

The content of this chapter has been published as Schmidt et al. [2019b].

2.1 Interdisciplinary Background

In this section an overview of the terminology used in affect recognition (AR) will be provided. For this purpose different psychological and physiological constructs of affective states will be presented and summarized.

2.1.1 Working Definitions of Affective Phenomena

In order to tackle AR, working definitions of different affective states are required. Psychologists have been studying human emotions intensively. Hence, the emotional models and terms employed in AR are "borrowed" from psychology. In this section terms commonly used in AR are defined and models for emotions and stress are introduced.

Despite a growing body of research, it is still difficult to define the terms affect, emotion, and mood in a precise way. Below working definitions are provided and differences between the constructs are highlighted. Russell [2003] defines **affect** to be a neurophysiological state. This neurophysiological state is consciously accessible as simple raw (nonreflective) primitive feeling Liu [2017]. Affect is not directed at a specific event or object and lasts only for a very short time. In contrast, **emotions** are intense and directed feelings, which have a short duration. Emotions are an indicator of affect, and arise from a cognitive process evaluating a stimulus (e.g., a specific object, an affect, or a thought). Hence, emotions are directed at a stimulus. To illustrate these aspects, Liu [2017] uses the example of watching a scary movie: If you are affected, the movie elicits the feeling of being scared. The mind processes this feeling (*scared*), adds an evaluation (*'this is really spooky'*), and expresses it to you and your surroundings as an emotion (*fear*) by, e.g., crying Liu [2017]. In AR literature, the terms mood and emotion are often used interchangeably. However, in contrast to emotions (and affects), **mood** is commonly defined to be less intense, more diffuse, and to last for a longer time period. This difference between mood and emotion is best illustrated by considering the following example: One can get angry very quickly, but it is hard to stay angry for a longer time period. However, the emotion *anger* might lead to an *irritable* mood, which can last for a long time Liu [2017].

In the remainder of this thesis the term **affective state** will be used to describe the internal state of a person, which can be referred to as emotion, mood, and/or affect.

2.1.2 Emotion Models

In this section, emotional models frequently employed in AR literature are detailed. These are grouped into two distinct types:

1. **Categorical models:** Here different emotions are represented in discrete categories.
2. **Dimensional models:** Following this approach, emotions are mapped into a multidimensional space, where each of the axes represents a continuous variable.

Categorical models date back to ancient Greek and Roman philosophers Poria et al. [2017]. Cicero, for instance, distinguished four basic categories of emotions, namely *fear, pain, lust*, and *pleasure*. Darwin also conducted studies on emotions, and came to the conclusion that emotions have an evolutionary history and, hence, are shared across cultures. Similar to Darwin, Ekman [1992] argues that basic emotions are shared across cultures and appear to be universally recognised. Following Ekman and Friesen, six basic emotions can be distinguished: *Joy, Sadness, Anger, Fear, Disgust,* and *Surprise* Ekman and Friesen [1978, 1976]. These basic emotions are discrete and have distinct physiological patterns, e.g., facial muscle movement. Being able to express basic emotions can be attributed with a number of (evolutionary evolved) physiological and communicative functions: *Disgust*, for example, is often expressed by a certain facial expression and a wrinkled nose. On a physiological level this facial expression limits inhalation of malodorous particles. On the communicative level, this distinct facial expression, performed for instance as reaction to rotten food, has the potential to warn others.

In 1980, Plutchik [1980] developed another taxonomy to classify discrete emotions. The so-called 'wheel of emotions' comprises of eight primary emotions: *Grief, Amazement, Terror, Admiration, Ecstasy, Vigilance, Rage,* and *Loathing*. Following Plutchik [1980], the primary emotions mix, and give rise to more complex emotions. In addition, emotions are expressed at different intensity levels. In the domain of wearable-base AR, categorical models were for instance used by Zenonos et al. [2016]. In their study the authors presented an approach to distinguish eight different emotions and moods (*excited, happy, calm, tired, bored, sad, stressed* and *angry*).

The above presented model of basic emotions is not unquestioned and one point of criticism is that some languages do not have words for certain basic emotions Russell [1979]. According to Soleymani et al. [2012a] in polish, for instance, there is no exact translation for the English word *disgust*.

Dimensional models where emotions are mapped into a multidimensional space, mitigate this shortcoming. The first dimensional approach dates back to Wundt [1863], who describes momentary emotions as a single point in a three-dimensional space Becker-Asano [2008]. Wundt's emotional space is spanned by the pleasure-displeasure, excitement-inhibition, and tension-relaxation axes. At the end of the 1970s, Russell [1979] postulated a two-dimensional model, namely the circumplex model (see Figure 2.1). This model has been very impactful and in the circumplex model, affective states are represented as discrete points in a two-dimensional space,

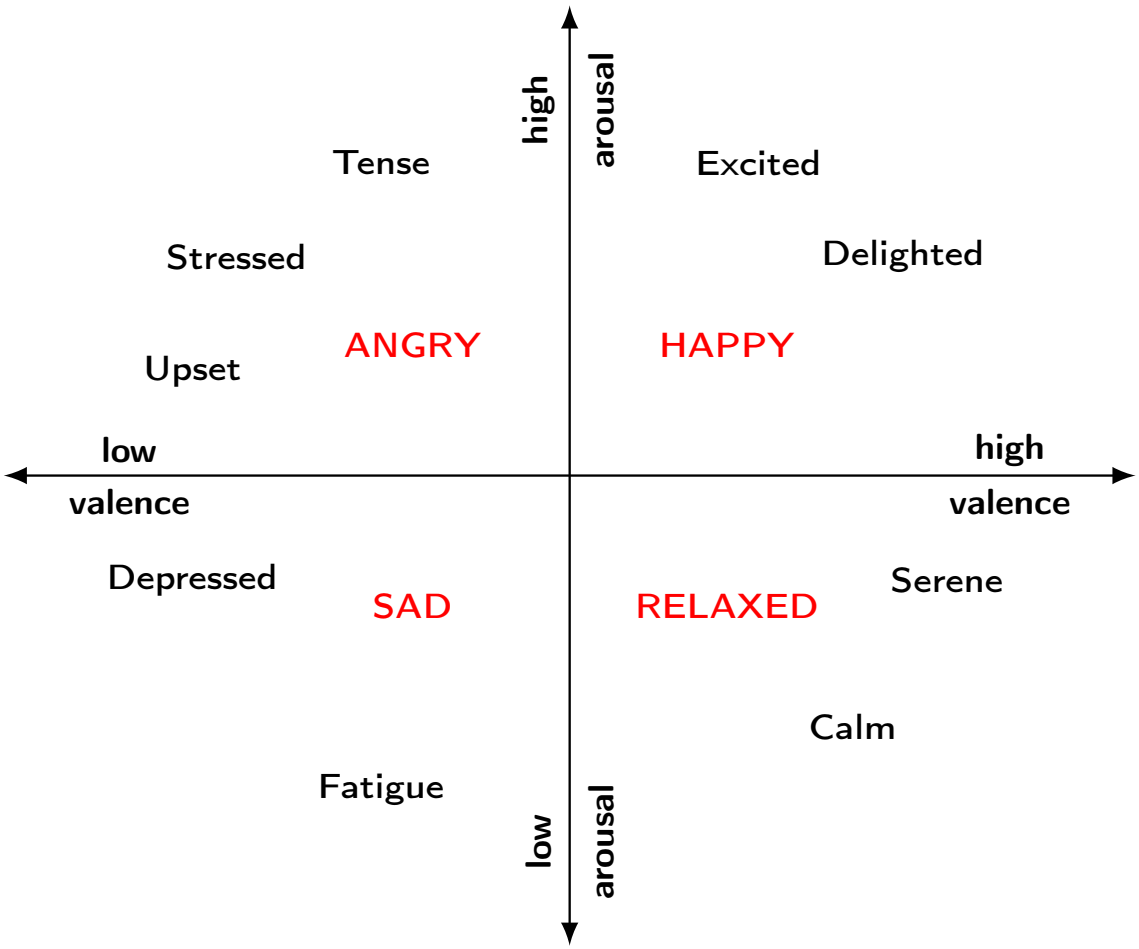


Figure 2.1: Schematic representation of the circumplex (valence-arousal) model. Adapted from Valenza et al. [2014].

spanned by the axes valence and arousal. The valence axis indicates the perception on how positive or negative the current affective state is. On the arousal axis, the state is rated in terms of the activation level, e.g., how energised or enervated one feels. The four quadrants of the circumplex model (low arousal/low valence (LALV), low arousal/high valence (LAHV), high arousal/low valence (HALV) and high arousal/high valence (HAHV)) can be attributed with *sad, relaxed, angry,* and *happy*. By adding further orthogonal axes, e.g., dominance, the circumplex model is easily extended. In AR, the circumplex model and its variants are frequently employed, see Kim and André [2008], Koelstra et al. [2012], Valenza et al. [2014], Abadi et al. [2015]. Using the Self-Assessment Manikins (SAM) of Morris [1995], the circumplex model can easily be assessed. These Manikins offer an easy graphical way for subjects to report their current affective states (see Figure 2.2). In addition, the SAM are easily understood across cultures, due to their simple graphical representa-

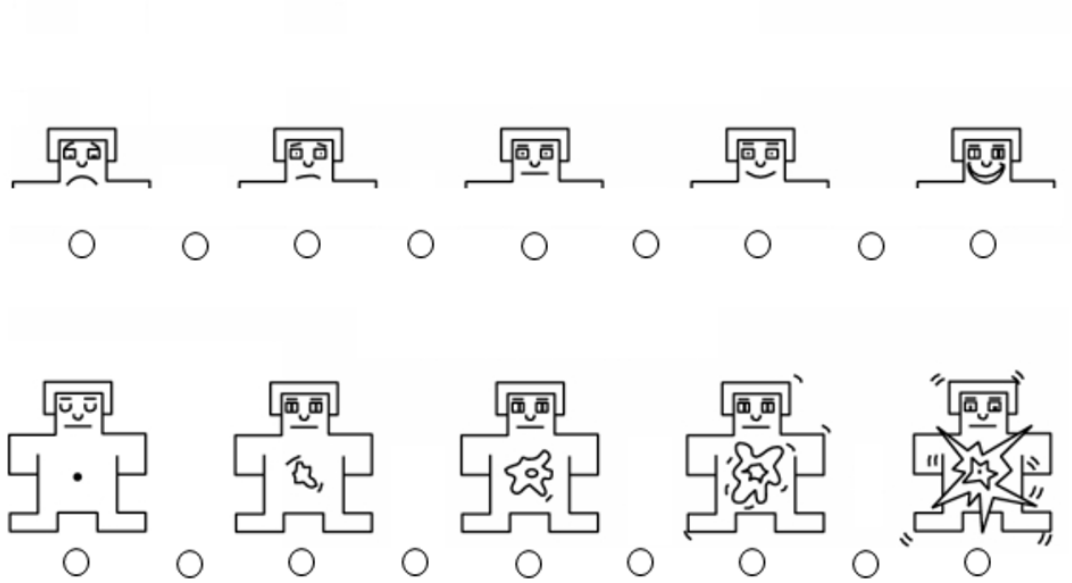


Figure 2.2: Exemplary Self-Assessment Manikins Morris [1995], used to generate labels in the valence-arousal space. Adapted from Jirayucharoensak et al. [2014].

tion. Another possible reason for the popularity of dimensional models might arise from a machine learning (ML) point of view. The (at least two) independent axes of the circumplex model offer an interesting set of different classification tasks: The valence and arousal axes, for instance, can be binned into multiclass classification problems, e.g., low/medium/high arousal or valence. In addition, posing classification problems based on the four quadrants named above is a frequently pursued task in AR, see for instance Kim and André [2008], Subramanian et al. [2017].

2.1.3 Stress Models

In everyday life, *stress* or *being stressed* are terms used to describe the feeling of being under pressure. Stress is commonly elicited by an external and/or internal stimulus called stressor. However, from a scientific point of view, stress is primarily a physiological response. At the beginning of the 20th century Cannon [1929] coined the terms homeostasis and "fight or flight" response. Homeostasis describes a balanced state of the organism where its physiological parameters stay within an acceptable range (e.g., a body temperature of $37\,C^{\circ}$). Following Cannon [1929], both physiological and psychological stimuli can pose threats to homeostasis. Stressors can be seen as threats, disrupting homeostasis. In order to maintain homeostasis,

even under extreme conditions, feedback loops (e.g., a fight or flight response) are triggered, Cannon [1929].

Selye [1974] defined stress to be or result in a 'nonspecific response of the body to any demand upon it'. Following this definition, 'nonspecific' refers to a shared set of responses triggered regardless of the nature of the stressor, e.g., physical or psychological. Recent stress models, e.g., McEwen and Stellar [1993], incorporate multiple effectors and advocate that the stress response is to some degree specific. The stress response is mainly influenced by two aspects: First, the stressor itself and, second, the organism's perceived ability to cope with the posed threat Goldstein and Kopin [2007]. Depending on the coping ability of the organism and estimated chances for success, **eustress** (positive outcome) and **distress** (negative outcome) are distinguished Lu et al. [2012]. Eustress can have a positive (motivating) effect, while distress is perceived to be hindering (feeling worried or anxious). In order to illustrate this the following example can be used: Assume a person has to take an exam. Here, this exam represents an external stressor and the body reacts with a physiological stress response, e.g., by increasing the blood glucose level. If the person feels well prepared for the exam and is looking forward to the challenge ahead, this can be interpreted as eustress. In contrast, if the person is not well prepared and feels like failing the exam, this can result in distress. Considering wearable stress recognition, distinguishing between eustress and distress is a largely unsolved problem due to the lack of adequate physiological indicators. However, long-term stress in general is associated with many severe health implications ranging from troubled sleeping and headaches to an increased risk for cardiovascular diseases, see McEwen and Stellar [1993], Chrousos and Gold [1992], Rosmond and Björntorp [1998]. Due to these severe side effects of long-term stress, the detection of stress is a frequent task in AR: Mozos et al. [2017], Plarre et al. [2011], for instance target binary stress recognition tasks (*stress* vs. *no stress*) and Gjoreski et al. aimed at distinguishing different levels of stress (*no stress* vs. *low stress* vs. *high stress*).

Above different emotion and stress models were summarised. Although stress is not an emotion, a link between dimensional models and stress is readily established: Following Sanches et al. [2010], a direct link between stress and arousal can be drawn. Valenza et al. [2014] maps stress into the high arousal/negative valence (quadrant II) of the circumplex model (see Figure 2.1). Following Thayer [1990] and later Schimmack and Reisenzein [2002], the arousal dimension of the 'classical circumplex' model can be split into tense arousal (stressed-relaxed) and energetic arousal (sleepy-active). According to Schimmack and Reisenzein [2002], this split is justified by the observation that only the energetic arousal component is influenced by the sleep-wake cycle. Considering the wearable affect and stress recognition literature, a recent study conducted by Mehrotra et al. [2017] uses this three-dimensional emotion model (valence, tense arousal, and energetic arousal) to investigate correlation and causation between emotional states and cell phone interaction.

2.2 Physiological Changes and Objective Measures

In this section the affect-related changes in physiology and devices to measure these are presented. Section 2.2.1 provides background on the physiological changes and in section Section 2.2.2 commonly used sensors are presented.

2.2.1 Affective States and their Physiological Indicators

Affective states and physiological changes are clearly linked, e.g., if someone cracks a good joke we laugh or at least smile. With this physiological response we express *amusement*. Negative emotional states have even stronger physiological indicators. For instance, when being *afraid* or *anxious* one might start sweating, get a dry mouth, or feel sick.

Stress was characterised primarily as a physiological response to a stimulus, see Section 2.1.3. The most severe physiological reaction to a stressor is the so called 'fight or flight' response Cannon [1929]. During this response the body prepares for a severe action, like fight or flight, releasing a mixture of hormones, like cortisol and adrenaline. This leads, for instance, to an increased breathing/heart rate, pupil dilation, and muscle tension. The induced physiological responses are quite distinct and are a good example for the link between affective states and physiological changes.

Above the link between affective states and physiological responses was established using examples. The direction/causality, e.g., do affective states cause physiological changes or vice versa, is still an open research question: At the end of the 19th century James [1884] postulated, that physiological changes precede emotions and that emotions arise from these changes. This is best illustrated considering the following example: Picture someone encountering a gigantic poisonous spider. Following this encounter the heart rate and the activity of the sweat glands of the subject would increase. Following this **James-Lange-Theory**, these physiological changes are not symptoms of *fear/disgust*, but rather involuntary physiological responses. According to James [1884] these physiological responses, become an emotion/feeling, like *fear/disgust*, once a cognitive evaluation occurred. Hence, the subject could describe the process as "I feel afraid, because I have a racing heart". This theory is supported, for instance, by experiments conducted by Levenson et al. [1990], who found evidence that performing voluntary facial muscle movements exhibit similar changes in peripheral physiology as if the corresponding emotion is experienced. For instance, when the subjects were asked to make an angry face the heart rate was found to increase. This theory, is not unchallenged. Following **common sense**, a stimulus is perceived, it elicits an feeling, and then the physiological responses are triggered. Hence, the subject could describe the process as "I have a racing heart, because I'm afraid of the poisonous spider". Following the **Cannon-Bard-Theory**, the perceived stimulus is processed in the brain and the physiological response and affective states arise simultaneously Friedman [2010]. Hence, the subject could de-

Table 2.1: Major functions of the sympathetic nervous system and parasympathetic nervous system.

Sympathetic nervous system (SNS)	**Parasympathetic nervous system (PNS)**
• associated with 'fight or flight'	• associated with 'rest and digest'
• pupils dilate	• pupils constrict
• decreased salivation and digestion	• increased salivation and digestion
• increased heart and respiration rate	• decreased heart and respiration rate
• increased electrodermal activity	
• increased muscle activity	
• adrenalin and glucose release	

scribe the process as "The spider makes me feel afraid and I have a racing heart". The debate outlined above is, from a theoretical point of view, very interesting. However, it is out of scope of this thesis. Wearable-based AR utilizes these affect-related changes in physiology.

Affective states occur spontaneously and are accompanied by certain physiological pattern. These physiological responses are hard or even impossible to control for humans. The **autonomic nervous system (ANS)** directs these unconscious actions of the organism. Hence, the ANS plays a key role in directing the physiological response to an external (e.g., event) or internal (e.g., thought) affective stimulus. The ANS has two major branches: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). In Table 2.1, the key contributions of the SNS and PNS are displayed. As the SNS is mainly associated with the 'fight or flight' response, an increased activity of the SNS indicates high arousal states. In other words, the main function of the SNS is to provide energy by increasing a number of physiological parameters (e.g., respiration rate, glucose level, etc.). The PNS, in contrast, regulates the 'rest and digest' functions McCorry [2007].

The interplay of SNS and PNS is best illustrated considering the cardiovascular system. In reaction to a potential threat, the SNS increases the heart rate (HR). Once the threat is over, the PNS reduces the HR, bringing it back to normal Choi et al. [2012]. A common measure to quantify the interaction of SNS and PNS is the **heart rate variability (HRV)**. The HRV is defined as the variation in the beat-to-beat intervals. An increased/decreased HRV indicates increased activity of the PNS/SNS, respectively. As a result, the HRV is a rather simple but efficient measure to quantify the contributions of the PNS/SNS. Hence in related work, the HRV is employed to detect stress Choi et al. [2012]. Changes in the **electrodermal activity (EDA)** are another simple but effective measure to assess the SNS activity, too. This is due to the fact, that changes in EDA are governed by the SNS Choi et al. [2012]. Hence, following Dawson et al. [2000] the EDA is particularly sensitive to high arousal states, like *fear, anger,* and *stress*. EDA has two main components, namely the skin conductance level (SCL) and the skin conductance response (SCR). The SCL, also known as tonic component, represents a slowly varying baseline conductivity. In contrast, the SCR, also called phasic component, refers to peaks in the

Table 2.2: Four exemplary affective states and their physiological response Kreibig [2010]. Abbreviations: ↓ indicate a decrease, ↑ indicates an increase, ↑↓ indicate both increase and decrease (depending on the study), − indicates no change in the parameter under consideration, # number of.

	Anger	**Sadness** (non-crying)	**Amusement**	**Happiness**
Cardiovascular:				
Heart rate (HR)	↑	↓	↑↓	↑
Heart rate variability (HRV)	↓	↓	↑	↓
Electrodermal:				
Skin conductance level (SCL)	↑	↓	↑	↑ −
# Skin conductance responses (SCRs)	↑	↓	↑	↑
Respiration:				
Respiration rate	↑	↑	↑	↑

EDA signal. For most other vital parameters, the contributions of PNS and SNS are more interleaved. Hence, their responses are less specific. Nevertheless, also considering **respiration and muscle activity**, certain patterns can be attributed to different affective states. For instance, the respiration rate increases and becomes more irregular when a subject is more aroused Kim and André [2008]. Later, in Section 2.4, a detailed description of physiological features will be provided.

As outlined above, the SNS contributions to high arousal states are quite distinct. In a recent meta analysis, Kreibig [2010] investigated the **specificity of the ANS** response to certain affective states. A subset of these findings, including two positive and two negative affective states, is presented in Table 2.2. Considering for instance *anger*: A majority of the analysed studies showed that it coincides with an increased HR, SCL, number of SCRs, and a higher breathing rate. Since *anger* represents a high arousal state, governed by the SNS, these reactions were expected. Non-crying *sadness* was found to decrease HR, SCL and number of SCRs, while increasing the respiration rate. In the circumplex model (see Figure 2.1), *sadness* is mapped into the third quadrant (low valence, low arousal). Hence, the arousal level is expected to drop which is confirmed by Table 2.2. *Amusement* and *happiness* are both positive affective states with a similar arousal level. Hence, it is not surprising that they have a similar physiological fingerprint.

The findings of Kreibig [2010] suggest that affective states have certain physiological fingerprints which are to some degree specific. These findings are promising, as they indicate that distinguishing affective states based on physiological indicators is feasible. However, in the context of **wearable-based AR**, the following aspects should be considered Broek et al. [2009]:

1. Physiological measures are *indirect* measures of an affective state.
2. Emotions are subjective, but physiological data are not.

3. Although some physiological patterns are shared across subjects, individual responses to a stimulus can differ strongly.
4. According to D'mello and Kory [2015], multimodal affect detecting systems reach higher accuracies than unimodal systems.
5. The physiological signal quality often suffers from noise, induced by motion artefacts and misplacement.

2.2.2 Frequently Employed Sensors

This section provides an overview of the sensor modalities frequently employed in wearable-based AR. The clear aim of AR is to find robust methods assessing the affective state of a user in everyday life. Hence, as detailed in Section 1.1, a major goal is to use sensor setups which are worn directly on the body of the subjects and are only minimally intrusive, posing only minor limitations to the mobility of the user. As defined in Table 2.1 and Table 2.2, physiological changes in the cardiac system and electrodermal activity are key indicators for affective states. Therefore, most studies utilise these modalities. Nevertheless, sensors measuring other physiological parameter, like respiration or muscle activity, can also contain valuable information on the affective state of a person Kreibig [2010]. Table 2.3 lists the most relevant sensors, grouped according to their placement on the human body. Further, each of the listed modalities is discussed, detailing advantages and limitations.

In order to assess the heart rate (HR), heart rate variability (HRV) and other parameters related to the cardiac cycle, the **electrocardiogram (ECG)** serves as gold standard. For a standard three-point ECG, three electrodes are placed on the subject's torso, measuring the depolarisation and repolarisation of the heart tissue during each heartbeat. ECG samples are collected with frequencies up to 1024 Hz. However, when acquired with such high frequency the signal can be downsampled to 256 Hz without loss of information Soleymani et al. [2012a]. Furthermore, experiments of Mahdiani et al. [2015] indicate that a 50 Hz ECG sampling rate is sufficient to obtain HRV-related parameters with a reasonable error. Using **photoplethysmogram (PPG)** also provides information about the cardiac cycles. In contrast to ECG, PPG utilises an optical method: The skin voxel, beneath the sensor, is illuminated by a LED and a photodiode measures the amount of backscattered light. Alternatively if the detector is on the opposite side of the respective body part (e.g., fingertip or earlobe), the amount of transmitted light is measured. Hence, the cardiac cycle is captured by the PPG signal, where the pulsatile part of the PPG signal reflects the pulsatile component in arterial blood flow Tamura et al. [2014]. Data obtained from a PPG sensor tends to be noisier than ECG data. This is due to artefacts caused by motion, light from external sources, or different skin tones, which influence the reflection/absorption properties of the skin. PPG sensors can be attached to the ear, wrist Gjoreski et al. [2017] or the finger tip Lin et al.

Table 2.3: Sensor modalities and derived indicators used in the wearable-based AR.

	Physiological Signal Type	**Derived Indicators**
Head/Face	Electroencephalogram	Electric potential changes of brain neurons
	Electromyogram	Facial muscle activity (e.g., zygomaticus major)
	Electrooculography	Eye movements
	Photoplethysmogram (ear)	HR and HRV
Torso/Back	Electrocardiogram	HR and HRV
	Electrodermal activity	Tonic and phasic component
	Electromyogram	Muscle activity
	Inertial sensor	Physical activity/body pose
	Respiratory inductive Plethysmograph	Respiration rate and volume
	Body thermometer	Temperature
Hand/Wrist	Electrodermal activity meter	Tonic and phasic component
	Blood Oxymeter	Blood oxygen saturation
	Sphygmomanometer	Blood pressure
	Inertial sensor	Physical activity
	Photoplethysmogram	HR and HRV
	Thermometer	Temperature
Feet/Ankle	Electrodermal activity	Tonic and phasic component
	Inertial sensor	Physical activity
Context	Sensors of a mobile phone (GPS, microphone, etc.)	Location, Sound, Activity, Interaction

[2014] of subjects. The PPG modality finds broad application in fitness trackers and smartwatches, which can be attributed to the small form factor of the sensory setup. Typical sampling rates of PPG devices are below 100 Hz.

The **electrodermal activity (EDA)** is commonly measured at locations with a high density of sweat glands, e.g., palm/finger Choi et al. [2012] or feet Healey and Picard [2005]. Alternative locations to measure an EDA signal is the wrist Gjoreski et al. [2017]. In order to assess EDA, the resistance between two electrodes is measured. From a technical point of view, see Dawson et al. [2000], EDA data is recorded employing either constant-current (measuring skin *resistance*) or constant-voltage systems (recording skin *conductance*). However, due to the more linear relationship between the skin conductance and the number of active sweat glands, Lykken and Venables [1971] argues strongly for a direct measure of the skin conductance using constant-voltage systems. In recent AR research the *Empatica E4* is a frequently employed device to collect EDA data Gjoreski et al. [2017], Di Lascio et al. [2019], Heinisch et al. [2019]. Having the form factor of a smartwatch, the E4 samples the EDA signal at 4 Hz, which is sufficient to distinguish the SCR from the SCL. Although the EDA is strongly influenced by the SNS, external parameters such as humidity, temperature, or the physical activity have a strong influence.

Although respiration can be assessed indirectly from measuring the blood oxygen level, a direct measurement contains more information about the actual respiration pattern. Commonly, a chest belt (**respiratory inductive plethysmograph (RIP)** Plarre et al. [2011]), which is either worn thoracically or abdominally, is utilised to measure the respiration pattern directly. During a respiration cycle (inhalation and exhalation), the thorax expands and constricts. Hence, the chest belt experiences a sinusoidal stretching and destretching process, from which different physiological parameters like respiration rate and volume can be derived. Healey and Picard [2005] sampled their respiration sensor at 31 Hz. However, following the Nyquist theorem a lower bound on the sampling rate of a RIP setup can be around 10-15 Hz. Nowadays, chest belts are mainly used by athletes monitoring their training progress. However, these devices have not found broad applications outside this domain.

Muscle activity is measured using surface **electromyogram (EMG)**. For this purpose, a pair (or array) of electrodes is attached to the skin above the muscle under consideration. The electrical potential is generated when the muscle cells are activated, and the surface electrodes are used to recorded changes in the electric potential. The frequency range of the muscle activity ranges from 15 to 500 Hz van Boxtel [2001]. Hence, in order to capture the full spectral range, the minimal sampling rate of the EMG modality should be around 1000 Hz. One source of noise in surface EMG are potential changes in adjacent muscles and heart rate activities. Depending on the measurement position, the QRS complex (indicating depolarization of the cardiac ventricles and the following contraction) can cause artefacts which require postprocessing beyond normal filtering. Considering related work in AR literature, EMG electrodes are often placed in the face (e.g. on the zygomaticus major Koelstra et al. [2012]) or on the shoulder (e.g. on the upper trapezius muscle Kim and André [2008], Wijsman et al. [2010], Koelstra et al. [2012]).

As the blood flow to the extremities is restricted during a 'fight or flight' response, changes in peripheral temperature is an interesting parameter. These changes in **skin-temperature (TEMP)** can be measured using either an infrared thermopile or a temperature-dependent resistor. A common confounding variable for body temperature measurements is the ambient temperature, which can have a strong influence on the recording depending on the location of the thermopile. As changes of the body temperature are low-frequent, a sampling rate of 1 Hz is sufficient.

The physiological modalities detailed above are only minimally intrusive. Hence, they are frequently employed in AR lab and field studies Lisetti and Nasoz [2004], Choi et al. [2012], Healey and Picard [2005], Kim et al. [2004]. In addition to the modalities listed above **electroencephalogram (EEG)** and **electrooculography (EOG)** are also often applied in AR lab studies. EEG, measuring the ionic current of brain neurons using electrodes placed on the scalp, was for instance employed by Soleymani et al. [2012b] to detect video-elicited emotions. EOG, which records horizontal and vertical eye movements by placing electrodes above/below and left/right

of the eye, has been used by Koelstra et al. [2012]. In our opinion, these modalities have the following disadvantages:

- Both require the placement of electrodes on the face/scalp. Hence, both EEG and EOG are quite intrusive.
- They pose strong limitations on the movement of the participants and, hence, are not really applicable in real world scenarios.
- EOG and EEG are prone to noise generated by muscle activity.

As stated in Section 1.1, EEG and EOG are not in scope of this work. Therefore, these modalities will be given very little attention in the remainder of this chapter and both modalities will be not employed in the studies presented later in Chapter 3 and Chapter 4.

Inertial sensors, incorporating a 3-axes acceleration (ACC), gyroscope, and magnetometer, are commonly used in human activity recognition (HAR). In AR field studies the ACC signal can provide context information about the physical activity of the user. Gjoreski et al. [2017], for instance, used ACC data to classify six different activity types (*lying, sitting, standing, walking, running*, and *cycling*). These activities, were then used as an additional input into a stress detection system. This certainly highlights the value of contextual information. However, results of Ramos et al. [2014] indicate that in order to detect stress it is sufficient to estimate the intensity level of an activity instead of performing an exact activity classification.

Finally, following Muaremi et al. [2013], smartphones offer an ideal platform to collect **context information**. This contextual data is aggregated by utilising position (GPS), sound snippets, calendar events, ambient light, and user interaction with the phone Muaremi et al. [2013], Mozos et al. [2017], Kanjo et al. [2019].

Table 2.4: Affective states and sensor signals frequently employed in wearable-based AR. Table 2.9 provides further detail on algorithms, location and performance. Abbreviations: 3-axes acceleration (ACC), blood pressure (BP), electrocardiogram (ECG), electrodermal activity (EDA), electroencephalogram (EEG), electromyogram (EMG), electrooculography (EOG), heart rate (HR), magnetoencephalogram (MEG), pupil diameter (PD), photoplethysmogram (PPG), respiration (RESP), skin-temperature (TEMP), arterial oxygen level (SpO2), low arousal/low valence (LALV), low arousal/high valence (LAHV), high arousal/low valence (HALV), high arousal/high valence (HAHV)

	Author	Affective States	Sensor Signals
<2005	Picard et al.	Neutral, anger, hate, grief, joy, platonic/romantic love, reverence	EDA, EMG, PPG, RESP
	Haag et al.	Low/medium/high arousal and positive/negative valence	ECG, EDA, EMG, TEMP, PPG, RESP
	Lisetti and Nasoz	Sadness, anger, fear, surprise, frustration, amusement	ECG, EDA, TEMP
	Liu et al.	Anxiety, boredom, engagement, frustration, anger	ECG, EDA, EMG

2005	Wagner et al.	Joy, anger, pleasure, sadness	ECG, EDA, EMG, RESP
	Healey and Picard	Three stress levels	ECG, EDA, EMG, RESP
07	Leon et al.	Neutral/positive/negative valence	EDA, HR, BP
2008	Zhai and Barreto	Relaxed and stressed	EDA, PD, PPG, TEMP
	Kim et al.	Distinguish high/low stress group of individuals	PPG
	Kim and André	Four quadrants in valence-arousal space	ECG, EDA, EMG, RESP
	Katsis et al.	High/low stress, disappointment, euphoria	ECG, EDA, EMG, RESP
2009	Calvo et al.	Neutral, anger, hate, grief, joy, platonic/romantic love, reverence	ECG, EMG
	Chanel et al.	Positively/negatively excited, calm-neutral (in valence-arousal space)	BP, EEG, EDA, PPG, RESP
	Khalili and Moradi	Positively/negatively excited, calm (in valence-arousal space)	BP, EEG, EDA, RESP, TEMP
10	Healey et al.	Points in valence arousal space. moods	ACC, EDA, HR, audio
2011	Plarre et al.	Baseline, different types of stress (social, cognitive, and physical), perceived stress	ACC, ECG, EDA, RESP, TEMP, ambient temperature
	Hernandez et al.	Detect stressful calls	EDA
2012	Valenza et al.	Five classes of arousal and five valence levels	ECG, EDA, RESP
	Hamdi et al.	Joy, sadness, disgust, anger, fear, surprise	ECG, EEG, EMG
	Agrafioti et al.	Neutral, gore, fear, disgust, excitement, erotica, game elicited mental arousal	ECG
	Koelstra et al.	Four quadrants in valence-arousal space	ECG, EDA, EEG, EMG, EOG, RESP, TEMP, facial video
	Soleymani et al.	Neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, fear	ECG, EDA, EEG, RESP, TEMP
2013	Sano and Picard	Stress vs. neutral	ACC, EDA, phone usage
	Martinez et al.	Relaxation, anxiety, excitement, fun	EDA, PPG
2014	Valenza et al.	Four quadrants in valence-arousal space	ECG
	Adams et al.	Stress vs. neutral (aroused vs. non-aroused)	EDA, audio
2015	Hovsepian et al.	Stress vs. neutral	ECG, RESP
	Abadi et al.	High/Low valence, arousal, and dominance	ECG, EOG, EMG, near-infrared face video, MEG
2016	Rubin et al.	Panic attack	ACC, ECG, RESP
	Jaques et al.	Stress, happiness, health values	EDA,TEMP, ACC, phone usage
	Rathod et al.	Normal, happy, sad, fear, anger	EDA, PPG

2016	Zenonos et al.	Excited, happy, calm, tired, bored, sad, stressed, angry	ACC, ECG, PPG, TEMP
	Zhu et al.	Angle in valence arousal space	ACC, phone context
	Birjandtalab et al.	Relaxation, different types of stress (physical, emotional, cognitive)	ACC, EDA, TEMP, HR, SpO2
2017	Gjoreski et al.	Lab: no/low/high stress; Field: stress vs. neutral	ACC, EDA, PPG, TEMP
	Mozos et al.	Stress vs. neutral	ACC, EDA, PPG, audio
	Taylor et al.	Tomorrow's mood, stress, health	ACC, EDA, context
	Girardi et al.	High vs. low valence and arousal	EEG,EDA,EMG
2018	Zhao et al.	LALV, LAHV, HALV, HAHV	EDA, PPG, TEMP
	Santamaria-Granados et al.	LALV, LAHV, HALV, HAHV	ECG, EDA
2019	Heinisch et al.	High positive pleasure high arousal, high negative pleasure high arousal, and neutral	EMG, PPG, TEMP
	Hassan et al.	Happy, relaxed, disgust, sad, and neutral	EDA, PPG, EMG (from DEAP)
	Kanjo et al.	Five valence classes	ACC, EDA, HR, TEMP, environmental, GPS
	Di Lascio et al.	Detect laughter episodes	ACC, EDA, PPG

Table 2.4 summarises recent wearable-based AR studies aspiring to detect different affective states, based on wearable-based data. A detailed comparison of the employed classification algorithm, number of target classes, setting (e.g., lab or field), number of subjects, validation procedure and obtained accuracy, will be presented in Table 2.9. In the studies presented in Table 2.4, the target affective states are rather diverse: Almost 39% of the presented studies aimed to detect stress. For this purpose, different types of stressors (e.g., mental, physical or social Plarre et al. [2011], Birjandtalab et al. [2016]) or different stress levels Gjoreski et al. [2017] are distinguished. Both the severe health implications and the strong physiological stress response (see Section 2.1.3), explain the popularity of stress recognition. According to Table 2.4, various studies aim to recognise different emotional categories, distinguishing up to eight different affective states. Dimensional models of emotions (e.g., valence-arousal space) were used in 36% of the analysed studies. Only in 14% of the considered studies EEG was recorded. Nevertheless, there exists a large body of work, utilizing EEG data to classify different affective states. However, as mentioned in Chapter 1 this modality is not in scope of this review. As a result, studies utilizing EEG data are given less attention here. Concluding from Table 2.4, sensor modalities monitoring cardiac activity are employed in 86% of the studies. EDA data was recorded in 75% of the studies. The popularity of these signals, certainly is linked to the strong impact of arousal-related changes on cardiac and electroder-

mal activity (see Section 2.2.1). In 32% of the considered studies, respiration data was acquired. Kim and André [2008] pointed out that increased arousal can lead to irregular respiration pattern. EMG and TEMP data were recorded in 32% of the studies. Finally, ACC was employed in 30% of the studies presented above. In summary, it is observed that sensors measuring parameters directly influenced by the SNS are most popular. Sensory setups recording less distinct changes are employed less frequently.

2.3 Affect-related User Studies

Picard et al. [2001] pointed out that, in order to generate high quality physiological data for affect detection, carefully designed study protocols are required. In order to reduce subject bias it might be necessary to disguise the true purpose of the study. However, if a deception is necessary for the protocol it is essential to uncover the true aim at the end of the protocol. Moreover, every study should be reviewed and approved by an ethics (or a similar) committee.

The arguably most important decision is whether the experiment is to be conducted in a laboratory setting or in the wild. A key issue when designing a field study is accurate label generation. In contrast, during a lab study, obtaining high quality labels is a minor issue as either the study protocol can be used or dedicated time slots for questionnaires can be reserved. However, considering lab studies, the desired affective states have to be elicited by a carefully chosen set of stimuli. If these stimuli are not appropriate, the desired effects might not occur. On the other hand, during field studies, affective stimuli do not have to be designed, as different affective states occur naturally. Section 2.3.1 provides an overview of protocols employed for user studies in the lab. Section 2.3.2 summarises related work on how to plan and conduct affect-related field studies. Finally, as conducting an own user study is always a time consuming task, publicly available datasets are described.

2.3.1 Affect-related User Studies in Laboratory Settings

Humans differ in their personality. Hence, generating data that corresponds to a particular emotional state is a challenging task Hamdi et al. [2012]. However, due to the controlled lab environment, researchers can conduct studies following well-designed protocols. Another advantage of lab studies is that their replication is possible, due to the well defined experimental protocol. Below a detailed overview of stimuli frequently employed to elicit affective states in AR lab studies is provided:

Images: The International Affective Picture System (IAPS) Lang et al. [1999] is a dataset comprised of colour photographs. The IAPS was compiled such that each image elicits an emotional reaction. Each image was rated multiple times by study participants, providing labels in the valence and arousal space. Mikels et al. [2005] identified a subset of IAPS images, which elicits certain discrete emotions. Hence, depending on the desired emotion, one can choose particularly strong images from this subset. In the AR domain, the IAPS has for instance been used by Leon et al. [2007] and by Hamdi et al. [2012]. In the experiments presented by Leon et al. [2007], 21 images from the IAPS were used to elicit three different affective states (*neutral, positive, negative*). Hamdi et al. [2012] exposed their study participants to ten images from the IAPS and aimed at recognising six basic emotions (*disgust, joy, surprise, sadness, fear, anger*) based on physiological data.

Videos: According to Gross and Levenson [1995], short audiovisual clips are very

suitable to elicit discrete emotions. Hence, video clips are frequently, e.g., Soleymani et al. [2012a], Abadi et al. [2015], Koelstra et al. [2012], employed as stimuli. A common procedure to select a set of videos evoking certain target emotions is to chose them from a large pool of videos. The process of identifying the most appropriate subset often happens in two steps: First, the clips are watched and rated by a large number of individuals. Second, the clips which elicit a certain emotion most reliably are chosen as stimuli in the study, see for instance Soleymani et al. [2012b], Koelstra et al. [2012]. Recently, Samson et al. [2016] published a study on 199 short amateur clips which were rated by 411 subjects with respect to three affective categories (*neutral, positive, negative*). In AR literature, there are many examples where audiovisual clips have been used to elicit different affective states. Koelstra et al. [2012] chose in their experiments music clips with a length of 60 seconds. After each stimulus, the progress was displayed and a 5 second baseline was recorded. Soleymani et al. [2012b] showed their participants 60 to 120 seconds long excerpts from movies and after each clip a short neutral clip (15 seconds) was displayed.

Acted Emotions: In the above detailed protocols, emotions are event-elicited. Another way of generating affective states is to ask the subjects to purposefully elicit emotions, e.g., act an emotion. For instance, Hanai and Ghassemi [2017] asked the study participants to tell at least one happy and one sad story. Other researchers Castellano et al. [2008], Dobriek et al. [2013] asked trained actors to perform certain emotions. These types of approaches are frequently employed in sentiment analysis and emotion recognition from audio/video data.

Game elicited emotions: Another way to elicit a target affective state is to ask the subjects to perform a certain task. Using a Breakout engine and applying latency between the user's input and the reaction in the game, Taylor et al. [2015a] elicited frustration in their study participants. Martinez et al. [2013] used four different versions of a Maze-Ball game to generate pairwise preference scores. The scores were generated by asking the subjects which of two games felt more *anxious, exciting, frustrating, fun,* and *relaxing.*

Stress inducing study protocols: There are numerous protocols aiming at eliciting stress in the study participants. Mason [1968] showed that in order to trigger a (physiological) stress response, the situation has to be either novel, and/or unpredictable, and/or beyond control for the subject Lupien et al. [2007]. Stressors frequently employed in AR literature can be categorised as follows:

- **C1** *Social-evaluative Stressors:* A task creating a socially relevant situation for the subject. For example, performing a task in front of a panel which evaluates the subject.
- **C2** *Cognitive Stressors:* A task demanding significant mental engagement and attention. For example, performing an (challenging) arithmetic task under time pressure.
- **C3** *Physical Stressors*: A task creating a physically uncomfortable situation. For example, being exposed to extreme hot or cold.

A well-studied and frequently employed stress elicitation protocol is the *Trier Social Stress Test* Kirschbaum et al. [1993]. The Trier Social Stress Test (TSST) has two conditions: A public speaking/job interview type of situation and a mental arithmetic task. Hence, the TSST incorporates both a social-evaluative (**C1**) and cognitive stressor (**C2**). Due to its reliability and easy set-up, the TSST was administered in numerous AR studies, e.g., Mozos et al. [2017], Plarre et al. [2011], Hovsepian et al. [2015], Gjoreski et al. [2016]. Another stressor employed to target cognitive load is the so called *Stroop color test* Stroop [1935]. In this condition, the subjects have to read out loud a sequence of colours written on a screen. However, the font colour does not match the written colour (e.g., green, blue, etc.). As a result, the task inflicts a high cognitive load and, hence, is a **C2** stressor. The Stroop colour test has for instance been employed by Choi et al. [2012], who aimed for the development of a wearable-based stress monitoring system.

Using *computer tasks*, stress can also be elicited reliable. Wijsman et al. [2013], for instance, asked the subjects to perform a calculation, to solve a logical puzzle, and to do a memorisation task. These tasks can all be seen as **C2** stressors. These tasks had to be completed under time pressure. In addition, the subjects were distracted with sounds and parts of the protocol (memorisation task) were also recorded on video. Furthermore, as the participants of Wijsman et al. [2013] were told that their scores would be made available to their colleagues, the study protocol also had a social-evaluative component (see **C1**).

The *cold pressor* test, applied by Plarre et al. [2011], can be used to evoke physical stress, corresponding to a **C3** stressor. Following this test, the subjects are asked to place their hand into a bucket of ice cold water and leave it there for a predefined time (e.g., 60 seconds).

Now as a common set of stimuli has been detailed, the issue of **obtaining ground truth in a lab setting** is discussed briefly. Following for instance Plarre et al. [2011], employed conditions (e.g., stressors) can be used as ground truth. One way to ensure the validity of the employed stimulus is to utilize *exactly the same* set up as in a related study. In addition, questionnaires integrated into the protocol can be used to verify that the desired affective states were successfully evoked. Typically, these questionnaires are used directly after each affective stimulus or condition. Ramos et al. [2014], for instance, collected subjective stress levels after each stressor. In addition, the State-Trait Anxiety Inventory (STAI) also has been used to capture different stress levels Gjoreski et al. [2017]. In order to generate labels in valence-arousal space the SAM are employed frequently Koelstra et al. [2012], Soleymani et al. [2012b]. In addition, as the perception of a stimulus can be influenced by **personality traits**, collecting this type of information, can be useful too Subramanian et al. [2017].

2.3.2 Affect-related User Studies in the Field

To develop affect-aware systems designed for everyday usage, data collection in the wild is essential. However, as the affective states occur naturally, the generation of a reliable ground truth has to be ensured differently. In this setting one can distinguish between questionnaires used in ecological-momentary-assessments (EMAs) and questionnaires employed during the pre- and post study phase. In the latter case constructs which are said to be constant for a longer time period (e.g., personality traits) are being queried. To assess the momentary affective state of a user, **EMAs**, also known as the experience sampling method, are employed. EMAs are a short set of questionnaires which the study participants file occasionally, to report their current affective state.

Using EMAs, an important trade-off has to be considered. On one hand the affective state of the subject should be probed frequently. On the other hand, the subject should not be overloaded with questionnaires. The scheduling of EMAs can be either done *interval-based* (e.g., at certain/random times during the day) or *event-triggered*. In a study of Zenonos et al. [2016], for instance, the subjects were prompted every two hours during their working hours. The EMAs employed by Zenonos et al. [2016], inquired eight different moods, asking for each the question *How have you been feeling for the last two hours?*. Another approach is to *distribute* a defined number of EMAs *randomly* over a time period. Muaremi et al. [2013], for instance, divided the day into four sections, and during each section subjects had to complete a randomly scheduled self-report. If the focus of a study lies on certain affective states or events, *event-triggered* self-reports can be utilized. In a study conducted by Hernandez et al. [2011] call centre employees rated personal stress level after each call. Another example of event-based scheduling can be found by Rubin et al. [2015]: Here subjects were asked to file an EMA once they became aware of the symptoms of a panic attack. In order to gain a deeper understanding of EMAs filed by the subjects daily screenings can be conducted. Following Healey et al. [2010], these screenings can be used to correct/extend participants' annotations.

Besides the frequency of EMAs, the length and complexity of each single questionnaire are also important factors defining the burden for the subjects. In order to avoid overloading study participants, EMAs should focus on the main goal of the study and their completion should require only little effort.

In Table 2.5 questionnaires used during the pre- and post study as well as questionnaires employed in EMAs are displayed. As mentioned earlier the pre- and post study questionnaires, are used to aggregate information about longer time periods or traits of the subjects. Subjects' **personality traits** can have an influence on their affective perception and physiological response Subramanian et al. [2017]. Therefore, completing a personality-related questionnaire can provide valuable insights.

Table 2.5: Questionnaires utilized in recent wearable-based AR field studies. Abbreviations: Number of Items (I), Big Five Inventory (BFI), Photo Affect Meter (PAM), Positive and Negative Affect Schedule (PANAS), Patient Health Questionnaire (PHQ-9), Pittsburgh Sleep Quality Index (PSQI), Perceived Stress Scale (PSS), Self-Assessment Manikins (SAM), Stress Response Inventory (SRI), State-Trait Anxiety Inventory (STAI).

Questionnaires employed *prior or after* the study.				
Goal	**Tool and description**	**I**	**Source**	**Example use**
Stress level	PSS: subject's perception and awareness of stress	10	Cohen et al. [1983]	Sano and Picard [2013]
	SRI: score severity of stress-related symptoms within time interval	22	Koh et al. [2001]	Kim et al. [2008]
Depression level	PHQ-9: score DSM-IV manual	9	Kroenke et al. [2001]	Wang et al. [2014]
Loneliness level	UCLA loneliness scale: addressing loneliness and social isolation.	20	Russell [1996]	Wang et al. [2014]
Sleep behaviour and quality	PSQI: Providing information about sleep quality	19	Buysse et al. [1989]	Sano and Picard [2013]
Measure success areas	Flourishing scale: measure success, self-esteem, purpose, and optimism	8	Diener et al. [2010]	Wang et al. [2014]
Personality traits	BFI: indicating personality traits	44	John and Srivastava [1999]	Taylor et al. [2017], Sano et al. [2015]
Questionnaires employed in ecological-momentary-assessment (during study).				
Valence-arousal representation	Mood Map: a translation of the circumplex model of emotion	2	Morris and Guilak [2009]	Healey et al. [2010]
Positive and negative affect	Shortened PANAS	10	Muaremi et al. [2013]	Muaremi et al. [2013]
Positive Affect of PANAS	PAM: choose one of 16 images, mapped to the valence-arousal space	1	Pollak et al. [2011]	Wang et al. [2014]
Subjective mood	Smartphone app querying user's mood	8	HealthyOffice app	Zenonos et al. [2016]
Stress level assessment	Adaptation of PSS for ambulatory settings	5	Hovsepian et al. [2015]	Hovsepian et al. [2015]
	Log current Stress Level	1	Gjoreski et al. [2017] Hernandez et al. [2011]	Gjoreski et al. [2017] Hernandez et al. [2011]
Severity of panic attack symptoms	Symptoms from the DSM-IV and Panic Disorder Severity Scale standard instrument	15	Shear et al. [1997]	Rubin et al. [2015]

These BFI personality traits were, for instance, used by Sano et al. [2015] as features for predicting subjects' mood. In addition, Taylor et al. [2017] used personality traits to perform a groupwise personalization. Moreover, Wang et al. [2014] used questionnaires assessing the mental health of their participants. For this purpose, the depression level (e.g., PHQ-9) and loneliness level (UCLA loneliness scale) were recorded. As shown by Sano and Picard [2013] and Sano et al. [2015], information on subjects' **sleep quality** can be useful in affect-related studies. The PSQI, inquiring information about the past four weeks, can serve as a suitable questionnaire for sleep behaviour and quality assessment. In order to assess the overall stress level of the study participants the PSS, measuring the perception and awareness of stress, can be employed. The PSS has been used in field studies (e.g., Sano and Picard [2013], Wang et al. [2014]) and in ambulatory setting Hovsepian et al. [2015]. The severity of stress-related symptoms can be scored using the SRI, or a simplified version of it, as shown by Kim et al. [2008].

As detailed in Table 2.4, wearable-based AR studies, typically rely on well-known psychological constructs. Hence, in order to generate labels using EMAs these constructs are employed, too. However, standard questionnaires are often quite long and as a result not really applicable in EMAs. In order to mitigate this issue, standard questionnaires can be shortened, e.g., using only a subset of items with the highest factor loads on the targeted construct. Such an approach was for instance presented by Muaremi et al. [2013] using a shortened version of the PANAS as EMA, which consisted of five positive affect items (*relaxed, happy, concentrated, interested*, and *active*) and five negative affect items (*tired, stressed, sleepy, angry*, and *depressed*). One particularly frequently employed construct is the valence-arousal space. In order to generate **valence and arousal labels**, Healey et al. [2010], for instance, used a tool called Mood Map. Furthermore, Wang et al. [2014] used the PAM, assessing a similar construct. The PAM is implemented as smartphone app, and the user selects from a set of 16 images the one that corresponds best to his/her current affective state. Zenonos et al. [2016] provides an example for a custom EMA tool used for overall mood assessment: Participants were asked to rate eight different moods on a scale from 0-100. The stress level of subjects can be assessed using a Likert-scale Hernandez et al. [2011], Gjoreski et al. [2017]. Moreover, the severity of a certain event can be scored using its' symptoms. Rubin et al. [2015], for instance, aimed to quantify the severity of panic attacks. Hence, they created a questionnaire including 15 panic attack symptoms. In case a panic attack occurred, subjects were asked to rate the severity of each of the 15 symptoms, using a severity rating of 1 (none) to 5 (extreme).

Historically, personal notebooks or journals were used for EMAs. However, these tools have been predominantly replaced by smartphone apps, as they offer an ideal platform to facilitate self-reports: Subjects do not need to carry a study-specific device, EMAs are automatically scheduled and uploaded, and contextual information available on the smartphone can be logged together with the ground truth informa-

Table 2.6: Questionnaires employed during recent field studies, focusing on the applied scheduling (Pre-, During, or Post-study).

	Author	Employed Questionnaires and their scheduling.
Emotion	Healey et al. [2010]	*During study:* Participants completed EMAs whenever they felt a change in their affective/physiological state. EMAs included a form of the circumplex model and a field for free text. Conducted Interviews at the end of each workday to generate additional labels and revision.
	Rubin et al. [2015]	*During study:* Start/stop time and severity ratings of 15 panic attack symptoms were reported by the subject using a mobile app.
	Jaques et al. [2016]	*During study:* Students reported health, stress and happiness twice a day (morning and evening).
Stress	Hernandez et al. [2011]	*During study:* Nine employees of a call center rated all their incoming calls on a 7 point likert scale (endpoints marked as "extremely good/bad").
	Muaremi et al. [2013]	*During:* Participants were asked to fill in a shortened PANAS four times between 8 a.m and 8 p.m. Before going to sleep they answered the question: "How stressful have you felt today?"
	Kim et al. [2008]	*Pre-study:* In order to divide the subjects into two groups they filled out a simplified SRI.
	Sano and Picard [2013]	*Pre-study:* Participants filled in a PSS, PSQI, and BFI. *During study:* Morning/evening EMAs on sleep, mood, stress level, health, etc. *Post-study:* Participants filled in questionnaires on health, mood, and stress.
	Adams et al. [2014]	*Pre-study:* Participants completed a PANAS, PSS, and a measure of mindfulness. *During study:* Self-reports approximately every 30 min. (with small random variations). Participants reported on momentary stress and affect. Additional reports and a small free text field were available too. *Post-study:* Semi-structured interview at the end of the end data collection.
	Hovsepian et al. [2015]	*During study:* EMAs randomly scheduled approximately 15 times. During each EMA subjects filled in a shortened version of the PSS containing 6 items.
	Gjoreski et al. [2017]	*During study:* Subjects replied to 4 to 6 randomly scheduled EMAs. During each EMA subjects reported on their current stress level.
Mood	Wang et al. [2014]	*Pre-study:* Subject filled in a number of behavioural and health surveys. *During study:* Every participant filled in 8 EMAs every day. The EMAs include measures on mood, health, stress and other affective states. *Post-study:* Interviews and the same set of behavioural and health surveys were administered.
	Sano et al. [2015]	*Pre-study:* subjects filed BFI, PSQI, and Morningness-Eveningness Horne and Ostberg [1976] questionnaire. *During study:* similar to Sano and Picard [2013] subject filled EMAs in morning and evening reporting on: activities, sleep, social interaction, health,mood, stress level and tiredness. *Post-study:* Subjects filed in a PSS, STAI, and other questionnaires related to physical and mental health.
	Zenonos et al. [2016]	*During study:* EMAs were scheduled every two hours. For the EMAs an app was used, containing sliders from 0-100 for 8 moods. Additionally, a free text field was provided.

tion. A key to both frequency and completeness of EMA is participant's motivation and using an appropriate **reward system** was proven to be beneficial: Participants of the study conducted by Healey et al. [2010] received a base reward and an incremental reward, depending on the number of annotations made per day. Another reward structure was introduced by Wang et al. [2014]: They offered all subjects a base reward, and the participants who completed most EMAs had the chance to win additional prizes.

In Table 2.6 an overview of recent **wearable-based AR field studies** is provided and the employed EMAs as well as their scheduling is summarized. This table illustrates that commonly a combination of pre-/post-study questionnaires are used. The pre-/post-study questionnaires can be employed as additional features or to group the participants Taylor et al. [2017], Kim et al. [2008]. In contrast, the data gathered via EMAs is often used as a subjective ground truth Rubin et al. [2016], Gjoreski et al. [2017].

2.3.3 Publicly Available Datasets

Conducting a user study is both a time consuming and challenging task. However, there are a number of publicly available datasets. Depending on the research idea these datasets make the overhead of recording an own dataset obsolete. Furthermore, considering research question **RQ 2**, these datasets facilitate benchmarking and allow a direct comparison of different approaches. Up-to-date the wearable-based AR community has only a handful of publicly available datasets containing data *solely* gathered via wearables. Therefore, we extend the scope of this section to datasets with a broader relevance to wearable AR. Below we present datasets which meet one of the following criteria: a) being publicly available, b) including data recorded from study participants being subject either to emotional stimuli or a stressor, and c) including at least a few sensor modalities which can be (theoretically) integrated into consumer-grade wearables, which are applicable in everyday life. The datasets included in our analysis are summarized in Table 2.7. Considering the population column in Table 2.7 it becomes apparent, that the data available originates mostly from a young cohort of subjects. Only the data set recorded by Taamneh et al. [2017], features two different age groups, namely an elderly (>60) and a young group (between 18 and 27). This is certainly a limitation that needs to be considered when working with these datasets. Below we describe the datasets in detail.

The **Eight-Emotion** dataset Picard et al. [2001] includes data of one (female) study participant who was subject to the same set of stimuli over a time span of 20 days. The stimuli, a set of personally-significant imagery, were chosen by the subject to elicit the affective states *neutral, anger, hate, grief, platonic love, romantic love, joy,* and *reverence*. The physiological signals (ECG, EDA, EMG, and RESP) were sampled at 20 Hz. Major limitations of this dataset are: a) only one subject is included, and b) due to the low sampling rate aliasing artefacts are likely to occur.

DEAP (Database for Emotion Analysis using Physiological signals), recorded by Koelstra et al. [2012], features physiological data of 32 study participants. In DEAP, one minute excerpts of music videos were used as stimuli. In total 40 clips were selected from a larger pool according to valence, arousal, and dominance ratings gathered during a pre-study. The physiological signals were all sampled with 512 Hz and later downsampled to 256 Hz. DEAP includes subjects' ratings of the videos (valence, arousal, dominance, and liking). However, due to the employed protocol and the sensor setup, the DEAP participants were very limited in terms of movement. Therefore, one can expect that models trained on the DEAP dataset will have a limited performance in real-life settings.

The **MAHNOB-HCI** dataset, Soleymani et al. [2012a], includes physiological data from 27 study participants (16 female). The dataset includes face and body video data from six cameras, and data from an eye gaze tracker and audio. The physiological data (ECG, EDA, EEG, RESP and TEMP) was sampled at 1024 Hz. Apart from EEG data, the physiological data was downsampled to 256 Hz. The MAHNOB-HCI dataset includes data from two experiments: First, study participants watched a set of 20 video clips, each associated with an emotional keyword (*disgust, amusement, joy, fear, sadness*, and *neutral*). The goal of the second experiment was implicit tagging: Subjects were exposed to 28 images and 14 videos, and reported on the agreement with the displayed tags. For the AR community, especially the first experiment is of interest.

DECAF (DECoding user physiological responses to AFfective multimedia content) Abadi et al. [2015] was recorded in a laboratory setting with 30 subjects (14 female). The data recording consisted of two sessions for each subject, presenting music videos and movie clips, respectively. In the first session (music videos) the same set of clips as in DEAP were employed. For the second session, 36 movie clips were used as stimuli. From this pool of videos always nine correspond to a quadrant in the valence-arousal space. These 36 movie clips were selected from a larger pool during a pre-study based on valence-arousal ratings from 42 participants. For a detailed description, we refer the reader to Abadi et al. [2015]. DECAF contains image (near-infrared face videos), magnetoencephalogram (MEG), and peripheral sensory data (ECG, EOG, and EMG). A clear limitation of DECAF is that, due to the MEG recordings, subjects were very restricted in their movements. Therefore, in contrast to real-life data DECAF is almost free from motion artefacts.

In **ASCERTAIN** (multimodal databASe for impliCit pERsonaliTy and Affect recognitIoN using commercial physiological sensors), recorded by Subramanian et al. [2017], the same 36 movie clips as in DECAF were employed as stimuli. ASCERTAIN provides data from 58 subjects (21 female), and includes physiological modalities (ECG, EDA, EEG) as well as data recorded from a facial feature tracker. In addition, self-reports including arousal, valence, engagement, liking, and familiarity obtained for each video are included. Moreover, the dataset contains the Big Five personality traits for each subject. Hence, based on the recorded data, not only mo-

Table 2.7: Publicly available datasets relevant for wearable affect and stress recognition. Abbreviations: Number of subjects (Sub), Location (Loc), Lab (L), Field (F), Field with constraint (FC), Population (Pop) reported as mean age or as category, College Student (CS), Graduate Student (GS), 3-axes acceleration (ACC), electrocardiogram (ECG), electrodermal activity (EDA), electroencephalogram (EEG), electromyogram (EMG), electrooculography (EOG), magnetoencephalogram (MEG), respiration (RESP), arterial oxygen level (SpO2), skin-temperature (TEMP)

	Name	Labels	Pop.	Sub.	Loc.	Included Modalities
Emotion (E)	Eight-Emotion[1]	Neutral, anger, hate, grief, joy, platonic love, romantic love, reverence	GS	1	L	ECG, EDA, EMG, RESP
	DEAP[2]	Continuous scale of valence, arousal, liking, dominance, Discrete scale of familiarity	26.9	32	L	ECG, EDA, EEG, EMG, EOG, RESP, TEMP, face video (not all subjects)
	MAHNOB-HCI[3]	Discrete scale of valence, arousal, dominance, predictability, Emotional keywords	26.06	27	L	ECG, EDA EEG, RESP, TEMP, face and body video, eye gaze tracker, audio
	DECAF[4]	Discrete scale of valence, arousal, dominance	27.3	30	L	ECG, EMG, EOG, MEG, near-infrared face video
	ASCERTAIN[5]	Discrete scale of valence, arousal, liking, engagement, familiarity, Big Five	30	58	L	ECG, EDA, EEG, facial activity data (facial landmark trajectories)
	USI_Laughs [6]	Detect and distinguish laughter from other events	26.70	34	L	ACC, EDA, PPG, TEMP
Stress (S)	Driver[7]	Stress levels: low, medium, high	-	24	FC	ECG, EDA, EMG, RESP
	Non-EEG[8]	Four types of stress (physical, emotional, cognitive, none)	CS	20	L	ACC, EDA, HR, TEMP, SpO2
	Distracted Driving[9]	Driving being subject to no, emotional, cognitive, and sensorimotor distraction	Elder + Young	68	L	EDA, heart and respiration rate, facial expressions, eye tracking
	StudentLife[10]	Sleep, activity, sociability, mental well-being, stress, academic performance	CS + GS	48	F	ACC, audio, context, GPS, smartphone usage

[1] Picard et al. [2001], [2] Koelstra et al. [2012], [3] Soleymani et al. [2012a], [4] Abadi et al. [2015], [5] Subramanian et al. [2017], [6] Di Lascio et al. [2019], [7] Healey and Picard [2005], [8] Birjandtalab et al. [2016], [9] Taamneh et al. [2017] [10] Wang et al. [2014],

dels predicting emotions can be created, but also personality traits can be assessed.

USI_Laughs has been recently published by Di Lascio et al. [2019]. The dataset contains physiological data recorded from 34 participants (6 female) recorded via an *Empatica E4* smartwatch (ACC, EDA, PPG, TEMP). Similar to prior work funny clips were used to induce laughter. Following Di Lascio et al. the main aim of the dataset is to facilitate the detection of laughter episodes based on physiological data. Here, the laughter episodes are to be considered as surrogate to positive emotions.

The **Driver stress** dataset Healey and Picard [2005] includes physiological data (ECG, EDA, EMG, and RESP) from 24 participants. The dataset was recorded during one *rest* condition and two driving tasks (*city* streets and on a *highway* near Boston, Massachusetts). Depending on traffic the two driving tasks had a duration between 50 and 90 minutes. Using questionnaires and a score derived from observable events, the three study conditions (*rest, highway, city*) were mapped onto the stress levels low, medium, and high. Therefore, the dataset facilitates the development of real-life stress monitoring approaches. However, one limitation of the dataset is that the sensor data was acquired at low sampling rates.

Distracted Driving, recorded by Taamneh et al. [2017], includes multimodal (physiological and eye tracking) data from 68 subjects driving in a simulator on a highway. All participants were subject to four different distractions: no, emotional, cognitive, and sensorimotor distraction. As the dataset includes among other modalities EDA, heart and respiration rate. This data can be used to study the influence of different distractions on these parameter.

Non-EEG Birjandtalab et al. [2016] is a dataset containing physiological data (EDA, HR, TEMP, SpO2, and ACC) from 20 subjects (4 female). The dataset was recorded during three different stress conditions (physical, cognitive, and emotional) and a relaxation task. Physical stress was evoked by asking the subjects to jog on a treadmill at three miles per hour. In order to elicit cognitive stress, the subjects had to count backwards from 2485 doing steps of seven. Lastly, emotional stress was triggered by anticipating and watching a clip from a zombie apocalypse movie. This dataset is particularly interesting as it contains only wearable-based data. Although the data collection was conducted in a lab setting, the subjects were (compared to the other datasets) less motion constrained due to the minimally intrusive nature of the sensors. However, a major limitation of the Non-EEG dataset is the low sampling rate of the employed devices (1 Hz and 8 Hz). In addition, as no ECG or PPG data was recorded, the HRV information can not be retrieved, a parameter shown to be relevant for stress recognition by various previous work (e.g., Kreibig [2010]).

StudentLife Wang et al. [2014] contains data from 48 college students (10 female). All participants were monitored over one academic semester (10 weeks). Unlike the afore described datasets StudentLife was recorded in the field. Considering the progress of the semester, it is expected that the students were more stressed towards the end of the data collection. This can be attributed to the examination pe-

riod. StudentLife contains data recorded from the students' smartphones (e.g., ACC, microphone, light sensor, and GPS/Bluetooth data). Moreover, various information related to the students' context (e.g., class attendance) and smartphone usage (e.g., conversation frequency and duration) were recorded. In addition, StudentLife includes a large number of self-reports targeting physical activity, sleep, perceived stress, mood, mental well-being, etc. Due to the popularity of smartphones, the dataset is certainly of interest by facilitating affect and stress recognition purely based on smartphone usage patterns.' However, a drawback of StudentLife is that it does not include any physiological data.

2.4 Data Processing and Classification

In wearable-based AR similar methods as in HAR are employed. Following the classical time series analysis pipeline, presented by Bulling et al. [2014], the raw data is first synchronised, filtered, segmented, features are computed, and finally feature-based classifiers are employed. The remainder of this section is structured as follows: In Section 2.4.1 the preprocessing of the raw data and segmentation is described. Section 2.4.2 provides an overview of features commonly used in wearable-based AR. The last step in the standard data processing pipeline is the classification. During this step a mapping between the computed feature and labels (e.g. emotion classes) is learned. Section 2.4.3 details common classification methods, applied validation schemes, and the results achieved in related work.

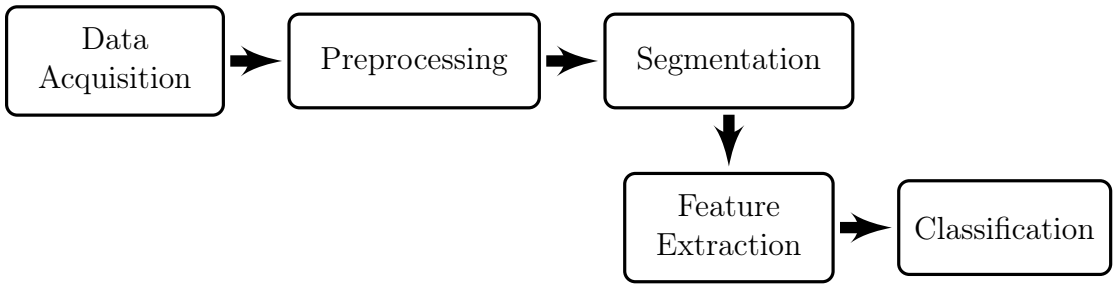


Figure 2.3: Standard time series classification chain.

2.4.1 Preprocessing and Segmentation

When multimodal systems are employed, synchronisation of the different raw data streams might be necessary as a first step. Clear events, e.g., pressing an event marker button or double tap gestures, can facilitate the synchronisation process. Depending on the transmission protocol of the recorded data, wireless data loss might be an issue. Different methods for handling missing values have been reviewed by García-Laencina et al. [2010]. Omitting cases with missing data, is arguably the simplest of these method. However, it comes at the cost of losing a lot of information. Imputation, estimation of missing data points is another more elaborate approach.

A common step in preprocessing is to apply denoising filters, in order to improve overall signal quality. The type of filtering strongly depends on the respective sensor modality. Therefore, below an overview of the different filtering and further preprocessing techniques, applied to the modalities in scope (see Section 2.2.2) are detailed.

1. **3-axes Acceleration Preprocessing:** A detailed analysis of preprocessing applied to ACC data can be found in Figo et al. [2010]. In AR, the ACC

data is often considered as a surrogate for the performed activity Mozos et al. [2017], Gjoreski et al. [2017].

2. **Electrocardiogram Preprocessing:** In the raw ECG signal the R-peaks need to be identified. For this purpose, the Pan and Tompkin's algorithm can be applied Pan and Tompkins [1985]. Once the R-peaks have been detected, the next step is to determine the RR intervals and assess their validity. For example, Hovsepian et al. [2015] present an algorithm to assess the validity of a candidate RR intervals. Behar et al. [2013], presented an approach to assess the ECG signal quality in regards to arrhythmia in the context of intensive care units. Similar approaches could also be utilized to assess the ECG quality during affect-related user studies.
3. **Photoplethysmogram Preprocessing:** A detailed description on PPG signal preprocessing methods applied to PPG data can be found in Elgendi [2012], Biswas et al. [2019]. In order to remove motion artefacts, adaptive (filtering) approaches can be applied Lee et al. [2010], Ram et al. [2012]. In more recent work, e.g., Reiss et al. [2019], Salehizadeh et al. [2016], peak matching approaches in the spectral domain were employed to remove movement artefacts. For the determination of RR intervals from identified R-peaks, similar algorithms as mentioned with ECG preprocessing can be applied. In addition, as shown by Li and Clifford [2012], the quality of a PPG signal can be assessed using a combination of dynamical time warping and multilayer perceptron (MLP).
4. **Electrodermal Activity Preprocessing:** In order to remove artefacts from EDA data different approaches were presented. The approaches can be grouped into filtering and machine learning-based approaches. Only changes in the low-frequency domain of the EDA signal are physiologically plausible. Hence, low-pass filtering with a cut-off of e.g. 5 Hz Setz et al. [2010] can be applied to remove high-frequency noise. After the noise removal, e.g., Soleymani et al. [2012a], detrended the EDA signal by subtracting a moving average, computed on smoothed version of the signal. Machine learning-based approaches, using support vector machines (SVMs) or convex optimization, to identify and remove artefacts in EDA data can be found in Taylor et al. [2015b], Greco et al. [2016]. As detailed in Section 2.2, the EDA signal consists of two components: A slowly varying baseline conductivity referred to as SCL and a series of peaks referred to as SCR. In literature different approaches to separate these two components can be found: Benedek and Kaernbach [2010], for instance, present an approach to separate SCL and SCR relying on nonnegative devolution. Alternatively, Choi et al. [2012] utilized, a regularized least-squares detrending method, to separate the two components.
5. **Electromyogram Preprocessing:** Raw EMG data is often filtered to remove noise. For example, Wijsman et al. [2010] report on a two step procedure. First, a bandpass filter, allowing frequencies from 20 to 450 Hz, was applied.

Then, in order to remove residual power line interference from data, notch filters were applied. The notch filters attenuated the 50, 100, 150, 200, 250 and 350 Hz components of the signal. Cardiac artefacts are another common source of noise in EMG data. Hence, Willigenburg et al. [2012] propose and compare different filtering procedures to remove ECG interference from the EMG signal.

6. **Respiration Preprocessing:** Depending on the signal quality, noise removal filtering techniques (e.g., bandpass filter with cut-off frequencies at 0.1 and 0.35 Hz) have to be applied. In addition, the raw respiration signal can be detrended by subtracting a moving average Khalili and Moradi [2009].

In the classical processing chain these preprocessing steps are followed by the **segmentation**. During this procedure the data is segmented using a sliding window of fixed size. The appropriate window size is crucial and depends on several aspects, such as the classification task or the considered sensor modality. Below appropriate choices for the window length of motion (ACC) and physiological data will be provided. In HAR, ACC data is most frequently employed to detect activities and there exists a body of work, e.g., Huynh and Schiele [2005], Reiss and Stricker [2012], Healey et al. [2010], identifying appropriate window sizes for HAR. A common finding is that in HAR the relevant patterns occur on short time scales. Therefore, window sizes of ~ 5 seconds are common.

The time scales on which physiological responses to emotional stimuli occur are hard to define. Hence, considering physiological signals, finding an appropriate window size is difficult Healey et al. [2010]. Moreover, due to inter-subject and inter-modality (e.g., ECG vs. EDA) differences, defining an appropriate window size becomes even more challenging. However, a meta analysis conducted by Kreibig [2010] found that physiological features are commonly aggregated over fixed window lengths of 30 to 60 seconds.

2.4.2 Physiological Feature Extraction

Following the classical time series classification pipeline, features are computed on the segmented data. These features aggregate information present in the signal, and serve as inputs into the classifier. Extracted features can be grouped in various ways, such as time- or frequency-domain features, linear or non-linear features, unimodal or multimodal features, etc. Considering computational complexity, extracted features range from simple statistical features (e.g., mean, standard deviation) to often modality-dependent complex features (e.g., number of SCR peaks). Table 2.8 gives an overview of features commonly extracted and applied in the wearable-based AR literature. In the remainder of this section, we give a brief description of features commonly extracted from different wearable sensors. As mentioned previously EEG

and EOG are not in scope of this work and, hence, will not be detailed here. For a comprehensive review on EEG-based AR we refer the reader to Kim et al. [2013].

From the HAR domain, a large set of ACC-based features is known. These features are often also employed in AR. Statistical features (mean, median, standard deviation, etc.) are often computed for each channel (x, y, z) separately and combined. Parkka et al. [2007] showed that the absolute integral of acceleration can be used to estimate the metabolic equivalent of physical activities, which can be an interesting feature for affect recognition as well. Mozos et al. [2017] used the first and second derivative of the accelerometer's energy as feature, e.g., to indicate the direction of change in activity level. Considering frequency-domain features, the power ratio of certain defined frequency bands, the peak frequency, or the entropy of the power spectral density (PSD) have been applied successfully.

From **ECG and PPG** data, various features related to cardiac activity are derived. Below, we provide a description of features commonly used in AR. For an in-depth analysis of features based on the cardiac cycle we refer to Malik [1996]. Commonly the HR is used as feature. Based on the location of the R-peaks (or the systolic peak in the PPG signal) the inter beat interval (IBI) can be computed. The IBI serves as a new time series signal, from which various HRV features can be derived, both in *time- and frequency-domain*. For instance, from the IBI the number and percentage of successive RR intervals differing by more than a certain amount of time (e.g. 20 or 50 milliseconds) can be computed. These feature are referred to as NNX and pNNX, where X is the time difference threshold in milliseconds. Based on the Fourier-transformation of the IBI time series, various frequency-domain features can be computed, which reflect the sympathetic and parasympathetic activities of the ANS. Four different frequency bands are established in this respect Rubin et al. [2016]. The ultra low frequency (ULF) and very low frequency (VLF) bands range from 0 to 0.003 Hz and from 0.003 to 0.03 Hz, respectively. Changes in low frequency (LF) band, ranged between 0.03 and 0.15 Hz, are mostly associated with the activity of the SNS. In contrast, the high frequency (HF) band, ranged from 0.15 to 0.4 Hz, is believed to reflect mostly the activity of the PNS Rubin et al. [2016]. Therefore, the LF/HF ratio quantizes is a descriptive feature indicating the influence of both, SNS and PNS, on the cardiac activity. In literature, e.g., Healey and Picard [2005], it was shown that the LF/HF ratio is a good indicator for stress. In addition to time and frequency domain-based features, *non-linear features* derived from ECG data were employed successfully wearable-based AR. Rubin et al. [2016], for instance, presents a detailed description of non-linear ECG features (e.g., maximal Lyapunov exponent, standard deviations (SD_1 and SD_2) along major axes of a Poincaré plot, the SD_1/SD_2 ratio, sample entropy, etc.). Moreover, Valenza et al. [2012], aiming to detect five levels of valence and arousal, compared the performance of linear and non-linear features. Their results indicate that non-linear features are able to improve classification scores significantly. Another class of features based on the cardiac cycle are referred to as *geometrical features*. An example

is the triangular interpolation index (TINN): A histogram of the RR intervals is computed, a triangular interpolation performed, and the baseline of the distribution is computed. The TINN is, for instance, used by Malik [1996], Valenza et al. [2012], and Rubin et al. [2016]. Finally, the respiration is known to have an impact on the ECG signal. In literature, there exist different approaches for quantifying the effect of the respiration on ECG data: Hovsepian et al. [2015], for instance, employed the respiratory sinus arrhythmia (RSA), which is calculated by subtracting the shortest RR interval from the longest RR interval within one respiration cycle. In addition, Choi et al. [2012] proposed a method of decomposing the HRV into a respiration- and a stress-driven component.

Considering the **EDA** signal, basic *statistical features* (e.g., mean, standard deviation, min, max) are commonly used, e.g., Setz et al. [2010]. In addition, Koelstra et al. [2012] provide a list of statistical (e.g., average rising time and decay rate) and *frequency domain-based* (spectral power values in the 0-2.4 Hz frequency bands) EDA features. Furthermore, the EDA is known to consist of two components - the skin conductance level (SCL) and skin conductance response (SCR) component. Approaches to separate these components were, for instance, presented by Choi et al. [2012] or Lim et al. [1997]. Following Choi et al. [2012] the degree of linearity of the SCL component was shown to be a useful feature. Considering the SCR component, the identified SCR segments are counted and further statistical features derived: sum of the SCR startle magnitudes and response durations, area under the identified SCRs Healey and Picard [2005]. The SCR-related features were found to be particularly interesting as they are closely linked to high arousal states Kim and André [2008].

From the **EMG** signal, various *time- and frequency-domain* features can be extracted. Christy et al. [2012], working on the DEAP dataset, computed statistical features such as mean, median, standard deviation, and interquartile ranges on the EMG data. Other researchers used frequency-based features such as peak or mean frequencies Kollia [2016], Wijsman et al. [2013]. Another frequently used feature is the signal energy of either the complete signal, see Koelstra et al. [2012], or specific frequency ranges (e.g. 55-95 Hz, 105-145 Hz), as in Abadi et al. [2015]. Wijsman et al. [2013] performed a reference voluntary contraction measurement to compute a personalised EMG gap feature. This feature is defined as the relative time the EMG amplitude is below a specific percentage of the amplitude of the reference measurements.

Soleymani et al. [2012a] pointed out that slow respiration is linked to relaxation. In contrast, irregular and quickly varying **breathing patterns** correspond to more aroused states like, anger or fear Rainville et al. [2006], Kim and André [2008]. Therefore, different respiration patterns can provide valuable information for the detection of affective states. Plarre et al. [2011] describe a number of *time-domain*

Table 2.8: Features commonly extracted and applied in wearable-based AR.

	Features
ACC	**Time-domain:** Statistical features (e.g. mean, median, standard deviation, absolute integral, correlation between axes), first and second derivative of acceleration energy **Frequency-domain:** Power ratio (0-2.75 Hz and 0-5 Hz band), peak frequency, entropy of the normalised PSD **References:** Reiss and Stricker [2012], Figo et al. [2010], Parkka et al. [2007], Mozos et al. [2017]
ECG/ PPG	**Time-domain:** Statistical features (e.g. mean, median, 20th and 80th percentile), HR, HRV, statistical features on HRV (e.g., Root Mean Square of Successive Differences (RMSSD), Standard Deviation of the RR Intervals (SDNN)), number and percentage of successive RR intervals differing by more than 20 ms (NN20, pNN20) or 50 ms (NN50, pNN50), pNN50/pNN20 ratio, **Frequency-domain:** Ultra low (ULF, $0 - 0.003\,Hz$), very low (VLF, $0.003 - 0.03\,Hz$), low (LF, $0.03 - 0.15\,Hz$), and high (HF, $0.15 - 0.4\,Hz$) frequency bands of HRV, normalised LF and HF, LF/HF ratio **Non-linear:** Lyapunov exponent, standard deviations (SD_1 and SD_2) from Poincaré plot, SD_1/SD_2 ratio, sample entropy **Geometrical:** TINN **Multimodal:** Respiratory sinus arrhythmia, motion compensated HR , respiration-based HRV decomposition **References:** Malik [1996], Healey and Picard [2005], Choi et al. [2012], Valenza et al. [2012], Hovsepian et al. [2015], Rubin et al. [2016]
EDA	**Time-domain:** Statistical features (mean, standard deviation, min, max, slope, average rising time, mean of derivative, etc.) **Frequency-domain**: 10 spectral power in the 0-2.4 Hz bands **SCL features**: Statistical features, degree of linearity **SCR features**: Number of identified SCR segments, sum of SCR startle magnitude and response durations, area under the identified SCRs **References:** Lim et al. [1997], Healey and Picard [2005], Setz et al. [2010], Choi et al. [2012], Taylor et al. [2015b], Greco et al. [2016]
EMG	**Time-domain:** Statistical features, number of myoresponses **Frequency-domain:** Mean and median frequency, energy **References:** Kim and André [2008], Wijsman et al. [2010], Koelstra et al. [2012]
RESP	**Time-domain:** Statistical features (e.g. mean, median, 80th percentile) applied to: inhalation (I) and exhalation (E) duration, ratio between I/E, stretch, volume of air inhaled/exhaled **Frequency-domain:** Breathing rate,mean power values of four subbands (0-0.1 Hz, 0.1-0.2 Hz, 0.2-0.3 Hz and 0.3-0.4 Hz) **Multimodal:** Respiratory sinus arrhythmia **References:** Rainville et al. [2006], Kim and André [2008], Plarre et al. [2011], Kukolja et al. [2014], Hovsepian et al. [2015]
TEMP	**Time-domain:** Statistical features (e.g., mean, slope), intersection of the y-axis with a linear regression applied to the signal **References:** Gjoreski et al. [2017], Taylor et al. [2015a]

features which aggregate information about breathing cycles: breathing rate, inhalation (I) and exhalation (E) duration, ratio between I/E, stretch (the difference between the peak and the minimum amplitude of a respiration cycle), and the volume of air inhaled/exhaled. Considering *frequency-domain* features, Kukolja et al. [2014] used mean power values of four frequency subbands (0-0.1 Hz, 0.1-0.2 Hz, 0.2-0.3 Hz and 0.3-0.4 Hz) in order to classify different types of emotions. As discussed previously features relation cardiac and respiratory activities (like RSA are frequently employed Plarre et al. [2011], Hovsepian et al. [2015].

Changes in **body temperature** might be attributed to the 'fight or flight' response (see Section 2.2). During this physiological state, the blood flow to the extremities is restricted in favour of an increased blood flow to vital organs. Hence, temperature-based features can be relevant indicators for a severe stress response. Gjoreski et al. [2017], for instance, extract the mean temperature, the slope, and the intersection of a linear regression line with the y-axis as features.

2.4.3 Classification

In AR the classification is either done using statistical approaches (e.g., ANOVA) or machine learning (ML) methods (e.g., support vector machine (SVM), k-nearest neighbour (kNN)). For both types of analyses, features similar to the ones described in Section 2.4.2 are combined into a feature vector, associated with a label and used as inputs. Since statistical analysis plays only a minor role in wearable-based AR literature, we focus in this section on classification approaches utilising ML techniques. In Table 2.9 the same studies are presented as in Table 2.4. However, here we focus on the employed classification algorithms, number of target affective classes, setting of the study, number of participants, evaluation schemes, and achieved classification performance. The performance is, if possible, reported as *accuracy*, indicating the overall percentage of correctly classified instances. The rest of this section discusses and compares the different approaches and their performance.

The algorithm column in Table 2.9 indicates that the SVM is the most common classification algorithm. It is employed in 48% of the considered studies. This is to some degree surprising as the SVM requires careful adjustment of the kernel size γ and the trade-off parameter C. For this adjustment the recorded data has to be split into *training, validation,* and *test* sets. The best set of hyperparameters can be found by performing a grid-search Hovsepian et al. [2015], Mozos et al. [2017], evaluating the current hyperparameter on the *validation set*. The performance of the final model is then evaluated on the *test set*. Hence, when using a SVM, it is important to report the final test error (and *not* the validation error).

kNN and decision-tree (DT), are the second most popular classifiers both applied in 20% of the considered studies. In comparison to the SVM, kNN and DT require only little hyperparameter tuning and, hence, are applied (almost) in an off-the-shelf way.

Concluding from Table 2.9, ensemble-based methods (e.g., random forest or AdaBoost) are employed less frequently. This is astonishing as ensemble methods have been proven to be strong classifiers. Fernández-Delgado et al. [2014] evaluated 179 classifiers on more than hundred different datasets and found that the random forest family 'is clearly the best family of classifiers'. In the wearable-based AR community, Rubin et al. [2016] for instance employed random forests to detect *panic* and *pre-panic* states, reaching a 97% and 91% accuracy, respectively. Randomized decision trees (ET), introduced by Geurts et al. [2006], are another tree-based ensemble method, considering a random subset of features. In contrast to the random forest, where the most discriminative splits are found, for the ET classifiers the splits are drawn randomly for each feature. In addition, boosting was found to be a strong classifier Fernández-Delgado et al. [2014], and Leo Breiman even considered it to be the 'best off-the-shelf classifier in the world' Friedman et al. [2000]. Mozos et al. [2017] applied the AdaBoost method to detect stress, reaching an accuracy of 94%. For a detailed description of random forests we refer the reader to Breiman [2001] and an introduction into boosting can be found in Freund and Schapire [1999].

Linear discriminant analysis and quadratic discriminant analysis also find application in the work presented in Table 2.9. These classifiers learn a linear or quadratic decision boundary and a detailed description can be found in Bishop [2006].

Fernández-Delgado et al. [2014] also found neural networks (NN) to be among the top-20 classifiers. Haag et al. [2004] and Jaques et al. [2016] used NN, in the form of multi-layered perceptrons, to detect different affective states.

Convolutional neural network and long short-term memory-based classification techniques, which are becoming popular in the field of human activity recognition Hammerla et al. [2016], Münzner et al. [2017], have not found broad application in the domain of wearable-based AR domain yet. Martinez et al. [2013] compare the performance of learned and hand-crafted features to detect the affective states *relaxation, anxiety, excitement* and *fun*. The learned features were extracted using a set of convolutional layers, and the final classification step was performed using a single-layer perceptron. The experiments of Martinez et al. [2013] indicate that learned features lead to an improved classification performance (compared to the hand crafted features). Another advantage of end-to-end trainable classifiers is that they easily facilitate unsupervised pre-training, using for instance auto-encoders. In the image analysis domain auto-encoders are a popular pre-training methods and in the time series classification domain auto-encoders were, for instance, employed by Zheng et al. [2016].

Table 2.9: Comprehensive comparison of algorithms, validation methods, and accuracies of recent wearable-based AR studies. If not stated differently, scores are reported as (mean) accuracy. Abbreviations: Setting (Set.), Lab (L), Field (F), Field with constraint (FC), Validation (Val), cross-validation (CV), Leave-One-Out (LOO), leave-one-subject-out (LOSO), Leave-One-Trial-Out (LOTO), Arousal (AR), Valence (VA), Dominance (DO), Liking (LI), AdaBoost (AB), Analysis of Variance (ANOVA), Bayesian network (BN), deep belief network (DBN), gradient boosting (GB), Gaussian Mixture Model (GMM), hidden Markow model (HMM), linear discriminant analysis (LDA), linear discriminant function (LDF), logistic regression (LR), naive Bayes (NB), neural networks (NN), passive aggressive classifier (PA), random forest (RF), Decision/Regression/Function Tree (DT/RT/FT), ridge regression (RR), quadratic discriminant analysis (QDA)

	Author	Algorithm	Classes	Set.	Sub.	Val.	Accuracy
<2005	Picard et al.	kNN	8	L	1	LOO	81%
	Haag et al.	NN	contin.	L	1	3-fold split	AR: <96%, VA: <90%
	Lisetti and Nasoz	kNN, LDA, NN	6	L	14	LOO	72%; 75%; 84%
2005	Liu et al.	BN, kNN RT, SVM	5	L	15	LOO	74%; 75%; 84%; 85%
	Wagner et al.	kNN, LDF, NN	4	L	1	LOO	81%; 80%; 81%
	Healey and Picard	LDF	3	FC	24	LOO	97%
07	Leon et al.	NN	3	L	8+1	LOSO	71%
2008	Zhai and Barreto	DT, NB, SVM	Bin.	L	32	20-fold CV	88%; 79%; 90%
	Kim et al.	LR	Bin.	FC	53	5-fold CV	$\sim$ 63%
	Kim and André	LDA	4	L	3	LOO	sub. dependent/independent: 95%/70%
	Katsis et al.	SVM	4	L	10	10-fold CV	79%
2009	Calvo et al.	BN, FT, LR, NB, NN, SVM	8	L	3	10-fold CV	one subject: 37%-98%, all subjects: 23%-71%
	Chanel et al.	LDA, QDA, SVM	3/Bin.	L	10	LOSO	<50%; <47%; <50%, Bin. <70%
	Khalili and Moradi	QDA	3	L	5	LOO	66.66%
10	Healey et al.	AB,DT, BN, NB	Bin.	F	19	10-fold CV	None[2]
2011	Plarre et al.	AB, DT, SVM/ HMM	Bin.	L/F	21/17	10-fold CV	82%; 88%; 88%/ 0.71[3]
	Hernandez et al.	SVM	Bin.	F	9	LOSO	73%
2012	Valenza et al.	QDA	5	L	35	40-fold CV	>90%
	Hamdi et al.	ANOVA	6	L	16	-	None [4]
	Agrafioti et al.	LDA	Bin.	L	31	LOO	Active/Pas AR: 78/52% Positive/Neg VA: <62%

	Koelstra et al.	NB	Bin.	L	32	LOO	AR/VA/LI: 57%/63%/59%
	Soleymani et al.	SVM	3	L	27	LOSO	VA: 46%, AR: 46%
2013	Sano and Picard	kNN, SVM	Bin.	F	18	10-fold CV	<88%
2013	Martinez et al.	convolutional neural network (CNN)	4[1]	L	36	3-fold CV	learned features: <75%, hand-crafted: <69%
2014	Valenza et al.	SVM	Bin.	L	30	LOO	VA: 79%, AR: 84%
2014	Adams et al.	GMM	Bin.	F	7	-	74%
2015	Hovsepian et al.	SVM/BN	Bin.	L/F	26/20	LOSO	92%/>40%
2015	Abadi et al.	NB, SVM	Bin.	L	30	LOTO	VA/AR/DO: 50-60%
2016	Rubin et al.	DT, GB, kNN, LR, PA, RF, RR, SVM	Bin.	F	10	10-fold CV	Bin. panic: 73%-97% Bin. pre-panic: 71%-91%
2016	Jaques et al.	LR, NN,SVM	Bin.	F	30	5-fold CV	<76%; <86%; <88%
2016	Rathod et al.	Rule-based	6	L	6	-	<87%
2016	Zenonos et al.	DT, kNN, RF	5	F	4	LOSO	58%; 57%; 62%
2016	Zhu et al.	RR	1	F	18	LOSO	$0.24\pi \approx 43^{\circ}$[5]
2016	Birjandtalab et al.	GMM	4	L	20	-	<85%
2017	Gjoreski et al.	AB, BN, DT, kNN, RF, SVM	3/Bin.	L/F	21/5	LOSO	<73%/ <90%
2017	Mozos et al.	AB, kNN, SVM	Bin.	L	18	CV	94%; 93%; 87%
2017	Taylor et al.	Single/Multitask LR, NN, SVM	Bin.	F	104	Cust.[6]	Mood:<78%, Stress/Health<82%
2017	Girardi et al.	DT, NB, SVM	Bin.	L	19	LOSO	$F1_{AR/VA} < 63.8/58.5\%$
2018	Zhao et al.	NB, NN, RF, SVM	4/Bin.	L	15	LOSO	76%
2018	Santamaria-Granados et al.	CNN	Bin.	L	40	-	Val: <75%, AR<82%
2019	Heinisch et al.	DT, kNN, RF	3	L	18	LOSO	<67%
2019	Hassan et al.	DBN+SVM	5	L	32	10-fold CV	89.53% use DEAP
2019	Kanjo et al.	CNN+LSTM	5	FC	34	User[7]	<95%
2019	Di Lascio et al.	LR, RF, SVM	Bin.	L	34	LOSO	<81%

[1] Given as pairwise preferences.
[2] DT overfit, other classifiers performed worse than random guessing.
[3] Correlation between self-reported and output of model.
[4] No significant differences could be found between the affective states.
[5] Mean absolute error of mood angle in circumplex model.
[6] 80/20% split of the entire data+5-fold CV.
[7] User specific models. Trained random on 70/30% splits with non-overlapping windows.

Considering the setting, three different types of studies are distinguished: *lab* (L), *field* (F), and *field with constraints* (FC) studies. Studies conducted in a vehicle on public roads are referred to as FC studies, as subjects are constrained in their movement. In addition, studies were subjects followed a specific (outdoor) path, e.g., Kanjo et al. [2019] are referred to as FC studies. Most, 29 out of 44, studies presented in Table 2.9, solely base their results on data recorded in a lab setting. The popularity of lab studies is easily explained: In lab studies the study protocol is designed to elicit a set of specific target affective states (see Section 2.3.1). Hence, the signal to noise ratio is much higher than in field studies. Furthermore, once the set of stimuli is chosen the same protocol is applied to multiple subjects, which makes lab studies very efficient. However, models trained on data gathered in constrained environments, are likely to exhibit a poor performance in an less constrained setting.

In order to overcome this, field studies have become more frequent over the past years. This 'out of the lab and into the fray' tendency, coined by Healey et al. [2010], is also related to recent advances in mobile sensor technology and the broad acceptance of smart devices (watches, phones, etc.) among users, see e.g., Rogers and Marshall [2017]. As wearable-based AR clearly aims to detect the affective state users in unconstrained environments, this trend is certainly desirable. Recent work aspiring to detect stress in lab and real life scenarios has for instance been conducted by Taylor et al. [2017], Gjoreski et al. [2017], Hovsepian et al. [2015], Plarre et al. [2011]. Their results indicate that stress detection, based on wearable-based data and context information, is feasible, even in mostly unconstrained settings.

Finally, considering the number of study participants there is large variation: The results reported in Table 2.9 are based on data originating from a single subject up to 104 subjects. Clearly, a large and diversified subject pool is desirable. This would allow to develop generalized models for wearable-based AR.

Judging from Table 2.9, n-fold cross-validation (CV) ($n \in [3, 5, 10, 20, 40]$) is frequently employed as validation method (30%). Following this method, the dataset is randomly partitioned into n equally sized subsets. Then, $n - 1$ subsets are used for training and the remaining one for testing. This procedure is repeated n times. Hence, each of the n subsets is used exactly once as test set. In case the trained model requires hyperparameter tuning, part of the training data can serve as validation set in each iteration. If features are extracted on overlapping windows and n-fold CV is used as validation methods the results are often overoptimistic. This is due to the strong correlation between the features extracted from overlapping windows. Leave-One-Out (LOO) CV is also used in several studies listed in Table 2.9. This is a specific version of the n-fold CV procedure, where n is equals to the total number of available feature vectors. In the LOO case each feature vector is used once for testing. A slightly different type of validation was performed by Abadi et al. [2015]: Leave-One-Trial-Out (LOTO) CV. During LOTO CV, the model is trained on the data of all subjects but leaving one trial/stimulus (e.g. video) aside. The trained algorithm is then evaluated on the left-out data, and the procedure is repeated for

each trial. LOO, LOTO, and n-fold CV lead to subject-dependent results. In order to obtain an subject independent score, corresponding to a more realistic results for real-life deployment, leave-one-subject-out (LOSO) CV should be applied. For this purpose, the algorithm under consideration is trained on the data of all but one subject. The data of the left-out subject is then used to evaluate the trained model. Repeating this procedure for all subjects in the dataset gives a realistic estimate of the model's generalisation properties on completely unseen data. As indicated by Table 2.9, nowadays LOSO CV is widely accepted and applied. From the results shown here, it can be concluded that using the LOSO validation method leads to lower classification scores than applying n-fold or LOO CV. However, only LOSO provides the information on how good the trained model is able to perform on completely unseen data (e.g. data of a new user). Hence, we recommend using this validation scheme.

The affect and stress recognition approaches presented in Table 2.9 report accuracies between 40 % and 95 %. Due to the lack of benchmarking datasets, the results obtained in different studies are hard to compare. On average, the classification accuracies obtained using lab data are higher than the ones obtained in field study data. Hovsepian et al. [2015], who conducted both a lab and a field study, report on a 92 % mean accuracy in detecting stress based on lab data. However, when field data is considered, the accuracy drops to 62 %. Moreover, Healey et al. [2010] conducted a field study and trained different classifiers on the collected data, but none of them was able to perform better than random guessing. This indicates that wearable-based AR in the field is very challenging. As indicated in Table 2.4, most studies were conducted recording multimodal datasets. This might be motivated by a recent review of D'mello and Kory [2015], who pointed out that the classifiers relying on multimodal input reach on average higher classification scores than their unimodal counterparts. Considering the accuracy of classifiers detecting high/low arousal and high/low valence separately it becomes apparent, that arousal is classified more reliably Haag et al. [2004], Valenza et al. [2012], Agrafioti et al. [2012], Abadi et al. [2015]. High arousal states are, from a physiological point of view, directed by the sympathetic nervous system (SNS) (see Section 2.2). Physiological changes directed by the SNS are quite distinct (e.g., increased HR, sweat production, etc.). Hence, detecting high arousal states using these physiological indicators is a feasible task. In contrast, detecting changes in a subject's valence based on physiological data is a more challenging.

The performance of standard ML classifiers depend strongly on the employed features. Hence, the benefits of a careful feature selection can be threefold:

1. Feature selection can help to improve classification results.
2. Feature selection identifies cost-effective and yet strong predictors.
3. It provides a better understanding of the processes generating the data, Guyon and Elisseeff [2003].

According to Guyon and Elisseeff [2003], feature selection methods are grouped into filter-based methods, wrappers, and embedded methods. Filter-based methods select a subset of features (e.g., based on statistical a criterion) and do not take the used classifier into account. Wrapper-based methods (e.g., sequential feature selection) treat the learning algorithm as black box and assess the quality of a subset of features based on the final classification score Guyon and Elisseeff [2003]. Finally, embedded methods perform variable selection during training. Hence, the selection is commonly specific to the used classifier Guyon and Elisseeff [2003]. Feature selection methods also find application in AR. Kim and André [2008], for instance, perform feature selection to improve the classification. Valenza et al. [2012] used principal component analysis to project the features onto a lower dimensional space. This linear method has the advantage that the features are condensed with only a minimal loss of information. For a detailed review of feature selection methods, see Guyon and Elisseeff [2003].

2.5 Conclusion

In this chapter, **RQ 1** (*What is the current state-of-the-art in wearable-based affect recognition?*) has been addressed by providing a comprehensive, detailed, and tutorial-style analysis of the state-of-the-art in wearable-based affect recognition (AR). For this purpose, in Section 2.1 working definitions of the terms affect, emotion, and mood were provided. In addition, different psychological models for emotions and stress were introduced. Furthermore, in Section 2.2 the effect of different affective states on physiological parameters and the sensors employed to measure them were presented. Judging from Table 2.4, sensors monitoring cardiac and electrodermal activity are frequently employed in wearable-based AR. The popularity of these two modalities presumably originates from two considerations: First, high arousal states have a particularly strong effect on both the cardiac and the electrodermal activity. Second, these changes can be easily detected using commercial off-the-shelf devices.

Section 2.3 presents the stimuli frequently employed to elicit different affective states in lab settings. Related work conducted in the field was reviewed in subsection 2.3.2. Questionnaires used for both long-term assessment and in ecological-momentary-assessment (EMA) were presented in Table 2.5 and their scheduling has been detailed in Table 2.6. Based on these two tables it becomes apparent that there is little standardization in terms of labelling procedures for AR studies in the wild. This open point is focused on in the first part of Chapter 4. Moreover, publicly available datasets were presented in Table 2.7. Concluding, from Table 2.7 we identified a lack of commonly used *purely* wearable-based datasets for AR. This limitation is addressed in Chapter 3.

In Section 2.4, the steps following the data collection and leading to classification were detailed. For this purpose modality specific preprocessing (e.g., filtering), segmentation, and commonly computed features were presented. In Table 2.9, the most popular machine learning classifiers and validation methods were summarized. In general, wearable-based AR utilizes mostly classical, feature-based, and supervised approaches. So far only little work has been done, employing semi-supervised training methods or utilizing personalisation methods. Judging from Table 2.9, subject dependent validation schemes like n-fold cross-validation or Leave-One-Out, were found to be the common validation schemes in wearable-based AR. This is to some degree astonishing, as these subject dependent validations schemes are to some degree over-optimistic and do not reflect the generalization properties of the trained classifiers on completely independent data. As a result, subject independent schemes like leave-one-subject-out should be applied. Up-to-date automated feature extraction methods, e.g., convolutional neural networks, have only received little attention in AR. In the second part of Chapter 4, their performance is investigated and compared to classical methods.

3

Study I: Wearable-based Affect Recognition in the Lab

The extensive literature review presented in Chapter 2, identified a lack of commonly used benchmarking datasets for purely wearable-based affect recognition (AR). As detailed in **RQ 2** (*How is benchmarking and direct comparison of different algorithmic approaches for wearable-based affect recognition feasible?*), such a dataset should meet the following demands:

I. The recorded data should originate purely from high quality wearables. The dataset should incorporate physiological "raw" data from multiple redundant sources sampled at high frequencies, ideally recorded from different locations.
II. The desired benchmarking dataset can be recorded in a lab setting but the participants should have some freedom with regards to their postures.
III. The dataset should include data of at least three different affective states. The employed stimuli should be reproducible and lead to distinct affective states.
IV. Data should be recorded from more than 10 subjects.
V. The dataset should be benchmarked using a standard set of features and well-established machine learning classifiers.

This has been addressed by recording, benchmarking, and publishing **WESAD**, a dataset for **We**arable **S**tress and **A**ffect **D**etection[1]. In this chapter, the recording and benchmarking procedure is detailed. The remainder of this chapter is structured in the following way: First, in section 3.1, a quick introduction and overview of related work is provided. Then, in section 3.2, the employed study protocol is outlined. In Section 3.4, the used sensors, their placement, and the extracted features are described. Section 3.4.3 details the obtained results and in section 3.5 we discuss the methods and results critically.

Some passages in this chapter have been quoted verbatim from the following peer reviewed source: Schmidt et al. [2018a].

[1]Publicly available from: https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/

3.1 Related Work

As detailed in Chapter 2, a lot of research has been done aspiring to detect stress and other affective states (e.g., emotions) in different settings. In the body of related work, different approaches to detect emotions and stress based on physiological data can be found (see Table 2.4 and Table 2.9). Due to the severe side effects of stress this emotional state is certainly worth targeting. Distinguishing between stressful and non-stressful states based on physiological data is feasible with high accuracies (at least in constrained environments), see for instance Gjoreski et al. [2016]. However, the physiological responses to high arousal states (e.g., stress) are quite pronounced (see Section 2.2.1). Hence, in order to be of a practical relevance at least three different affective states should be considered. Up-to-date combining stress and emotion detection systems has only received little attention, e.g., Zenonos et al. [2016].

Although there is intensive research in the domain of wearable-based AR, there is only little data publicly available. An overview over these datasets has been provided in Section 2.3.3. However, the list of available datasets narrows if the exclusion criteria, defined in Chapter 1 and the list of demands stated above are applied. Considering the datasets containing different emotions, the "Eight-Emotion" dataset, see Picard et al. [2001], contains multiple affective states but was recorded from just one subject. Further, the "DEAP", "DECAF", "ASCERTAIN", and "MAHNOB-HCI" datasets were all recorded in constrained settings and most presented analyses rely strongly on the electroencephalogram data. Considering USI_Laughs, presented by Di Lascio et al. [2019], the main purpose of the dataset is to distinguish laughter from other events and, hence, it is also not suitable for our purpose. Next considering the datasets containing stress data, the "Driver Stress" dataset, presented by Healey and Picard [2005], contains different modalities but they are all sampled at rather low (15.5-496 Hz) frequencies. Further the "Non-EEG" and "Distracted Driving" dataset contain data, where the participants are subject to different stressors. However, different affective states are not targeted. Finally, the "StudentLife" dataset was recorded using smartphone data only. Hence, a major shortcoming (with respect to this thesis) is that no physiological has been recorded. Based on the considerations detailed above none of the available datasets meets the formulated demands and as a result we chose to record WESAD, a dataset for Wearable Stress and Affect Detection. This dataset combines stress and emotion detection, by containing three different affective states. These states are a neutral, stressed, and an amused state.

3.2 Lab Study Protocol, Self-reports, and Sensors

In this section, the employed lab study protocol, used questionnaires, and the sensory setup are detailed. The goal of the lab study was to elicit three different affective states namely: *neutral, stress*, and *amusement*. After the stress and amusement conditions, the subjects were asked to follow a guided meditation in order to de-excite them. Below the different parts of the study protocol are presented in detail:

1. **Preparation:** Prior to the study, the participants read and signed a consent form. In addition, in the hour before the experiment was to begin, the participants were asked to avoid caffeine and tobacco. Further, the subjects were asked to do no strenuous exercise on the day of the study. Upon arrival at the study location, the participants were equipped with the sensors. Next, a short sensor test was conducted and then the sensor devices (*Emaptica E4* and *RespiBan*) were synchronised manually via a double tap gesture.
2. **Baseline condition:** After the subjects had been equipped with the sensors, a 20 minute baseline was recorded. The baseline condition aimed at inducing a *neutral* affective state, hence, neutral reading material (magazines) was provided.
3. **Amusement condition:** During the *amusement* condition, the subjects watched a set of eleven funny video clips. Each clip was followed by a short neutral sequence of five seconds. Eight of the short clips were chosen from the corpus presented by Samson et al. [2016]. The remaining three videos were chosen by the authors. In total, the *amusement* condition had a length of 392 seconds.
4. **Stress condition:** The subjects were exposed to the well-studied Trier Social Stress Test (TSST), which consists of a public speaking and a mental arithmetic task Kirschbaum et al. [1993]. Reflecting on Section 2.3.1 the TSST incorporates stressors of category **C1** (Social-evaluative nature) and **C2** (Cognitive load). These tasks are known to elicit stress reliably, as they are social evaluative and inflict a high mental load on the subjects Plarre et al. [2011]. In our version of the TSST, the study participants first had to deliver a five minute speech on their personal traits in front of a three-person panel, focusing on strengths and weaknesses. The subjects were told that the three panel members were human resources specialists from our research facility. In order to boost their career options, the subjects, all students at our facility, should try to leave the best possible impression. The study participants had three minutes to prepare their speech but they were not allowed to use their notes during the presentation. After the speech, the panel instructed the subjects to count from 2023 to zero, doing steps of 17. Whenever the subjects made a mistake, they had to start over. For both tasks, the subjects were given five minutes by the panel. Hence, the TSST had a total length of about ten

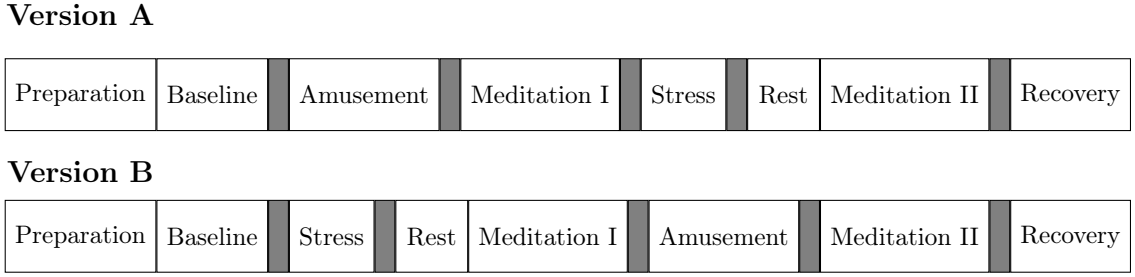


Figure 3.1: The two different versions of the lab study protocol. The gray boxes refer to points in the protocol where self-reports were filed by the subjects.

minutes. After the TSST, the study participants were given a ten-minute rest period.

5. **Meditation:** The amusement and stress conditions both aimed at exciting the subjects, either in a positive or a negative way. These conditions were both followed by a guided meditation. The aim of this meditation was to 'de-excite' the subjects and bring them back to a close to neutral affective state. The meditation was based on a controlled breathing exercise, instructed via an audio track found online. Subjects followed the instructions with closed eyes, while sitting in a comfortable position. The meditation had a duration of seven minutes.
6. **Recovery:** At the end of the protocol, the sensors were again synchronised via a double tap gesture. Then, the sensors were removed. In addition, the subjects were debriefed (e.g., informed that the panel members were not human resources specialists but just 'normal' researchers). Further, during the debriefing period subjects had the opportunity ask detailed questions about the study protocol and its aim.

In total, the study had a duration of about two hours and Figure 3.1 depicts the protocol. As detailed above, our lab protocol features two major stimuli: an *amusement* condition and a *stressful* condition. These two conditions were interchanged (see Figure 3.1) between different subjects in order to avoid effects of order. In addition to these conditions, a baseline and two meditation periods were recorded. In order to induce variance in the subjects' posture, the *baseline, amusement* and *stress* conditions were conducted either standing or sitting. For each condition, approximately half of the subjects were standing and the other half were sitting. During the meditation, however, all subjects were seated.

Self-reports: During the study five self-reports were collected from each participant. Their timing is indicated by the gray boxes in Figure 3.1. Each of the self-reports contained several questionnaires. First, participants filled in a Positive and Negative Affect Schedule (PANAS), which consists of 20 items (ten positive and ten negative items) each rated on a five point Likert scale. PANAS reliably

assesses positive (PA) and negative affect (NA), which are two largely independent dimensions, according to Watson et al. [1988]. PA reaches from 'sad and lethargic' (low value) to 'concentrated and energetic' (high value). NA ranges from 'calmness' (low value) to 'subjective distress' (high value). Furthermore, we added the items *Stressed?, Frustrated?, Happy?*, and *Sad?*, which were scored using the same scale as in the PANAS. These items can be used to generate the same labels as used by Plarre et al. [2011]. Second, similar to Gjoreski et al. [2016], we used six items from the State-Trait Anxiety Inventory (STAI) to gain insight into the anxiety level of the participants. The items were chosen according to their factor loads Barker et al. [1977], and scored on a four point Likert scale. Third, similar to Koelstra et al. [2012] we used the Self-Assessment Manikins (SAM) to generate labels in valence-arousal space. Finally, after the TSST, nine items from the Short Stress State Questionnaire (SSSQ), developed by Helton and Näswall [2015], were added to the questionnaires in order to identify which type of stress (worry, engagement, or distress) was most prevalent in the subjects. The values from these questionnaires can be seen as subjective reports on how the participants felt during a condition.

Physiological Data: For the physiological data collection, we used both a chest- and a wrist-worn device: As depicted in Figure 3.2, the *RespiBan* was placed around the subject's diaphragm. The *RespiBan*[2] itself is equipped with sensors to measure 3-axes acceleration (ACC) and respiration (RESP), and can function as a hub for up to four additional modalities. Using these four analogue ports, electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), and skin-temperature (TEMP) were recorded. All signals acquired by the *RespiBan* were sampled at 700 Hz. The RESP has been recorded via a respiratory inductive plethysmograph (RIP) sensor. ECG data was acquired via a standard three point ECG. For this purpose Ag/AgCl electrodes were employed and located at the following positions: The plus electrode was placed on the center of the chest (approximately 5 cm below the jugular notch of the sternum), the minus electrode on the lower left rib cage, and the ground electrode on lower part of the left abdomen. The EDA data was recorded on the rectus abdominis. This is due to the following consideration: Connecting the *RespiBan* hub to electrodes located on the subjects' palm or finger tips would pose strong limitations on the subjects' ability to move freely. Hence, following a recent literature review from Taylor and Machado-Moreira [2013], which indicates a high density of sweat glands on the abdomen, the EDA data was recorded at this location. The TEMP sensor was placed on the sternum. Similar to Wijsman et al. [2010] the EMG data was recorded on the upper trapezius muscle on both sides of the spine. Using medical tape the wires connecting the *RespiBan* hub to the electrodes/temperature sensors were attached firmly to the subject's torso. The entire *RespiBan* setup is a bit bulky, but the subjects were still able to move freely. A strong advantage of the *RespiBan* is its high sampling rate.

[2]Manufacturer website: https://www.biosignalsplux.com/en

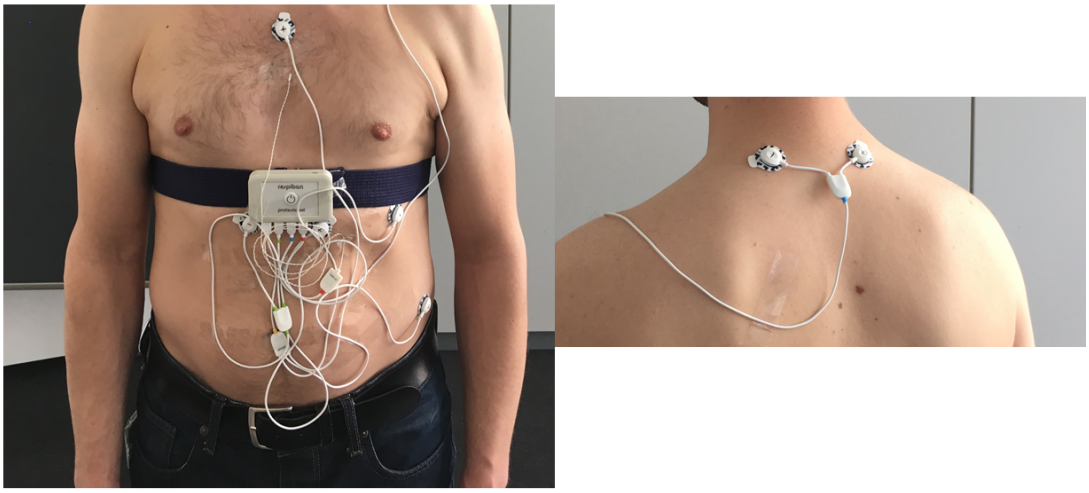


Figure 3.2: Placement of the *RespiBan* and the ECG, EDA, EMG, TEMP sensors.

Furthermore, in future smart fabrics might be able to integrate the sensors housed in *RespiBan*. In order to avoid wireless packet loss, the recorded data was stored locally and transferred to a computer for further processing after the experiments.

In addition to the data acquired using the *RespiBan*, a device with the formfactor of an smartwatch has been employed. This is due to the following observations: First, the *RespiBan* is, considering the size of the hub and attached cables/sensors, quite bulky. This poses strong limitations on the comfort and, hence, we believe that such a setup would not be used by many in real life. In contrast, smartwatches have almost the same formfactor as normal watches and are popular among users. Secondly, comparing the performance of classifiers trained on data originating from different devices is interesting. Especially, as in the desired setup the *RespiBan* incorporates the gold standard sensors (e.g., ECG) sampled at high frequencies, whereas a smartwatch is much closer to a consumer wearable. Due to the broad acceptance of the *Empatica E4*[3] in the AR community, e.g., Gjoreski et al. [2016], Heinisch et al. [2019], or Di Lascio et al. [2019], we chose to employ this device too. In the presented experiment all subjects wore the *Empatica E4* on their non-dominant hand. The *E4* records photoplethysmogram (PPG) (64 Hz), EDA (4 Hz), TEMP (4 Hz), and ACC (32 Hz).

Saliva Probes: In addition to the self-reports and physiological data, saliva probes were collected from the first six subjects. These samples were collected at five different occasions during the above presented lab study protocol:

[3]Manufacturer website: https://www.empatica.com/en-eu/

1. Before the recording of the baseline stated. This will be referred to as Pre-Base.
2. Immediately after the TSST, referred to as Post-TSST.
3. Ten minutes after the end of the TSST (referred to as TSST+10 and collected at the end of the rest period in Figure 3.1).
4. After the meditation, which followed the rest period. This is referred to as Post-Medi.
5. Directly after the amusement condition, called Post-Amusement.

The saliva probes were analysed with regards to their cortisol content, which is a well-established stress hormone Kirschbaum et al. [1993].

Study Participants: Due to the deception necessary for the TSST, only graduate students from our research facility were targeted as participants. The recruitment process happened via Email. Exclusion criteria, stated in the study Email, were pregnancy, heavy smoking, mental disorders, chronic conditions, and cardiovascular diseases. In total, 17 subjects participated in our study. However, due to sensor malfunction, the data of two participants had to be discarded. The remaining 15 subjects had a mean age of 27.5 $\pm$ 2.4 years. Twelve subjects were male and the other three subjects were female.

3.3 Evaluation of Self-reports and Saliva Samples

In this section a qualitative analysis of the study protocol is performed. This analysis is based on the self-reports collected from the participants and on the saliva cortisol measurements. The overall goal of this analysis is to validate the study protocol. Specifically, we evaluated if the employed experimental conditions were suitable to manipulate the subjects' affective state in the desired way.

For this purpose a statistical analysis (mean and standard deviation) of the collected self-reports is displayed in Table 3.1. Comparing the self-reports after the amusement and baseline condition reveals that the amusement condition had the desired effect: The subjects report slightly higher valence and arousal values and less anxiety (STAI), after the amusement condition. However, the effect of this condition is rather small.

In contrast, the impact of the stress (TSST) condition is pronounced: The TSST leads to a strong decrease in the mean valence value and an increase in the mean STAI and arousal values. In addition, the analysis of the SSSQ scores indicates that the subjects felt more engaged and worried than distressed during the TSST task (Engagement: 11.7 ± 2.3, Distress: 6.0 ± 2.9, Worry: 10.6 ± 2.3). The high 'Engagement' score might result from the subjects' high motivation to perform well in the given task. The high 'Worry' score suggests that the subjects were determined to leave a good impression on the panel. In our opinion, these scores demonstrate that most subjects believed the employed cover story of the TSST. After the stress condition, the PANAS showed increased scores with respect to positive (PA) and negative affect (NA). The high PA score indicates that subjects felt energised and concentrated during the TSST, which coincides with the high engagement values reported in the SSSQ. The elevated NA score indicates an increased level of subjective distress. The statistical difference between the baseline and stress conditions were confirmed with the Wilcoxon signed-rank test. Overall, based on the analysis of the collected self-reports, the experimental protocol (especially with respect to the stress condition) is suitable to induce the desired affective states.

Cortisol is a well-known stress hormone. Hence, an increase of the saliva cortisol concentration can be interpreted as an indicator of a stressful event. During the above presented study protocol saliva samples were collected from the first six study

Table 3.1: Evaluation of the questionnaires employed during the Lab study.

	PANAS		STAI	DIM	
	positive	negative		valence	arousal
Baseline	25.5±6.0	12.3±2.0	10.8±1.9	6.7±0.9	2.5±0.9
Stress	31.3±4.7	22.0±6.4	18.5±2.0	4.5±1.6	6.8±1.8
Amusement	25.8±5.1	11.4±2.1	9.3±2.0	7.5± 0.6	3.0±1.6

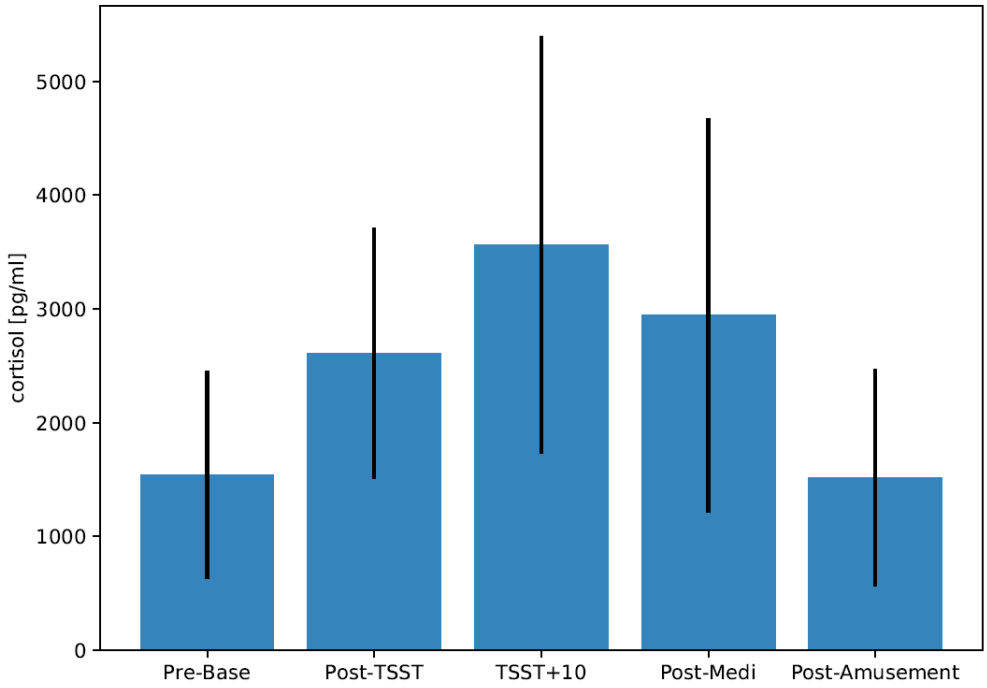


Figure 3.3: Saliva samples obtained from the first six subjects. The cortisol level is plotted over the course of the study protocol, the vertical bars indicate the standard deviation.

participants. Figure 3.3 displays the saliva cortisol level plotted against the different conditions of the lab study protocol. The plot was generated by computing the mean of the saliva cortisol level of the first six subjects for each condition and then arranging the values according to version B of the study protocol (see Figure 3.1). Although the sensor data of the first subject had to be discarded due to sensor malfunction, the saliva probes of this subject are fully available and hence were included in this evaluation.

From Figure 3.3 it becomes apparent that the highest mean saliva cortisol concentration was reached for the TSST+10 measurement. For this measuring point the average saliva cortisol level is more than a factor two larger than mean value collected prior to the baseline recording (indicated by Pre-Baseline in Figure 3.3). In addition, it was found that the saliva cortisol concentration increases between the end of the TSST (referred to as Post-TSST) and the sample collected at TSST+10. These observations are in accordance with the findings of Kirschbaum et al. [1993].

Nevertheless, these saliva cortisol reports have to be interpreted with caution. This is due to the following considerations: First, only from a fraction (N=6) of the study participants saliva samples were collected. Second, for each of these subjects only a five samples were collected. Having these limitations in mind, the analysis of the saliva probes still serves as plausibility check and indicates that the stress condition was working well and that (physiological) stress was elicited successfully.

3.4 Employed Sensors and Feature-based Evaluation

The analysis and evaluation of the WESAD dataset follows the classical time series data processing chain, presented in Section 2.4, consisting of the following steps: data collection, preprocessing, segmentation, feature extraction, and classification Bulling et al. [2014]. In this section focuses on the quantitative analysis of the data. For this purpose the employed features, preprocessing, segmentation, and feature extraction are detailed in a modality specific way. In addition, the recorded dataset is benchmarked and the results are presented in Table 3.3 and Table 3.4.

3.4.1 Feature Extraction

Segmentation of the (preprocessed) sensor signals was done using a sliding window, with a window shift of 0.25 seconds. The 3-axes acceleration (ACC)-features were computed with a window size of five seconds, as similar window lengths are broadly applied for acceleration-based activity recognition (e.g., Reiss and Stricker [2012]). All features (except for statistical- and frequency-domain electromyogram (EMG)-features, see below) based on physiological signals were computed with a window size of 60 seconds. This window size was chosen following recent review by Kreibig [2010]. In Table 3.2, the features extracted from the different modalities are displayed.

On the raw ACC signal different statistical features, e.g., the mean $\mu_{ACC,i}$ and standard deviation $\sigma_{ACC,i}$ were computed. These features were computed both for each axis separately ($i \in \{x, y, z\}$) and as absolute magnitudes, summed over all axes ($3\,D$). In addition, the peak frequency was computed for each axis separately $f^{peak}_{ACC,i}$.

From the ECG and PPG signal features related to the cardiac activity were computed. For this purpose the R-peaks were identified first. In the ECG data the R-peaks were identified with the help of Biosppy [2017]. For the PPG data the minima in the signal were used as surrogate R-peaks. Using these R-peaks, the heart rate (HR) and corresponding statistical features (mean, standard deviation) were computed. Moreover, from the location of the R-peaks the heart rate variability (HRV) was derived, which is an important starting point for additional features. For instance, the energy in different frequency bands (f^{x}_{HRV}) was computed. The frequency bands (x) used, were the very low (VLF: 0.01-0.04 Hz), low (LF: 0.04-0.15 Hz), high (HF: 0.15-0.4 Hz) and ultra high (UHF: 0.4-1.0 Hz) band. In Malik [1996] the HR and HRV are described in detail.

As detailed in Section 2.2, the EDA is controlled by the sympathetic nervous system (SNS). Hence, this signal is particularly sensitive to high arousal states. First, similar to related work by Setz et al. [2010] and Sun et al. [2012], a 5 Hz lowpass filter was applied to the raw EDA signal. Then, statistical features

Table 3.2: List of extracted features. Abbreviations: # = number of, $\sum$ = sum of, STD = standard deviation.

	Feature	Description
ACC	$\mu_{ACC,i}$, $\sigma_{ACC,i}$ $i \in \{x,y,z,3D\}$	Mean, STD for each axis and summed over axes
	$\| \int_{ACC,i} \|$ $i \in \{x,y,z,3D\}$	Absolute integral for each/all axes
	$f^{peak}_{ACC,j}$ $j \in \{x,y,z\}$	Peak frequency for each axis i
ECG and PPG	μ_{HR}, σ_{HR}, μ_{HRV}, σ_{HRV}	Mean, STD of the HR, Mean, STD of the HRV
	$NN50$, $pNN50$	# and percentage of HRV intervals differing by more than 50 ms
	$TINN$	Triangular interpolation index
	rms_{HRV}	Root mean square of the HRV
	f^{x}_{HRV}, $x \in \{VLF,LF,HF,UHF\}$	Energy in different frequency bands of the HRV
	$f^{LF/HF}_{HRV}$	Ratio of LF and HF component
	$\sum^{f}_{x}$ $x \in \{ULF,LF,HF,UHF\}$	$\sum$ the freq. components in ULF-HF
	rel^{f}_{x}	Relative power of freq. component
	LF_{norm}, HF_{norm}	Normalised LF and HF component
EDA	μ_{EDA}, σ_{EDA}	Mean, STD of the EDA signal
	min_{EDA}, max_{EDA}	Min and max value
	∂_{EDA}, $range_{EDA}$	Slope and dynamic range
	μ_{SCL}, σ_{SCL}, σ_{SCR}	Mean, STD of the SCR/SCL
	$corr(SCL,t)$	Correlation btw SCL and time
	$\#_{SCR}$	# identified SCR segments
	$\sum^{Amp}_{SCR}$, $\sum^{t}_{SCR}$	$\sum$ SCR startle magnitudes and response durations
	$\int_{SCR}$	Area under the identified SCRs
EMG	μ_{EMG}, σ_{EMG}	Mean, STD of EMG signal
	$range_{EMG}$	Dynamic range
	$\| \int_{EMG} \|$	Absolute integral
	$\widetilde{\pi}_{EMG}$	Median of the EMG signal
	P^{10}_{EMG}, P^{90}_{EMG}	10th and 90th percentile
	μ^{f}_{EMG}, $\widetilde{f}_{EMG}$,	Mean, median and
	f^{peak}_{EMG}	Peak frequency
	$PSD(f_{EMG})$	Energy in seven bands
	$\#^{peaks}_{EMG}$	# peaks
	μ^{Amp}_{EMG}, σ^{Amp}_{EMG}	Mean, STD of peak amplitudes
	$\sum^{Amp}_{EMG}$, $\bar{\sum}^{Amp}_{EMG}$	$\sum$ and normalised $\sum$ of peak amplitudes
RESP	μ_x, σ_x $x \in \{I,E\}$	Mean, STD of inhalation (I) and exhalation (E) duration
	I/E	Inhalation/exhalation ratio
	$range_{RESP}$, vol_{insp}	Stretch, Volume
	$rate_{RESP}$	Breath rate
	$\sum_{RESP}$	Respiration duration
TEMP	μ_{TEMP}, σ_{TEMP}	Mean, STD of the TEMP
	min_{TEMP}, max_{TEMP}	Min, max TEMP
	$range_{TEMP}$	Dynamic range
	∂_{TEMP}	Slope

were computed (e.g., mean, standard deviation, dynamic range, etc.). In order to separate the tonic (skin conductance level (SCL)) and a phasic (skin conductance response (SCR)) component of the EDA signal, the method proposed by Choi et al. [2012] was applied. After separating the SCL and SCR, additional features, e.g., number of peaks in the SCR ($\#_{SCR}$), were computed. Details about the EDA-related features can be found in Choi et al. [2012] and Healey and Picard [2005].

Two different processing chains were applied to the raw EMG signal. In the first chain, the DC component was removed by applying a highpass filter. Then, the filtered signal was cut into 5-second windows, and statistical and frequency-domain features (e.g., peak frequency) were computed. In addition, the spectral energy ($PSD(f_{EMG})$) was computed in seven evenly spaced frequency bands from 0 to 350 Hz. Following the second processing chain, a lowpass filter (50 Hz) was applied to the raw EMG signal. Next, the processed signal was segmented into 60-second windows. On these windows different peak features, e.g., number $\#_{EMG}^{peaks}$ and mean amplitude μ_{EMG}^{Amp}, were computed. For a more detailed description of EMG-based features, we refer the reader to Wijsman et al. [2010].

Before computing features on the RESP signal, a bandpass filter (cut off frequencies: 0.1 and 0.35 Hz) was applied. Next, a peak detector was used to identify minima and maxima. Following Plarre et al. [2011] the mean and standard deviation of the inhalation/exhalation (μ_I, σ_I, μ_E, and σ_E) were computed. In addition, as also detailed by Plarre et al. [2011], the ratio between inhalation and exhalation (I/E), stretch $range_{RESP}$, inspiration volume vol_{insp}, respiration rate $rate_{RESP}$, and respiration duration were derived $\sum_{RESP}$.

On the raw TEMP signal common statistical features (mean, standard deviation, min, max, etc.) were computed. In addition, the slope of the signal ∂_{TEMP} is used as a feature.

3.4.2 Classification Algorithms and Evaluation Metric

The extracted features, detailed above, serve as input for the classification. Within the benchmark presented below five different feature-based machine learning (ML) classifiers were compared: decision-tree (DT), random forest (RF), AdaBoost (AB), linear discriminant analysis (LDA), and k-nearest neighbour (kNN). As the entire data processing chain was implemented in Python, the scikit-learn implementation, see Pedregosa et al. [2011], of the aforementioned algorithms has been used. For the AB ensemble learner, DTs were used as base estimators. For each of the decision-tree-based classification algorithms (DT, RF, AB), information gain was used to measure the quality of splitting decision nodes, and the minimum number of samples required to split a node was set to 20. The number of base estimators was set to 100 for both of the ensemble learners (RF and AB). Moreover, a LDA and a kNN (with k=9) classifier were used for classification.

We used accuracy and F_1-score as evaluation metrics. Accuracy represents the

number of correctly classified instances out of all samples. The F_1-score is defined as the harmonic mean of precision, indicating the reliability of the results in a certain class, and recall, representing a measure of completeness. To obtain the final F_1-score, precision and recall were computed for each class separately and then averaged. Applying the F_1-score is recommended for unbalanced classification tasks, which is the case when using WESAD (since the various conditions were carried out at different lengths during the study protocol). All models were evaluated using the leave-one-subject-out (LOSO) cross-validation (CV) procedure. Hence, the results indicate how a model would generalise and perform on data of a previously unseen subject.

3.4.3 Classification Results

Based on the features described above classical feature-based ML classifiers are trained and their performance is compared. For the data analysis and evaluation presented here, we only consider the data recorded during the baseline, stress (Trier Social Stress Test (TSST)), and amusement conditions of the study protocol (see Figure 3.1). The analyses presented in section 3.3 indicated that the employed study protocol elicited the desired targeted affective states. Hence, for the quantitative analysis presented below the conditions of the study are used as ground truth. Based on the affective states of the study protocol (baseline, stress, and amusement condition), we distinguish two classification tasks. First, a three-class problem was defined: *baseline vs. stress vs. amusement*. Results on this classification task are presented in Table 3.3. Second, a binary classification task was defined by combining the states *baseline* and *amusement* to a *non-stress* class, posing the *stress vs. non-stress* classification problem. Results of this classification task are presented in Table 3.4. For both classification tasks, 16 different modality combinations are evaluated:

1. Each of the four modalities of the wrist-based device separately (ACC, PPG, EDA, and TEMP).
2. Each of the six modalities of the chest-based device separately (ACC, ECG, EDA, EMG, RESP, and TEMP).
3. All modalities of one device (wrist or chest).
4. All physiological modalities of one device (same as last entry, but without ACC).
5. All modalities from both devices (wrist and chest) together.
6. All physiological modalities from both devices together (same as last entry, but without ACC).

The evaluation was performed using each of the five machine learning algorithms, specified previously. We performed a LOSO evaluation and for each subject each setup (defined by the classification task, applied classifier, and included sensor

modalities) was run five times, to report mean and standard deviation of the evaluation metrics (F_1-score and accuracy). Since LDA and kNN are deterministic classifiers, only the mean values are reported.

The data considered here, belonging to the three affective states of interest, amount to approximately 36 minutes per subject. With 15 subjects and using a sliding window with a shift of 0.25 seconds, approximately 133000 windows were generated. Out of these windows, 53 % belong to the *baseline* class, 30 % represent the *stress* class, and 17 % originate from the *amusement* condition. In the last two rows of Table 3.3 the baseline F_1-score/accuracy of a random and a sophisticated guesser on the three-class problem are displayed. The random guesser is defined to choose one of the three possible classes at random, thus reaching an accuracy of 33 % and a F_1-score of 32 %. In contrast, the sophisticated guesser would always choose the majority class. Hence, a sophisticated guesser would reach an accuracy of 53%. However, its' F_1-score would only be 32 %. In the two last rows of Table 3.4, the same types of random and sophisticated guesser are presented for the binary classification task.

Comparing the performance of the employed algorithms, on the three-class task (Table 3.3) and binary classification task (Table 3.4), it becomes apparent that the ensemble-based methods (RF, AB) and the LDA reached similar classification scores. Depending on the input modalities, these classifiers reach scores up to 80 % for the three-class problem and up to 93 % for the binary task, respectively.

Concluding from Table 3.3 and Table 3.4, the kNN had the overall worst performance, reaching accuracies of at most 60 % on the three-class problem, and 78 % in the binary task.

Using only motion-based features (wrist and/or chest ACC) leads to considerably lower classification scores compared to results obtained using physiological features. This suggests that the physiology-based features provide a deeper insight into the affective states of the subjects than the motion patterns. Moreover, we can rule out the possibility that our classifiers only learned to distinguish between motion patterns or postures characteristic for the conditions of the protocol.

In the three-class problem the accuracies using one of the wrist-based physiological modalities range from 59 % to 70 %. Using one of the physiological chest-based modalities on the same classification problem, accuracies between 54 % and 72 % are reached. In the binary classification task the accuracies using a wrist-based input modality range from 69 % to 86 % and the accuracies using one of the chest-based modalities range from 67 % to 88 %. In both classification tasks the RESP is a particularly strong chest-based modality leading to the best result of a single modality. Besides the stress-related changes in the respiration, this can be partially explained considering the fact that the study participants spoke during the TSST. Hence, the classifiers might have partially learned to distinguish between speaking (stress condition) and non-speaking episodes (baseline and amusement condition). In both classification tasks, using only the TEMP data, either chest or wrist-based,

Table 3.3: Evaluation of the given modalities and classifiers on the **three-class** (*baseline vs. stress vs. amusement*) classification task. Abbreviations: decision-tree (DT), random forest (RF), AdaBoost (AB) DT, linear discriminant analysis (LDA), k-nearest neighbour (kNN)

	DT		RF		AB		LDA		kNN	
	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]
Motion:										
ACC wrist	43.91 ± 1.16	53.71 ± 0.91	46.50 ± 0.26	56.40 ± 0.16	46.38 ± 0.64	$\mathbf{57.20 \pm 0.57}$	36.27	47.73	37.20	45.54
ACC chest	42.18 ± 0.4	51.14 ± 0.29	41.96 ± 0.29	53.48 ± 0.29	44.28 ± 0.75	$\mathbf{56.56 \pm 0.70}$	34.61	48.84	31.07	40.29
Wrist:										
PPG	51.15 ± 0.31	57.57 ± 0.22	53.83 ± 0.11	64.09 ± 0.12	53.29 ± 0.16	64.46 ± 0.21	54.72	**70.17**	50.97	59.44
EDA	45.48 ± 0.17	54.36 ± 0.27	45.74 ± 0.06	56.57 ± 0.05	49.06 ± 0.59	59.85 ± 0.42	42.72	**62.32**	45.20	54.98
TEMP	41.46 ± 0.24	47.42 ± 0.36	41.85 ± 0.19	48.67 ± 0.21	41.19 ± 0.24	49.39 ± 0.23	40.89	**58.96**	38.97	44.32
Wrist physio	57.13 ± 0.86	63.34 ± 1.00	66.33 ± 0.36	$\mathbf{76.17 \pm 0.42}$	64.24 ± 0.39	73.62 ± 0.55	58.18	68.85	50.85	58.54
Chest:										
ECG	51.69 ± 0.35	57.81 ± 0.36	52.24 ± 0.33	60.36 ± 0.22	52.48 ± 0.38	61.71 ± 0.40	56.03	**66.29**	47.77	54.76
EDA	43.88 ± 0.20	48.49 ± 0.29	42.40 ± 0.55	45.00 ± 0.61	48.33 ± 0.31	54.06 ± 0.45	46.83	**67.07**	37.26	40.03
EMG	34.65 ± 0.21	41.00 ± 0.19	38.10 ± 0.47	48.20 ± 0.51	37.68 ± 0.24	48.03 ± 0.24	37.72	**53.99**	35.97	42.73
RESP	59.08 ± 0.21	65.97 ± 0.20	60.69 ± 0.15	70.27 ± 0.14	61.76 ± 0.34	71.94 ± 0.30	60.09	**72.37**	45.86	60.45
TEMP	41.27 ± 0.29	47.53 ± 0.28	42.46 ± 0.24	48.40 ± 0.26	40.76 ± 0.8	47.98 ± 0.60	30.96	**55.68**	35.18	43.32
Chest physio	55.10 ± 0.92	58.62 ± 1.07	64.60 ± 0.54	71.37 ± 0.58	72.51 ± 0.17	$\mathbf{80.34 \pm 0.43}$	74.43	79.35	51.09	57.31
All wrist	43.62 ± 1.33	53.98 ± 1.79	62.86 ± 0.65	74.85 ± 0.20	64.12 ± 0.98	$\mathbf{75.21 \pm 0.77}$	63.24	70.74	37.20	45.54
All chest	53.06 ± 0.50	57.68 ± 0.40	60.80 ± 1.00	68.76 ± 1.35	64.89 ± 0.81	74.74 ± 0.94	72.49	**76.50**	38.39	46.18
All physio	55.71 ± 0.93	62.57 ± 0.80	64.23 ± 0.97	73.33 ± 0.95	71.10 ± 0.78	$\mathbf{79.86 \pm 0.62}$	72.48	78.19	52.94	59.61
All modalities	58.05 ± 1.61	63.56 ± 1.73	64.08 ± 1.68	74.97 ± 1.11	68.85 ± 0.89	$\mathbf{79.57 \pm 0.93}$	71.56	75.80	48.70	56.14

Baseline	Random Guessing		Sophisticated guessing	
	F_1-score	Accuracy	F_1-score	Accuracy
	31.66	33.33	23.13	53.12

Table 3.4: Evaluation of the given modalities and classifiers on the **binary** (*stress vs. non-stress*) classification task.

	DT		RF		AB		LDA		kNN	
	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]	F_1-score in [%]	Accuracy in [%]
Motion:										
ACC wrist	55.36 ± 0.47	64.08 ± 0.49	59.02 ± 0.78	69.96 ± 0.55	61.70 ± 0.80	$\mathbf{71.69 \pm 0.45}$	44.93	60.02	52.72	63.80
ACC chest	61.92 ± 0.83	71.75 ± 0.53	59.91 ± 0.25	72.87 ± 0.08	62.17 ± 0.45	$\mathbf{73.87 \pm 0.30}$	57.52	72.05	47.79	57.81
Wrist:										
PPG	78.27 ± 0.17	81.39 ± 0.15	81.35 ± 0.15	84.18 ± 0.11	81.23 ± 0.15	84.10 ± 0.13	83.08	**85.83**	78.94	82.06
EDA wrist	70.95 ± 0.37	76.21 ± 0.27	70.88 ± 0.20	76.29 ± 0.14	75.34 ± 0.57	$\mathbf{79.71 \pm 0.43}$	69.86	78.08	68.30	73.13
TEMP wrist	63.15 ± 0.18	68.22 ± 0.19	62.90 ± 0.10	67.82 ± 0.11	62.27 ± 0.25	67.11 ± 0.34	56.37	**69.24**	60.18	64.46
Wrist physio	82.37 ± 0.21	84.88 ± 0.11	86.10 ± 0.29	$\mathbf{88.33 \pm 0.25}$	85.86 ± 0.20	88.05 ± 0.18	83.77	86.46	78.93	81.96
Chest:										
ECG	77.01 ± 0.37	80.17 ± 0.29	79.64 ± 0.15	82.78 ± 0.11	80.20 ± 0.25	83.37 ± 0.20	81.31	**85.44**	75.39	79.19
EDA chest	69.88 ± 0.41	73.55 ± 0.44	73.63 ± 0.18	77.51 ± 0.23	71.97 ± 0.26	75.50 ± 0.29	74.51	**81.70**	66.64	69.73
EMG	47.06 ± 0.20	56.25 ± 0.05	49.42 ± 0.35	63.44 ± 0.18	50.84 ± 0.44	62.88 ± 0.31	52.49	**67.10**	51.84	58.74
RESP	79.92 ± 0.19	83.03 ± 0.17	84.33 ± 0.10	86.63 ± 0.08	84.64 ± 0.06	86.87 ± 0.06	85.61	**88.09**	69.17	75.67
TEMP chest	57.40 ± 0.08	64.33 ± 0.07	56.75 ± 0.25	64.75 ± 0.28	55.03 ± 0.27	63.46 ± 0.21	41.00	**69.49**	51.64	58.25
Chest physio	81.29 ± 0.22	84.18 ± 0.20	90.44 ± 0.66	92.01 ± 0.51	87.11 ± 0.57	89.76 ± 0.48	91.47	**93.12**	77.27	81.05
All wrist	78.71 ± 0.53	82.19 ± 0.44	84.11 ± 0.31	$\mathbf{87.12 \pm 0.24}$	80.11 ± 0.93	83.98 ± 0.75	84.05	86.88	52.72	63.80
All chest	78.26 ± 0.46	81.29 ± 0.38	90.04 ± 0.84	91.70 ± 0.75	89.57 ± 0.61	91.58 ± 0.46	91.07	**92.83**	64.20	69.70
All physio	83.03 ± 1.61	85.16 ± 1.50	86.02 ± 0.55	87.91 ± 0.54	87.78 ± 1.38	89.77 ± 1.17	90.93	**92.51**	79.44	83.16
All modalities	80.83 ± 1.13	83.60 ± 1.08	85.71 ± 0.63	87.74 ± 0.60	83.88 ± 0.93	87.00 ± 0.78	90.74	**92.28**	69.14	74.20

Baseline	Random Guessing		Sophisticated guessing	
	F_1-score	Accuracy	F_1-score	Accuracy
	47.96	50.00	41.15	69.94

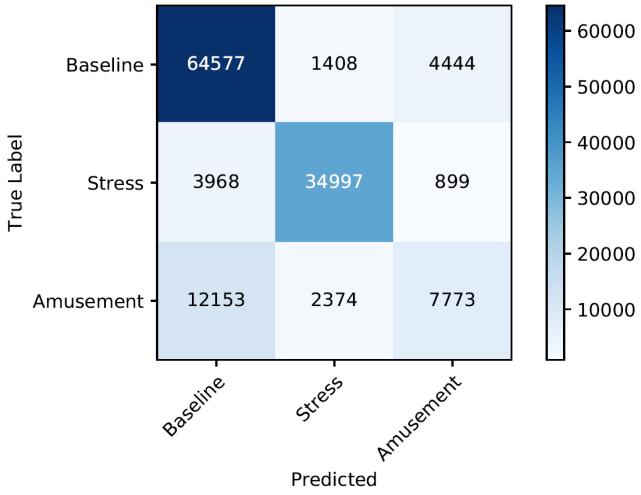


Figure 3.4: Confusion matrix of the best setup (AB classifier trained using the chest-based physiological features) for the three-class problem.

as input leads to low classification scores. This shows that, TEMP is not a well-suited modality to solely base the classification of affective states upon. Comparing the results obtained using only the wrist- or chest-based EDA data, the latter seems to hold more relevant information leading to somewhat higher accuracies in both classification tasks. In contrast, comparing the performance of classifiers solely relying on the PPG or ECG data, the former leads to slightly higher accuracies. The results reached using all physiological chest-based modalities (three-class accuracy: 80 %, binary accuracy: 93 %) are higher than the ones obtained using all physiological wrist-based modalities (three-class accuracy: 76 %, binary accuracy: 88 %). When both wrist- and chest-based physiological modalities are combined, an accuracy of 79 %/92 % is reached for the three-class/binary problem, respectively. This is no improvement compared to results achieved using only the chest-based physiological modalities. This indicates that if the chest-based modalities are available, the wrist-based modalities become redundant. Nevertheless, the classification scores reached using only the physiological wrist-based modalities are very promising, especially considering the minimal intrusive nature of the device used. Overall, the best performance result (in terms of accuracy) on each of the classification task is:

- 80.34 % (three-class problem, using all chest-based physiological modalities, AB classifier)
- 93.12 % (binary case, using all chest-based physiological modalities, LDA)

These results are comparable to the work of Gjoreski et al. [2016], who reported an accuracy of 72 % on a three-class problem (no, low, and high stress) and an accuracy of 83 % in the binary case.

In Figure 3.4, the confusion matrix of the best setup, AB classifier trained using the chest-based physiological features, on the three-class classification problem is displayed. The values indicate that the classifier was able to distinguish well between the baseline and the stress class. However, distinguishing between the classes baseline and amusement was difficult. The explanation for this is twofold. First, the physiological changes elicited by amusement are subtle. Second, the self-reports indicate (see Table 3.1) that the subjects' affective state was less influenced by the amusement condition compared to the stress condition.

Using all physiological features and the DT classifier, the subject-specific accuracies range from 69% to 98% and from 82% to 100%, in the three-class classification problem and the binary case, respectively. However, only weak correlations were found between the subject-specific accuracies and the difference between the selfreports of the corresponding conditions (e.g., $Arousal_{Stress} - Arousal_{No\,Stress}$). Nevertheless, the large inter-subject differences emphasise the need for personalisation methods.

In order to assess the feature importance, a DT was trained for both the three-class and the binary classification task, using all available features. The feature importance is computed using the Gini criterion. The results of this experiment are displayed in Table 3.5. In both cases (three-classes and binary classification) the two most important features ($\sigma_E^{RESP,}$, and μ_{HR}^{ECG}) were alike. This suggests that the classifier in the three-class problem first learned to distinguish between *stress* and *non-stress* states, before it learned to classify the *baseline* and *amusement* classes.

Table 3.5: Feature importance for the three-class and binary classification task considering all modalities.

Importance	Three-class	Importance	Binary Task
0.23	$\sigma_E^{RESP,\,chest}$	0.35	$\sigma_E^{RESP,\,chest}$
0.11	$\mu_{HR}^{ECG,\,chest}$	0.20	$\mu_{HR}^{ECG,\,chest}$
0.07	min_{TEMP}^{wrist}	0.09	max_{TEMP}^{wrist}
0.06	$\mu_{ACC,\,3D}^{chest}$	0.07	$range_{EDA}^{wrist}$
0.05	$range_{EDA}^{wrist}$	0.05	$\#_{SCR}^{chest}$

3.5 Conclusion

In this chapter, a detailed description and analysis of WESAD, a dataset for multimodal wearable stress and affect detection, has been provided. Section 3.1 outlined related work, highlighting other available datasets and their limitations (e.g., relying strongly on EEG data). In Section 3.2, the study protocol has been presented and the its effectiveness was verified using the collected self-report data (see Section 3.3). Further, saliva samples collected from a subset of participants were used as additional validation of the lab study protocol. The analysis of these saliva samples indicated that the Trier Social Stress Test had successfully induced stress in the study participants. In Section 3.4 the employed sensors, their placement, derived features, and the quantitative classification results were presented.

In research question **RQ 2** (*How is benchmarking and direct comparison of different algorithmic approaches for wearable-based affect recognition feasible?*), five demands for a wearable-based benchmarking dataset were defined. These are all met by WESAD:

I. In contrast to other available datasets, WESAD contains all physiological modalities commonly integrated in commercial and medical devices: photoplethysmogram (PPG), electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RESP), skin-temperature (TEMP), and 3-axes acceleration (ACC). The data was recorded in a redundant fashion and at high sampling rates (up to 700 Hz). By using these modalities, we hope that our dataset will enable and support the development and evaluation of new affect recognition systems.
II. Although data was acquired in a lab setting, the sensor setup allowed the participants to move as freely as possible. In addition, during the data recording the effect of different postures (standing and sitting) was investigated.
III. The study protocol was tailored to elicit three different affective states (neutral, stress, amusement), using a set of well studied stimuli. This enables not only the detection of stress related arousal, but also of valence.
IV. WESAD contains data of 15 study participants.
V. A large set of standard features and classifiers has been used to benchmark WESAD.

In the benchmarking experiments both a three-class (*baseline vs. stress vs. amusement*, accuracy 80 %) and a binary (*stress vs. non-stress*, accuracy 93 %) were targeted. These results are promising, however, due to the limitations of WESAD, regarding the number of subjects, their age, and gender diversity, these results should be interpreted with caution. Nevertheless, since using the leave-one-subject-out (LOSO) evaluation scheme, the results indicate that generalisation is possible.

During the benchmarking of WESAD, a detailed analysis on the importance of the two device locations as well as the different sensor modalities has been performed.

In the experiments presented before, EDA, ECG, PPG and RESP were particularly strong modalities. ACC, EMG, and TEMP, however, were only of minor importance. Considering the three-class problem, the overall highest classification results (accuracy of 80 %) was archived using the chest-based physiological features (without the motion information). Adding data of the wrist-based device lead to no further improvement. Using only physiological data recorded from the wrist a three-class accuracy of 76 % was reached. This result is very promising and motivating. It indicates that using data gathered via a minimally intrusive device with the formfactor of a smartwatch can be used to detect affective states surprisingly well. For future (field) studies it seems that the burden of the quite bulky chest-based sensor can be lifted by using a wrist worn device.

In this chapter, the feasibility of affect recognition purely based on physiological data in the lab has been demonstrated. However, the laboratory environment represents a constrained setting. This is mainly due to the following considerations:

1. **Available Stimuli:** In a lab setting strong emotional stimuli, like videos or the Trier Social Stress Test are easily integrated into the study protocol. These stimuli are well known from psychological research (e.g., Kirschbaum et al. [1993], Lang et al. [1999], Samson et al. [2016]). Hence, the elicitation of certain affective states is feasible.
2. **Labelling:** Due to the known starting and end points of the different conditions, the ground truth is easily generated in a lab study (e.g., using the study protocol). In addition, questionnaires can be used after each stimulus to generate additional labels.
3. **Data quality:** Due to the constrained setting, the sources of noise which are present in the acquired physiological data can be reduced to a minimum.

The clear goal of wearable-based affect recognition is to detect a person's affective state in unconstrained environments. Hence, an appropriate next step is the conduction of a field study. However, the above listed points concerning stimuli, labelling, and data quality are more challenging in the field. So, as pointed out in related literature, e.g., Plarre et al. [2011], a decrease in the performance of the affect recognition systems is to be expected.

4

Study II: Wearable-based Affect Recognition in the Field

Many users of commercial mobile devices are interested in automatically logging information related to their physical health, e.g., counting steps, or tracking consumed/burned calories. Recently, a first generation of commercial devices promising insights into mental health, by detecting stress[1] [2], entered the consumer market. Providing users with such data-driven insights into their, especially negative, affective states could increase awareness and lead to an overall health improvement.

As presented in Chapter 3, the detection of different affective states can be done to a satisfying extent in lab environments, reaching high classification scores. This has been observed both related in literature, e.g., Wijsman et al. [2010], Plarre et al. [2011], and Soleymani et al. [2012b], and in own work (see Chapter 3). The overall goal of affect recognition (AR) is to detect affective states in unconstrained environments. Hence, a desirable next step is to investigate the performance of machine learning methods detecting different affective states from data and labels gathered in the field. For this purpose research question **RQ 3** (*What is the performance of machine learning systems that detect multiple affective states in unconstrained environments?*) has been formulated and a field study dataset, containing affective labels and physiological data, has been recorded.

In this chapter, the above stated research question is targeted and the remainder of this chapter is structured as follows: First, in section 4.1, a brief recapitulation of the current state-of-the-art in wearable-based AR field studies is provided. For a more detailed description please refer to Section 2.3.2. In section 4.2, the field study protocol, used sensors, and the labelling tool employed to target research question **RQ 3** are presented. In addition, paradigms and guidelines for the development of labelling tools are formulated and lessons learned are provided in Section 4.3. This

[1]Product website: www.apple.com/apple-watch-series-4/health/
[2]Product website: https://buy.garmin.com/en-US/US/p/567813

directly addresses research question *RQ 3a* (*What is an appropriate way to label affective states in everyday life reliably?*). The second part of this chapter (section 4.4) addresses research question *RQ 3b* (*What is the performance of classifiers trained on labels generated with an ecological-momentary-assessment tool?*) and presents different quantitative analyses. For this purpose, the labels generated via the labelling tool as well as the physiological and motion data acquired during the field study are used to train different machine learning classifiers. The employed classifiers range from classical feature-based algorithms to end-to-end trainable convolutional neural networks.

Some passages in this chapter have been quoted verbatim from the following sources: Schmidt et al. [2018b] and Schmidt et al. [2019a].

4.1 Related Work

Experiments conducted by Bower [1981] indicate that human decision making and memorization are strongly linked to their affective state. Hence, in order to build holistic user models, the reliable detection of affective states in everyday life is key. As discussed in Section 2.4.3, there has been a focus shift from lab to field studies in the AR community.

In general, the common approach to AR is data-driven: Given some sort of input data (e.g., physiological signals), machine learning models are trained to assess the affective state of a person. Wearables, like smartphones and watches, facilitate out-of-the-lab AR. This is due to three reasons: First, their passive, unobtrusive, and ubiquitous sensing capability, second, their computational power, and third, their broad acceptance by a large number of users. From a technical point of view, smartphones offer an ideal platform to collect data and labels in the wild, see for instance Healey et al. [2010], Muaremi et al. [2013], Gjoreski et al. [2016], Zenonos et al. [2016], and Taylor et al. [2017] . However, in field studies no objective ground truth, e.g., condition in a study protocol, is available. Hence, AR studies in the wild rely solely on self-reports of the participants and these self-reports have to be used in lieu of a protocol-based ground truth. These self-reports are often gathered via ecological-momentary-assessments (EMAs), a method to assess the *momentary* affective state of subjects in their natural environment using a set of questionnaires.

Table 4.1 presents a subset of recent AR field studies relying on EMAs. Due to the severe health implications of long-term stress, e.g., increased risk of cardiovascular diseases identified by McEwen and Stellar [1993], most of these studies focus on stress detection, see Gjoreski et al. [2016], Hovsepian et al. [2015], or Muaremi et al. [2013]. Emotions and mood, see Healey et al. [2010], Sano et al. [2015], Zenonos et al. [2016], and Taylor et al. [2017], were targeted less frequently in AR field studies.

Table 4.1: Overview of recent AR field studies.

Author	year	Target States
Muaremi et al. [2013]	2013	Stress
Hovsepian et al. [2015]	2015	Stress
Gjoreski et al. [2017]	2016	Stress
Healey et al. [2010]	2010	Emotion
Sano et al. [2015]	2015	Mood
Zenonos et al. [2016]	2016	Mood
Taylor et al. [2017]	2017	Mood & Stress forecasting

4.2 Field Study Protocol

The number of AR studies conducted in the wild is growing. Contributing to this body of work, we conducted an AR field study collecting physiological, motion, context, and affective data. The goal of this study was to investigate the performance of multimodal real-life affect recognition systems. In this section, the study protocol, and the employed labelling tool will be presented.

Study Protocol and Sensory setup: Physiological data was collected using the *Empatica E4*[3] smartwatch. This choice is due to the following three considerations:

1. The *E4* has the formfactor of a standard smartwatch, incorporates multiple sensors, and the "raw" data is accessible. Furthermore, the *E4* (unlike the *RespiBan*) does not require the placement of electrodes and/or the connection of any cables. Hence, using the *E4* as measurement device reduces the burden for the user to a minimum, while ensuring high data quality.
2. Based on the lab study results presented in Chapter 3, data acquired by the E4 can be used to classify the affective state to a satisfying extent (see Table 3.3).
3. The *E4* (or *Affectiva Q sensor*) has been used in many studies with a similar target, e.g., Gjoreski et al. [2016], Taylor et al. [2017], Gashi et al. [2018]. Hence, using the *E4* to collect physiological data makes our results comparable to related work.

The *E4* houses four different sensor types: 3-axes acceleration (ACC) (32 Hz), electrodermal activity (EDA) (4 Hz), photoplethysmogram (PPG) (64 Hz), and skin-temperature (TEMP) (4 Hz). In total, physiological data of 12 healthy users (all students at our facility) was recorded. The subjects were instructed to wear the *E4* on their non-dominant hand during their wake hours. During potentially harmful activities for the *E4* (e.g., showering, washing the dishes, etc.) the subjects were asked to take off the *E4*. However, during such a period participants were told that the *E4* should not be switched off. Hence, the subjects recorded up to 17 hours of physiological and motion data every day. The outline of the data collection protocol is depicted in Figure 4.1. During the week, the study leader met the study participants twice a day:

1. Once the subjects arrived at the research facility in the morning, the study leader met the participant to exchange the *E4*. During this procedure, the subject received a new *E4*, which was then worn for the remainder of the day. The study leader downloaded the data from the device worn during the prior day.
2. Around noon/early afternoon, the study leader and subject met again for a screening. During this screening, the study leader and the participant examined the collected data visually and a structured interview was conducted in order to gain additional context information.

[3]Manufacturer website: https://www.empatica.com/en-eu/

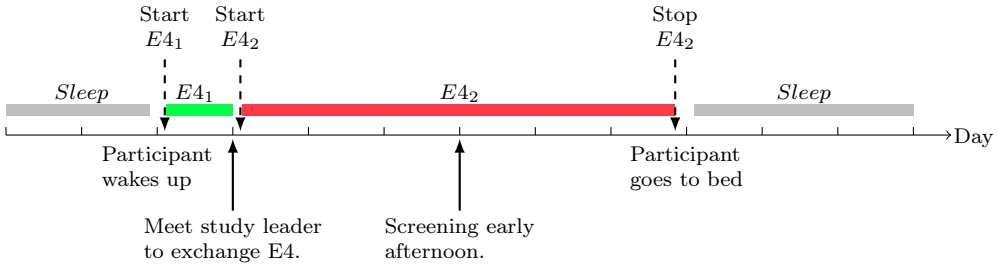


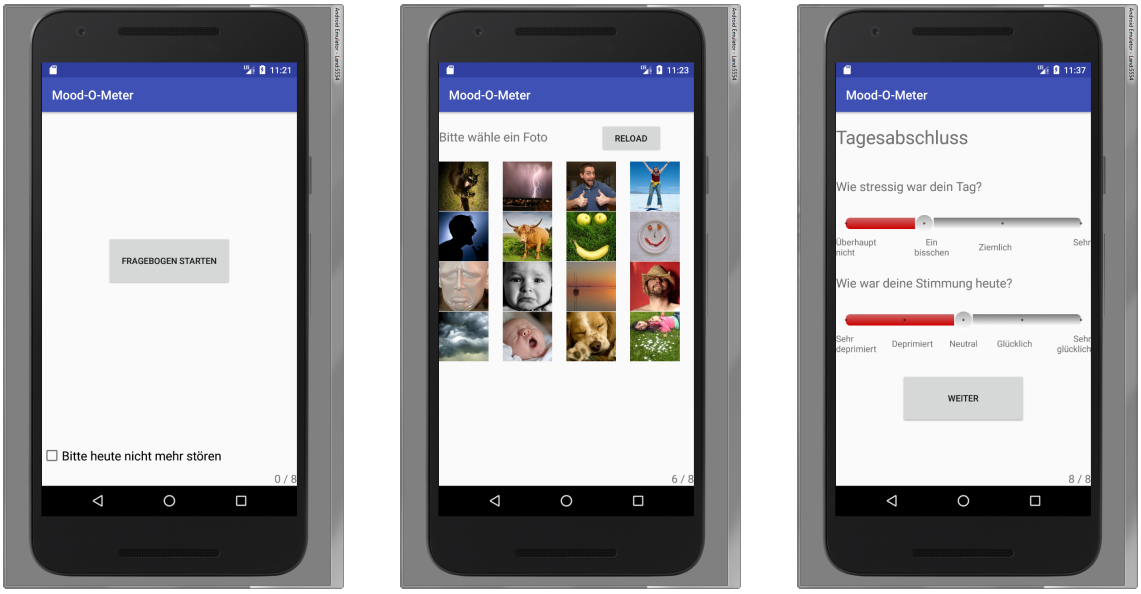
Figure 4.1: Weekday outline of the field study protocol.

During the weekend, this procedure was not feasible. Hence, on Friday the subjects received a second *E4* and were instructed to use it for the data collection on Sunday. Consequently, on Mondays the study leader collected two *E4s* from the participants and during Monday's screening the data from Friday, Saturday, and Sunday was inspected.

Context Logging and Labelling Tool: In order to aggregate context data, the subjects installed a so called ContextLogger app on their smartphone. This ContextLogger was designed to acquire context information in a passive fashion. Passive, in this case, means that no interaction between the subject and app was required. However, the ContextLogger was running for the entire period of the study on the subject's smartphone as a background application. The ContextLogger was used to aggregate contextual information like weather, location, screen events, and activities. As this context information is rather sensitive, the employed data logging consent form allowed the subjects to select the recorded modalities individually. All recorded contextual information was time-stamped and uploaded to a server at the end of a each day.

Prior to the study, all participants completed a Perceived Stress Scale (PSS) and a Pittsburgh Sleep Quality Index (PSQI) questionnaire. However, in order to capture the subjects' affective states on a regular basis, an Android EMA app was developed and deployed on the subjects' smartphone. All subjects were fluent in German, hence, the app was developed in German. The app incorporated several (shortened, well-established) questionnaires:

- The Self-Assessment Manikins (SAM), found in Morris [1995], and the Photo Affect Meter (PAM), developed by Pollak et al. [2011], (exemplary screen depicted in Figure 4.2b) were used to generate labels in valence-arousal space.
- One screen in the app was dedicated to emotional categories. Here the subjects had the choice between the basic emotions, see Ekman and Friesen [1978], namely anger, fear, surprise, happiness, disgust, and sadness or were able to select "None of them".
- Similar to Gjoreski et al. [2016] a shortened STAI was used and similar to Plarre et al. [2011] the subjects reported their current stress level on a four point Likert scale.



(a) First screen. (b) Photo Affect Meter Illustration. (c) Final question of the day.

Figure 4.2: Exemplary screens from the developed EMA app.

- Subjects reported the intensity of the physical activity they had been pursuing during the past 10 minutes.

During an initial face-to-face meeting, the functionality of the E4 and both apps (EMA app and ContextLogger) were explained to the subjects. Using the EMA app the subjects filed automatically and manually triggered self-reports on their affective states. Explanatory screens are displayed in Figure 4.2. For each subject, the EMA app was customised to match their diurnal rhythm. During the configured time span (e.g., 7.30 to 22.30) the EMA app was triggered automatically approximately every 2 hours. After such a trigger the subjects received a notification that an EMA should be filed. In addition to the standard questionnaires incorporated in each set of EMAs, the first and last set of questionnaires included an additional screen:

- **First questionnaire of the day:** Similar to Sano et al. [2015], the subjects were asked about their sleep duration and quality.
- **Last questionnaire of the day:** Like in Muaremi et al. [2013], the subjects were asked about their overall stress and mood level throughout the day.

As depicted in Figure 4.2a, the subjects had the freedom to terminate the EMA app by selecting the "Please no more disturbances for today" button. This button paused the remaining automatically scheduled questionnaires and added the final question screen (see Figure 4.2c) to the current set of questionnaires. During the

initial face-to-face meeting, the subjects were also instructed to trigger an EMA manually when they felt a change in their affective state. At the end of each day, the labels acquired using the EMA app were automatically uploaded to a server.

Ethics and Demographic information: The study was approved by both the workers council and the data security officer of our research facility. Participants were mostly recruited via Email and all participants were students. In total, 12 healthy subjects (7 male, 5 female) participated in the study. Due to sensor malfunction we had to exclude one participant (female). Hence, further analysis will be based on the remaining 11 subjects (mean and standard deviation age: 26 ± 2.5, mean and standard deviation participation duration: 16 ± 1 days).

4.3 EMA Guidelines, Implementation Details, and Lessons Learned

As no 'best practice' guidelines for wearable-based AR field studies are available, the ways these studies have been conducted are diverse. This makes the direct comparison of the results difficult. In this section, we provide practical guidelines for AR field studies, focusing especially on frequent and high quality affective labels generated via EMAs. In order to ensure optimal objectivity, reliability, and validity of EMA data, four paradigms for EMA-based labelling tools are formulated:

Paradigms for ecological-momentary-assessment in AR field studies

1. **Intrusiveness:** EMAs should be only minimally intrusive.
2. **Autonomy:** Subjects can decide when to file an EMA.
3. **Redundancy:** Multiple sources facilitate validity/plausibility checks.
4. **Motivation:** Motivated subjects file more EMAs.

Following these paradigms, seven guidelines for the design and application of EMA apps are provided below. In addition, the implementation of these guidelines into the presented AR field study is detailed. Based on the 1248 EMAs collected during the study the effectiveness of the formulated guidelines is evaluated and lessons learned are formulated.

1) Sampling Rate and Scheduling:
Guideline: The trade-off between overloading and sampling the affective state of a subject as frequently as possible needs to be balanced. In literature, scheduling an EMA every two hours, Zenonos et al. [2016], or approximately five times a day, Gjoreski et al. [2016], seems to be adequate.
Implementation: In accordance with Zenonos et al. [2016], we chose to trigger an EMA automatically every $120 \pm x,\ x \in (0 < x < 30)$ minutes. The lag x was introduced to add randomness to the sampling points. Following an automatic trigger, the subjects were notified that they should file an EMA. If subjects did not complete the EMA within 30 minutes after the trigger event, they received a second notification. However, the subjects had the freedom to ignore these notifications and file the EMA some time later.
Insights & lessons learned: Figure 4.3 displays the distribution of EMAs filed over a day. Our sampling rate ensures a mostly even distribution of EMAs. In addition, none of our subjects reported to feel overloaded by the number of scheduled EMAs. The deviations in the number of completed EMAs at the beginning (6.00-9.00) and end (21.00-23.00) of the day can be explained by the differences in the diurnal rhythm of the subjects. From Figure 4.3 it becomes apparent that many EMAs were completed around 22.00. This is easily explained by the fact that many subjects customized the EMA to not

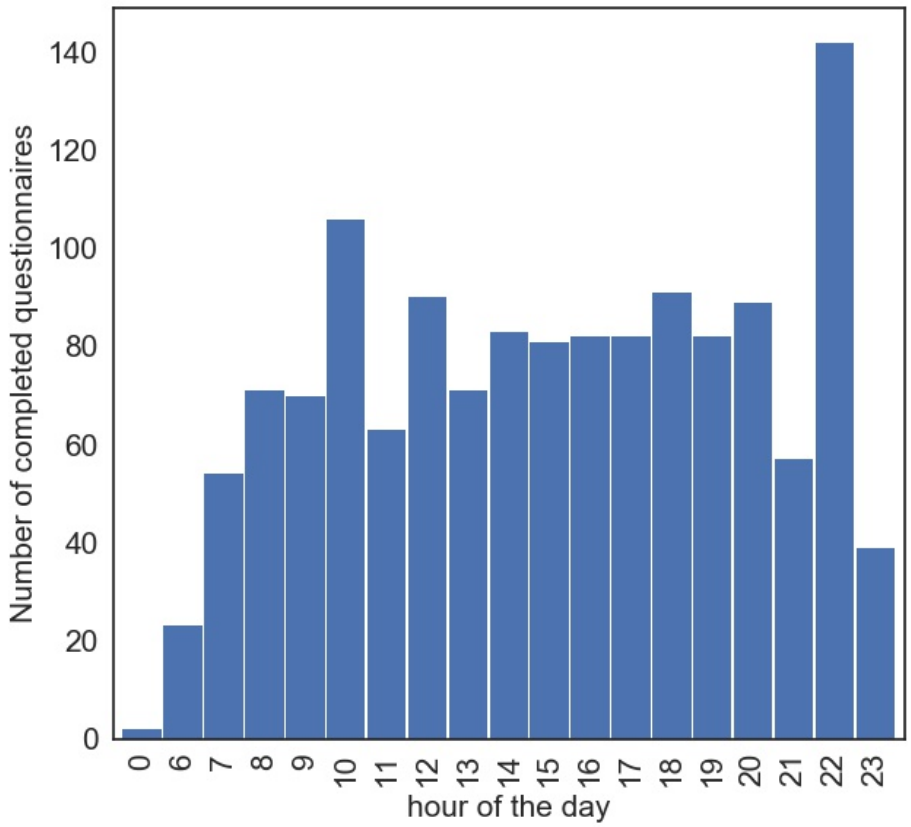


Figure 4.3: Distribution of questionnaires filed over a day.

be triggered after 22.00 and, hence, completed the last EMA of the day around 22.00.

2) Filing Time and Number of Items:
Guideline: EMAs should target the core goal of the study, and they should include as few items as possible. For example, Muaremi et al. [2013] report that they had to reduce their EMA to 10 items after receiving complaints of the study participants.
Implementation: While reporting various aspects of affective states (even in a redundant fashion), the number of items is kept as low as possible, e.g., by reducing the number of STAI items. In addition, all questions could be answered with a single click (no free text or audio reports were necessary).
Insights & lessons learned: In our study the median filing time of an EMA was 40 seconds. As none of our subjects complained about the EMA length,

Table 4.2: Comparison of automatically and manually triggered EMAs, with regards to the basic emotion label.

	Automatic scheduled	Manually triggered	Total
EMAs	880	368	1248
With Basic Emotion	204	126	330

we believe that both filing time and EMA length were appropriate.

3) Manual Trigger of EMAs by Subject:
Guideline: Since automatically triggered EMAs are completely independent of the affective state of the subjects, the chance of missing "interesting" events is high. Due to memorization effects, e.g., the perception of the event under consideration is influenced by the current affective state, labelling these "interesting" events in hindsight is difficult. Hence, in addition to randomly scheduled EMAs, subjects should also have the freedom to trigger EMAs manually (e.g., by opening the EMA app).
Implementation: In our field study, the subjects were encouraged to trigger an EMA whenever they felt an a change in their affective state by simply starting the EMA app. In order to avoid overlapping labels and ensure adequate spacing between the selfreports, after a manual trigger, the subsequent automatically triggered EMA was postponed.
Insights & lessons learned: Table 4.2 summarises the total number of EMAs filed during our field study. For most EMAs, the subjects reported no basic emotion (by selecting the "None of them" button in the corresponding screen). In addition, comparison of the absolute valence and arousal values shows higher valence and arousal values for manually triggered EMAs. Further, the fraction of reported "basic emotions" to "None of them" is substantially higher in the manually triggered EMAs (34 vs 23%). In our opinion this bias is plausible as discrete emotional are more constrained in their resolution. Hence, mixed or ambiguous emotions are captured less reliably using a discrete model, see for instance Eerola and Vuoskoski [2011]. Overall, these results suggest that manually triggered EMAs contain reports on more intense and less ambiguous emotional states. This supports our recommendation to allow the manual trigger of EMAs.

4) Validity and Redundancy of EMAs:
Guideline: Self-reports are subjective. However, using well-established questionnaires increases the validity of the results and enables a comparison to other studies. In addition, using multiple questionnaires assessing similar constructs (e.g., basic emotions and points in valence-arousal space) offers the possibility to check the EMA values for consistency.

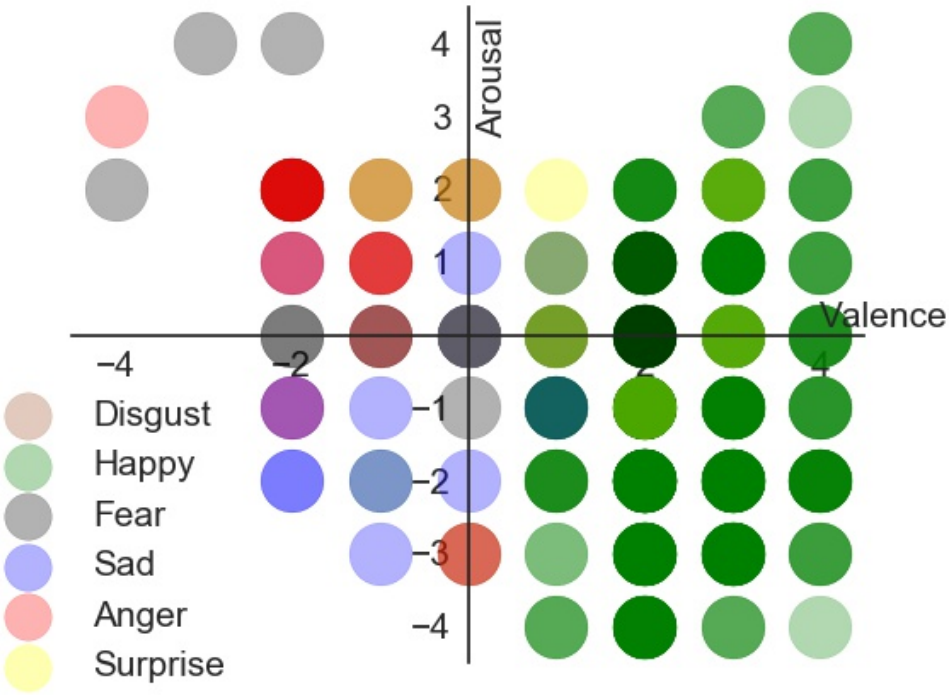


Figure 4.4: Basic emotions of all subjects mapped to valence-arousal space.

Implementation: During the presented field study several well-established scales (e.g., SAM, PAM, State-Trait Anxiety Inventory (STAI)) and a list of basic emotions were used to generate affective labels. In addition, subjects reported their stress level.
Insights & lessons learned: In Figure 4.4, the basic emotions reported by all study participants are mapped into the valence-arousal space. Plots like these help to facilitate plausibility checks: For instance, the subjects reported the basic emotion 'Happy' when having a positive valence. However, 'Happy' seems to be not affected by the arousal value. In contrast, subjects only reported 'Anger' and 'Fear' when being in a high arousal and low valence state. 'Sadness' was mostly reported when the subjects were in a low valence and low arousal state. This redundancy helps to check the labels for plausibility and the depicted mapping is in accordance with other research (e.g., Eerola and Vuoskoski [2011]).

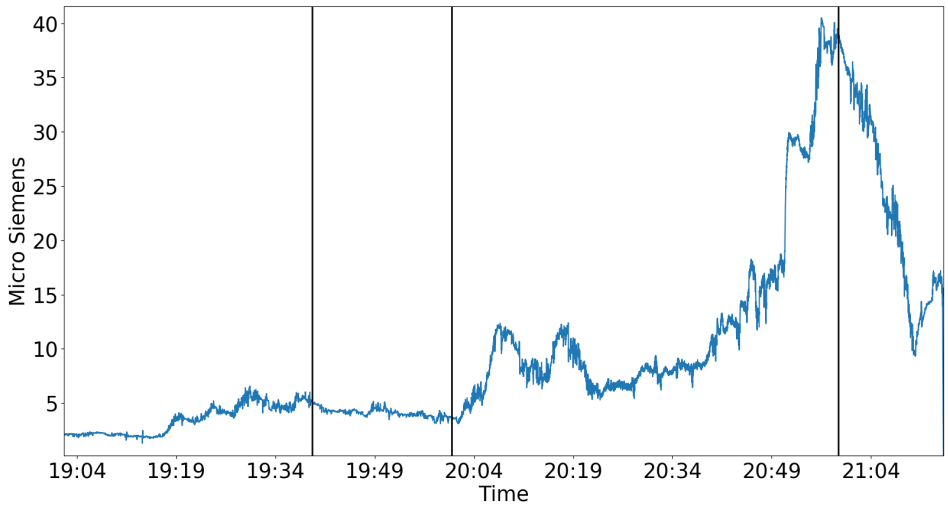


Figure 4.5: Electrodermal activity data of a subject prior and during a workout session. Vertical lines correspond to filed EMAs.

5) Context Information:
Guideline: In previous work, e.g., Gjoreski et al. [2017], Sano et al. [2015], it has been shown that physical activities and sleep quality are important context information in the domain of AR. Hence, we recommend to record this data either automatically, e.g., using the Android Activity Recognition API, or as part of the EMAs.
Implementation: In our field study, context information was gathered both automatically and manually. First, the ACC data, recorded by the *E4*, can be seen as some sort of context information as it can be employed to perform an activity classification or alternatively estimate the intensity of the activity. Second, using the subjects' smartphones, a number of context information was acquired both in a passive and an active fashion: Using location-based services (e.g., weather information) and the Android Activity Recognition API, context information was logged automatically. In addition, each EMA incorporated a question on the physical intensity of the activity pursued during the past 10 minutes. Moreover, the first EMA of the day included items on sleep quality and duration.
Insights & lessons learned: Based on the above detailed implementation the recorded dataset contains different forms of context data.

6) Daily Data-Driven Screening:

Guideline: Understanding field data in hindsight is often difficult. Therefore, related work suggests to conduct daily screenings to assess data quality, see Healey et al. [2010], Hovsepian et al. [2015], Muaremi et al. [2013], Sarker et al. [2016].

Implementation: Data-driven screenings on weekdays were an integral part of this field study. During these screenings, a structured interview (asking the the subjects about their health, workout sessions, etc.) was conducted. Plots of EMAs label, motion, and physiological data were used to understand the circumstances of important situations.

Insights & lessons learned: The plots helped to gather further context information on major physical and mental events of the day. Figure 4.5 displays the EDA of a subject during a workout. One immediately notices the strong increase in EDA values. Spotting events like this and incorporating them (as notes) into the structured interview provided a deep insight into labels and raw data. In addition, the screenings also allowed data quality assessment on a regular base. Hence, a reduced data quality (e.g., due to misplacement of the *E4*) would become apparent timely and could be corrected by re-instructing the subjects.

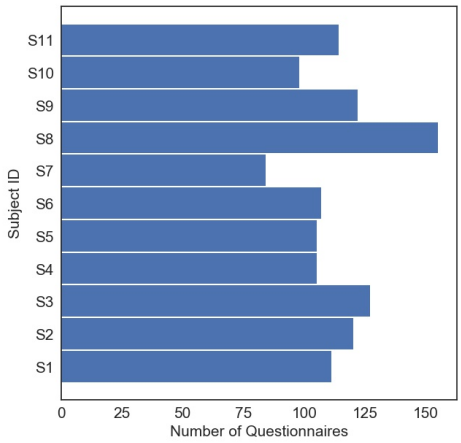
However, a drawback of these daily screenings is that they are rather time-consuming (between 30 and 45 min. per subject). Considering a large scale study with a large number of participants and/or a long duration (e.g., three months) this procedure quickly becomes infeasible. One way to mitigate this issue could be to use an active labelling approach: In such a setup, a trigger algorithm could be used to prompt the subject with an EMA whenever a certain set of parameters enters an unusual state, e.g., dramatic increase of the EDA value without a change in the ACC signal. In order to develop such trigger algorithms data acquired from lab studies could be used.

7) Subjects' Commitment:

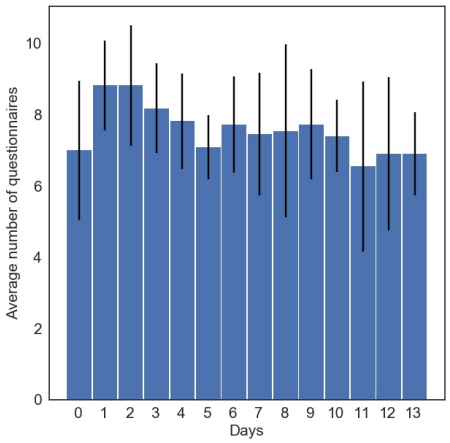
Guideline: In order to motivate study participants to file EMAs, incremental reward systems as in Healey et al. [2010], or the chance to win an additional price via a lottery can be employed, e.g., Wang et al. [2014]. Following Berkel et al. [2017], another way to increase subjects' motivation is the use of gamification. Keeping the subjects motivated will ensure high-quality labels, regarding both frequency and completeness.

Implementation: In the conducted field study, every participant received a base reward (20 € gift card for the completion of two full days). Further, among the five participants providing the most EMAs, two were selected randomly to receive an additional price.

Insights & lessons learned: Figure 4.6a displays the total number of EMAs completed by each subject. Apart from S7 and S10, all of our study par-



(a) Histogram of the total number of questionnaires filed per subject.



(b) Average number of EMA field by subjects over course of the study.

Figure 4.6: Plots indicating the subjects' motivation.

ticipants filed more than 100 EMAs during their participation. Subject S8, who participated the longest in the study, filed the highest total number of questionnaires. In Figure 4.6b the average number of filed EMAs per day of the study is displayed. It indicates that the number of filed EMAs stayed almost constant over the course of the study. Hence, it can be concluded that the participants stayed motivated and that the employed incentive system was working well.

Based on participants' feedback, two additional guidelines can be formulated:

(i) *Allow Hindsight Labelling:*
During a intense affective event, e.g., stressful exam, it is difficult or even impossible to complete an EMA. Relying on the subjects' autonomy, allowing short hindsight labelling could be beneficial to further improve label completeness and quality. Further, allowing the subjects to adjust the time span of a label, e.g., entire duration of the exam or just the final 15 min., could also help to increase label accuracy.

(ii) *Incorporate Reviewing Possibility:*
Some of the study participants pointed out that the option to review items and change the selection would have been beneficial. Relying again on the autonomy of the subjects, this issue could be mitigated by providing a back button in the EMA app. Or displaying the final selection/scores before terminating the EMA app and storing the questionnaire values.

However, as the guidelines i and ii were not incorporated in the EMA app employed during our AR field study, no implementation details or lessons learned can be provided.

4.3.1 Discussion

In order to make the design of EMA tools for affect recognition (AR) field studies more comparable, four paradigms, regarding *intrusiveness, autonomy, redundancy* and *motivation*, were formulated at the beginning of section 4.3. Based on these paradigms seven guidelines were defined and implemented in the presented AR field study protocol and EMA-based labelling tool. These guidelines refer to:

1 Sampling Rate and Scheduling
2 Filing Time and Number of Items
3 Manual Trigger of EMAs by Subject
4 Validity and Redundancy of EMAs
5 Context Information
6 Daily Data-Driven Screening
7 Subjects' Commitment

Using the data and insights gained during the conducted field study the effectiveness of the guidelines has been evaluated critically and lessons learned were formulated. The above presented analyses indicated that the guidelines were successfully implement into our field study. A key finding is that (at least for the studied cohort of participants) one can rely on the subjects' *autonomy* and their *motivation*. This finding is, for instance, emphasised by the subjects' feedback, which lead to two additional guidelines:

i Allow Hindsight Labelling
ii Incorporate Reviewing Possibility

However, as these analyses are of a descriptive nature, the evaluation is no strict proof of the guidelines. Nevertheless, the guidelines can be understood as a first approach to standardize EMA-based labelling tools. As the field of wearable-based AR is growing rapidly, we hope that the presented findings are helpful for other researchers. Due to the limitations of the study, considering gender and age, these guidelines have to be understood as pointers and we encourage the community to modify them based on their findings.

In the next section the recorded labels, physiological, and motion data shall be used to train affect recognition systems, detecting the affective state based on these indicators.

4.4 Quantitative Analysis of the Field Study Data

Based on the data gathered during our AR field study, presented in section 4.2, a quantitative analysis of the data is performed in this section. The objective here is to detect the affective state of a subject given only physiological and motion data. The affective states considered in this analysis are arousal, STAI, stress, and valence values. For this purpose both feature-based and end-to-end trainable classifiers are employed.

4.4.1 Considered Data and Label

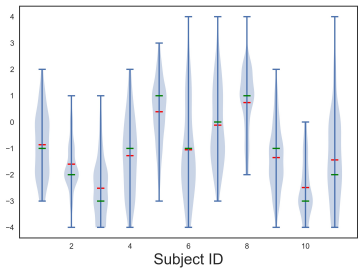
As detailed in section 4.2, labels have been generated by the subjects utilizing a self-developed EMA-based labelling tool. For the quantitative analysis presented here the following subset of labels is considered:

- **Valence** and **arousal** labels generated using the well-known self-assessment mannequins Morris [1995].
- A shortened version (six items) of the **STAI** Spielberger et al. [1983]. Items were chosen according to their factor loads, and scored on a four point Likert scale.
- **Stress** level scored on a four point Likert scale Gjoreski et al. [2017].

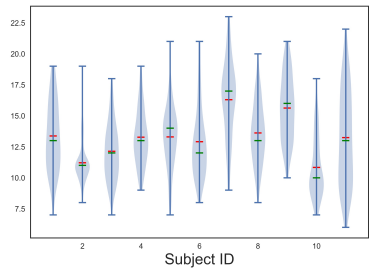
The basic emotion labels are not considered for two reasons: First, in only about 26% of the available EMAs contain a "real" basic emotion (see Table 4.2). Second, as discussed by Eerola and Vuoskoski [2011] basic emotions are coarser than the corresponding reports in valence-arousal space. Further, the labels generated via the PAM are also not considered in this work. This is due to the fact that the PAM also generates labels in valence-arousal space, which makes them redundant to the ones generated using the SAM. The reason for relying on the SAM and not on the PAM is their finer granularity.

For the quantitative analysis presented below the physiological and motion data generated using the *E4* will be utilized as inputs. As the motion data is strongly linked to the activities the subjects are performing this can be seen as a form of context information. As stated in the exclusion criteria detailed in Chapter 1, smartphones were excluded as sensing modality. Hence, the context information gathered using the sensors integrated into the subjects' smartphones will not be considered in this analysis. During the data acquisition this data was, nevertheless, recorded as it might be analysed in future work.

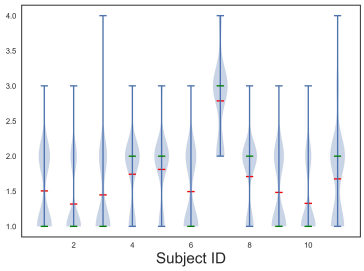
Using all EMAs filed by the subjects during the study period violin plots of the considered label (arousal, STAI, stress, and valence) distributions of each subject are depicted in Figure 4.7. These plots highlight strong inter-subject differences: S4, S5, and S7, for instance, tend to be more stressed than the other participants.



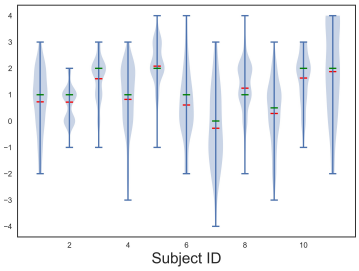
(a) Subject specific arousal values.



(b) Subject specific STAI values.

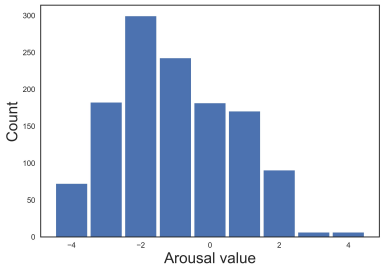


(c) Subject specific stress values.

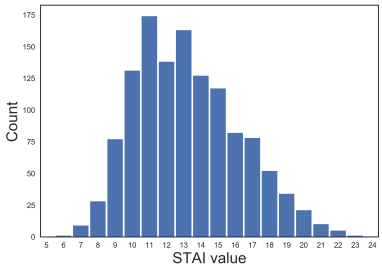


(d) Subject specific valence values.

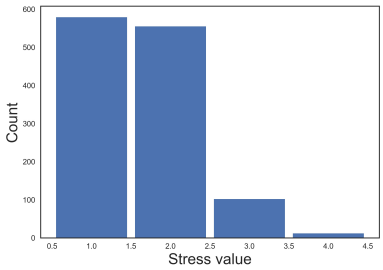
Figure 4.7: Violin plots depicting the label distributions for each label type and subject. Mean value displayed in red. Median value shown in green.



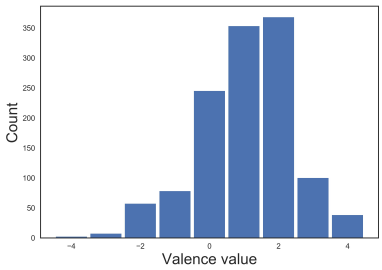
(a) Arousal Values.



(b) STAI values.



(c) Stress values.



(d) Valence values.

Figure 4.8: Cumulated arousal, STAI, stress, and valence values generated during the field study using our EMA app.

Further, the accumulated label distributions of all subjects are plotted in Figure 4.8. From these plots the skewness of the corresponding label distributions (arousal, stress, STAI, and valence) becomes apparent: In general, the histograms exhibit only little mass in the high arousal, STAI, and stress bins. Judging form the valence scale, the subjects are more often in a positive than in a negative valence state.

Using Pearson's R, a correlation analysis has been performed for the labels. A strong negative correlation (-.60) between the STAI and valence values was found. Furthermore, a moderate positive correlation was found between the arousal and STAI values (.44) and a strong correlation is also observed between the stress and STAI values (.68). The above detailed correlations are significant (2-tailed p-values < 0.001). We found no correlation between valence and arousal labels. This finding emphasises that valence and arousal are independent scales.

Windowing and questionnaire binning

For the quantitative analysis presented below, we labelled the time period from $X - 600$ sec to $X + 655$ sec with the affective states reported in the EMA started at time point X. The additional 55 seconds account for the completion of the entire set of questionnaires and was empirically verified prior to the data collection. In the next step, these time spans were segmented using a sliding window. Following, a review by *Kreibig*, the window size for the segmentation was set to 60 sec Kreibig [2010]. The window shift was 5 sec. Hence, for each valid time period, 240 windows were extracted.

During the segmentation we excluded time periods where the *E4* was either not worn or one of the sensors had a malfunction. Questionnaires with incomplete data in the corresponding interval have been rejected, too. After the above detailed cleaning procedure, a total of 1083 valid questionnaires were retained.

We formulated a three class classification problem for the arousal, STAI, and valence labels. The below detailed bins were chosen in order to establish equally sized bins for the three class classification tasks. Arousal was binned according to the following scheme: low (-5, -2], medium (-2, 1], and high (1, 5]. For valence the

Table 4.3: Number of questionnaires in the different bins.

	Low	Medium	High
Arousal	479	519	85
STAI	479	539	65
	negative	neutral	positive
Valence	56	593	434
	No stress	Stressed	
Stress	504	579	-

same bins were used, yielding a negative, neutral, and positive class. The employed STAI thresholds are (5, 12], (12, 18], and (18, 25]. Considering the skewness of the stress distribution (Figure 4.8c), a binary classification problem was formulated. For this purpose the labels with a value equal to 1 represent the "No stress" class. The other (2, 3, 4) were used to represent the "Stressed" class. In Table 4.3 the number of valid questionnaires per bin are displayed.

4.4.2 Evaluation Method and Metric

In order to validate the machine learning (ML) approaches, we employed leave-one-subject-out (LOSO) and leave-target-questionnaires-out (LTQO) as validation schemes. For LTQO a stratified N-fold 80%/10%/10% (Train/Test/Validation) split was performed using all valid questionnaires. The stratified nature of these splits ensures similar label distributions in the different sets. However, in contrast to simple N-fold cross-validation, this scheme ensures that all instances (segmented windows/features) belonging to specific *target questionnaires* are placed in the same set (e.g., training). Hence, this validation scheme provides an insight into the subject dependent performance of the different classifiers. This scheme has been employed to mitigate the large individual differences in the label distributions displayed in Figure 4.7.

Due to the skewness of the considered dataset the macro F_1 score, corresponding to the unweighted mean of the F_1 scores for the different labels/classes, is used as evaluation metric. The performance of the different ML classifiers is compared to the performance of a sophisticated guesser (also known as Zero Rule). This classifier always predicts the majority class found in the training data. Later, an investigation of different types of classifiers (feature-based and end-to-end learning) is presented and their performance is compared.

4.4.3 Classification Algorithms

Classical Approach: For the classical experiments we followed the human activity recognition pipeline presented by Bulling et al. [2014] and extracted features from windowed data (size 60 sec, shift 5 sec). In total 62 features were extracted and used as input for the classical classifiers. We used the same set of *E4* features as described in Table 3.2. These features range from plain statistical features (mean and standard deviation), to complex physiological features like heart rate, heart rate variability, or number of peaks in EDA data. Here, two different experiments were performed: In the first experiment, the features were used directly as input. In the second, a z-transformation, normalizing each feature to zero mean and unit variance has been applied. For the classical evaluation the sklearn, see Pedregosa et al. [2011], implementation of different tree-based classifiers (decision-tree (DT), randomized decision trees (ET), and random forest (RF)) have been used.

For ensembles (RF and ET) the number of trees were chosen to be N=101. In order to avoid overfitting, the minimal number of samples per split was set to 150 for all classical classifiers.

End-to-End Learning: In the end-to-end learning scenario, the windowed data served as direct input into convolutional neural networks (CNNs). Below we present an approach for affect recognition based purely on physiological and motion time series data utilizing CNNs. Starting with the single-task CNN (ST-CNN) formulation we extend this approach to a multi-task CNN (MT-CNN) classifier, predicting arousal, STAI, stress and valence simultaneously. The multitask approach is motivated by the correlations found between the different label distributions. The ST-CNNs and MT-CNNs architectures investigated here utilize four layer types: convolutional, max-pooling, global-average pooling, see Lin et al. [2013], and fully-connected (FC) layers. The CNNs receive the windowed *E4* data (ACC, EDA, PPG, TEMP) as input. As the PPG data has been down sampled by a factor of two, the CNNs deals with two sampling frequencies (4/32 Hz).

Feature Extraction: The CNNs employ sensor-based late fusion, see Münzner et al. [2017]. This enables the network to learn modality-specific filters. The feature extraction part of the CNNs is depicted in Figure 4.9. The architectural parameters (e.g., kernel size, stride, etc.) were chosen to be the same in branches with the same sampling frequency (ACC+PPG and EDA+TEMP) and a grid search has been performed to identify appropriate settings. In each of these branches convolution and max-pooling layers are alternated. Table 4.4 details the hyperparameters used in the feature extraction branches. Throughout the network, RELUs are employed as non-linearities. After the feature extraction a global average pooling operation is performed.

Classification: Both the ST-CNN and MT-CNN approach utilize the feature extraction architecture described above. The main difference between the two approaches lies in the classification part of the network: The ST-CNN uses a separate

Table 4.4: Overview of the numbers of convolutional layers in the different feature extraction branches and the corresponding parameters. Abbreviations: # number of fully-connected (FC), rectifying linear unit (RELU).

	ACC and PPG branches	EDA and TEMP branches
Sampling rate	32 Hz	4 Hz
# Filter per Layer	$4, 8, 16, 32$	$8, 16, 32$
Kernel size	$32, 16, 8, 4$	$4, 4, 2$
Stride	1	1
Padding	'same'	'same'
Max-pool	$4, 4, 2$	$2, 2$
Non-linearity	RELU	RELU
Neurons in FC layers	$64, 32, 16, N_{out}$	

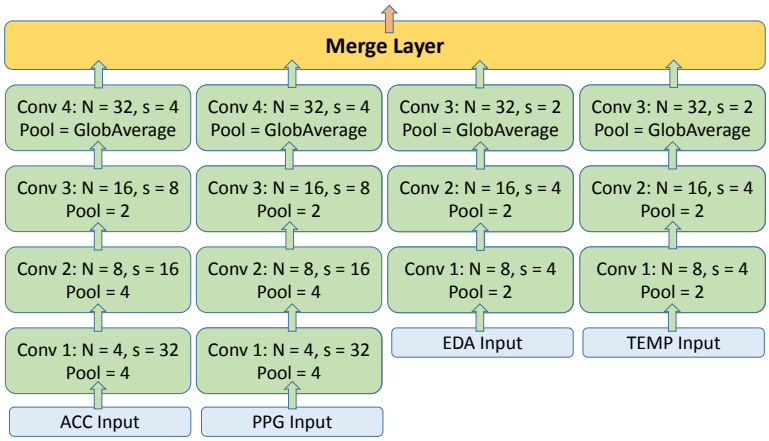


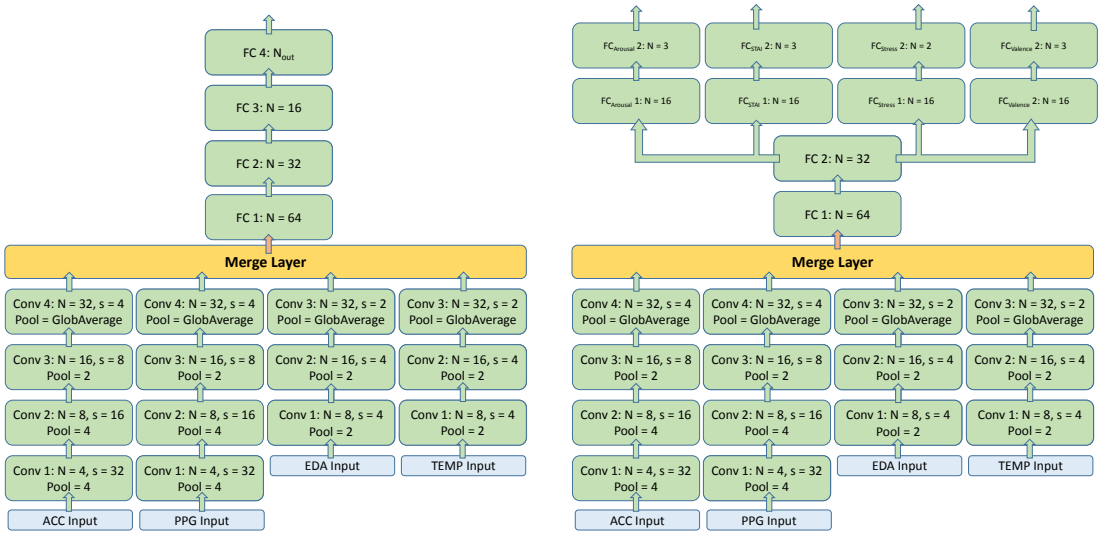
Figure 4.9: Feature extractor applying a four and three layered CNN architecture to the windowed data. Abbreviations: N = number of filters, s = Kernel size.

feature extractor for each classification task and on top four FC layers (depicted in Figure 4.10a), classifying *one specific* type of label each (e.g., valence). In contrast, the MT-CNN share two FC layers, have multiple output branches (see Figure 4.10b), and are trained to classify *all* labels types (arousal, STAI, stress, and valence) simultaneously. Apart from the last FC layer where a softmax is used, the FC layers also use RELUs as non-linearities. In both cases the CNNs were trained using a cross-entropy loss and mini-batches of size 1024 or 64. Following the hyperparameter settings of Hannink et al. [2017], ADAM was used as optimizer.

Following LTQO the ST-CNNs were trained, validated, and tested on stratified splits (80%/10%/10%) of the target questionnaire. In the MT-CNNs setup a stratified split was performed along the arousal labels and the split for the arousal values was then utilized for the other questionnaires, too. During training the F_1 score on the validation set was monitored. For prediction the weights corresponding to the highest score on the validation set were employed.

The number of trainable parameters differs between the ST-CNN and MT-CNN. The binary ST-CNN stress detector has 21946 parameters. Each ST-CNN architecture employed for arousal, STAI, and valence classification contains 21963 trainable parameters. Hence, in the ST-CNN formulation predicting arousal, STAI, stress, and valence would require four different CNNs with a total of 87835 parameters. In contrast, a MT-CNN has a total of 23683 parameters and predicts all targets of interest simultaneously. This is a factor 3.7 less parameters than in the ST-CNN approach.

Unsupervised Pre-training: Similar to Zheng et al. [2016] unsupervised pre-training, using convolutional auto-encoders (convolutional AE), has been investigated. Here convolutional AEs were trained for each sensor modality separately, using ADAM as optimizer and mean squared error loss. All convolutional AE were trained for 40



(a) ST-CNN network architecture.

(b) MT-CNN network architecture.

Figure 4.10: Illustration of the ST-CNN and MT-CNN network architectures. Abbreviations: N = number of filters, s = Kernel size, FC = Fully Connected.

epochs, using all available windowed data (80%/20% train/test split). Apart from the global average pooling operation the modality specific encoder employed the same type of convolution and pooling operations as the feature extractors described above. In the decoder part of the convolutional AE upscaling and convolutional layers (reversing the number of kernels and filter sizes) were applied. Using these convolutional AE weights two different experiments were performed:

1. The encoder weights were set to non-trainable during the fine-tuning (referred to as *frozen* below). Hence, only the final FC classification layers were updated during training.
2. Both the encoder weights and the classification layers were updated during the training. This setup is referred to as *not frozen*.

Here ADAM, with hyperparameter settings as before, has been used. The CNNs were implemented in keras using a tensorflow backend and trained on Nvidia GTX 1080 TI GPUs.

4.4.4 Classification Results

In this section the results obtained using different feature-based classification methods and end-to-end learning are compared.

Table 4.5: Mean F_1 score in [%] using feature-based classifiers and the subject independent validation scheme (LOSO). The results are averaged over the different subjects and five runs per subject. The last column displays the F_1 score averaged over the tasks. Abbreviations: Decision-tree (DT), Random forest (RF), Randomized decision trees (ET), z-normalisation (zN), and Sophisticated Guesser (Base).

	Arousal	STAI	Stress	Valence	Average
DT	30.9 ± 3.8	31.4 ± 3.4	46.7 ± 5.9	33.5 ± 1.7	35.6 ± 3.7
DT+zN	30.8 ± 3.9	31.3 ± 3.3	$\mathbf{46.8 \pm 5.7}$	33.7 ± 1.9	35.7 ± 3.7
ET	31.4 ± 8.1	33.3 ± 9.0	46.5 ± 7.4	41.1 ± 8.2	38.1 ± 8.2
ET+zN	$\mathbf{31.4 \pm 8.1}$	$\mathbf{33.6 \pm 9.3}$	46.5 ± 7.3	$\mathbf{42.2 \pm 9.5}$	$\mathbf{38.4 \pm 8.6}$
RF	30.6 ± 5.1	31.7 ± 7.6	46.2 ± 7.2	39.9 ± 8.3	37.1 ± 7.1
RF+zN	30.6 ± 5.2	32.1 ± 8.5	46.2 ± 7.3	39.9 ± 8.4	37.2 ± 7.3
Base	19.9 ± 9.5	21.3 ± 6.7	39.0 ± 20.3	26.4 ± 7.9	26.6 ± 11.1

Table 4.6: Mean F_1 score in [%] using feature-based classifiers and the subject dependent validation scheme (LTQO). The displayed results are averaged over five runs and the last column displays the F_1 score averaged over the tasks. Abbreviations: Decision-tree (DT), Random forest (RF), Randomized decision trees (ET), z-normalisation (zN), and Sophisticated Guesser (Base).

	Arousal	STAI	Stress	Valence	Average
DT	37.4 ± 0.8	34.9 ± 1.4	53.1 ± 0.7	40.1 ± 1.5	41.3 ± 1.1
DT+zN	38.7 ± 2.2	37.5 ± 1.0	52.3 ± 0.4	39.6 ± 1.4	42.0 ± 1.3
ET	38.4 ± 1.5	36.8 ± 1.7	56.3 ± 1.7	42.8 ± 3.1	43.6 ± 2.0
ET+zN	$\mathbf{38.8 \pm 1.7}$	35.2 ± 1.9	$\mathbf{57.9 \pm 3.2}$	$\mathbf{43.3 \pm 1.2}$	$\mathbf{43.8 \pm 2.0}$
RF	38.2 ± 2.2	$\mathbf{37.4 \pm 2.0}$	54.9 ± 1.6	42.2 ± 0.6	43.2 ± 1.6
RF+zN	38.0 ± 2.5	36.0 ± 2.0	55.8 ± 1.7	42.6 ± 1.6	43.1 ± 1.9
Base	21.7	22.2	34.9	23.6	25.6 ± 5.5

Feature-based Evaluation:
Using the above described features and evaluation schemes (LOSO and LTQO), the performance of different classical decision tree-based classifiers has been investigated. The performance of these classifiers is compared to a sophisticated guesser baseline, predicting only the majority class found in the training set. Table 4.5 displays the F_1 scores generated using LOSO. The results were averaged over all subjects and per subject five runs have been performed. Here, the decision tree-based classifiers are able to outperform the sophisticated guessing baseline by a large margin. The results of both the sophisticated guesser baseline and the decision tree-based classifiers have rather large standard deviations. This is to be attributed to the large inter-subject differences in the label distributions, which pose limitations on successful generalization. Averaging the obtained F_1 scores over the different tasks, the ET using the normalised features (ET+zN) reached the highest combined F_1 score.

In Table 4.6, the performance of the feature-based classifiers using the LTQO evaluation scheme is reported. In this setup the RF and ET reach similar averaged F_1 scores. The reasons for the increased F_1 scores following LTQO are twofold: First, LTQO is subject dependent, which simplifies the problem. Secondly, due to the stratified split the folds have the same label distributions, which decreases the

Table 4.7: Mean F_1 score in [%] using CNNs and LTQO as validations scheme. All results were averaged over three runs and the last column displays the F_1 score, averaged over the different tasks.

		Arousal	STAI	Stress	Valence	Average
$Batch_{size} = 1024$	Training CNNs from scratch					
	ST-CNN	44.3 ± 1.4	39.2 ± 0.5	55.5 ± 3.4	42.8 ± 2.4	45.4 ± 1.9
	MT-CNN	42.8 ± 3.8	37.4 ± 1.1	56.6 ± 1.6	44.0 ± 2.4	45.2 ± 2.2
	Fine-tuning: convolutional AE weights are *frozen*					
	ST-CNN	43.1 ± 4.1	36.9 ± 0.7	58.3 ± 0.8	43.2 ± 2.0	45.4 ± 1.9
	MT-CNN	39.4 ± 1.8	36.3 ± 0.5	56.8 ± 1.5	41.2 ± 0.8	43.4 ± 1.1
	Fine-tuning: convolutional AE weights are *not frozen*					
	ST-CNN	42.5 ± 3.3	38.1 ± 2.2	53.8 ± 6.0	40.4 ± 1.4	43.7 ± 1.9
	MT-CNN	43.9 ± 3.0	41.5 ± 2.0	55.7 ± 2.7	39.0 ± 0.5	45.0 ± 2.0
$Batch_{size} = 64$	Training CNNs from scratch					
	ST-CNN	42.9 ± 3.0	37.0 ± 0.9	56.9 ± 1.4	40.3 ± 0.9	44.3 ± 1.5
	MT-CNN	42.1 ± 1.0	38.3 ± 1.6	57.0 ± 1.1	44.6 ± 3.5	45.5 ± 1.8
	Fine-tuning: convolutional AE weights are *frozen*					
	ST-CNN	40.5 ± 1.6	35.1 ± 0.7	54.2 ± 2.0	43.4 ± 3.6	43.3 ± 2.0
	MT-CNN	42.8 ± 5.9	36.6 ± 1.2	56.6 ± 0.9	42.0 ± 1.6	44.5 ± 2.4
	Fine-tuning: convolutional AE weights are *not frozen*					
	ST-CNN	41.9 ± 3.9	38.8 ± 0.7	55.8 ± 1.2	41.7 ± 1.6	44.5 ± 1.8
	MT-CNN	40.6 ± 2.4	39.2 ± 3.3	57.7 ± 0.9	41.6 ± 0.9	44.8 ± 1.9

standard deviation.
From both Table 4.5 and Table 4.6, two major observations can be made: First, the normalisation has no crucial influence on the averaged F_1 scores. Second, the overall $F1$ scores are not satisfying. However, related work reported an F_1 score of 47%, using similar feature-based methods for a stress detection task, see Gjoreski et al. [2017]. Hence, it can be speculated that the employed classifiers might not be powerful enough to learn these relations. Therefore, the next step is to investigate the performance of CNNs, which established the current state-of-the-art on many human activity recognition tasks, see for instance Hammerla et al. [2016], Hannink et al. [2017], or Münzner et al. [2017].

End-to-End learning:
Table 4.7 displays the results from the CNN experiments using the LTQO validation scheme. For the binary stress classification task the highest F_1 scores were reached. Similar to the classical results presented in Table 4.6, the lowest scores are obtained for the multi-class STAI classification. Using CNNs the arousal and valence classification tasks are solved with a similar performance. However, comparing classical valence and arousal classification (see Table 4.6) to the CNN-based one, especially the arousal task, is solved with a higher F_1 score.

The mean $F1$ score over the different tasks using CNNs is on average 1.8 percent points better than the average $F1$ score of the feature-based classifiers. In general, this is only a minor improvement. The highest mean F_1 scores reached by

the CNNs is around 45.5%. This result is achieved by the ST-CNN trained from scratch with $N_{batch} = 1024$ and the fine-tuned ST-CNN, where the weights of the convolutional auto-encoders (convolutional AE) have been frozen. In addition, the MT-CNN trained from scratch setting $N_{batch} = 64$ reaches an averaged F_1 of 45.5%, too. Hence, these best CNNs outperform the best classical approach (ET, averaged $F_1 = 43.8$) by 1.6-1.7% percent points.
In general, the performance of the ST-CNNs and MT-CNNs are comparable. However, the MT-CNNs predict all labels simultaneously and require only little more parameters than a single ST-CNN.
Judging from Table 4.7 both investigated batch sizes $N_{batch} = [1024, 64]$ led to similar performances. In addition, utilizing the pre-trained convolutional AE weights did not improve the results. These observations hold for both experimental settings (frozen and not frozen weights in the feature extractor).

The F_1 scores presented in Table 4.7 are only a marginal improvement of the scores reached by the classical approaches. However, the utilized CNNs operate directly on the windowed data and, hence, make feature engineering obsolete. In addition, the presented CNN are small (less than 25k parameters). Restricted Boltzmann Machines, requiring up to a factor of 140 more parameters, were successfully deployed on a Snapdragon 400 platform Bhattacharya and Lane [2016]. Therefore, the deployment of our models should be feasible on a similar platform.

4.4.5 Limitations and Further Considerations

Above the performance of feature-based classifies and convolutional neural networks (CNNs), applied to data originating from an affect recognition (AR) field study have been investigated. Using a subject dependent evaluation scheme and state-of-the-art CNNs it was challenging to reach an average F_1 score higher than 45%. From literature, see for instance Gjoreski et al. [2017] or Healey et al. [2010], similar results were reported which underlines that this is a challenging problem.

4.4.5.1 Limitations of the Presented Approach

Assessing the above employed methods critically, the following algorithmic and data-driven limitations can be identified in the presented work:

Algorithmic: The pipeline was tailored to directly classify the affective state based on features, either learned or hand-crafted, from 60 sec data snippets. In this approach the temporal and sequential nature of the data are not captured. This could be improved by adding a refinement step, e.g., voting over multiple adjacent windows. Another approach would be to model the temporal nature of the data explicitly by employing a Hidden Markow Model or recurrent neuronal networks, for instance.

Data and label quality: Data obtained from field studies are intrinsically noisy and the labels are not completely reliable. Data noise ranges from sensor misplacement to movement artefacts. Furthermore, label fuzziness can be attributed to the subjective nature of ecological-momentary-assessments (EMAs). In Figure 4.8 an intrinsic bias towards positive labels is observed, this is reflected by the skewed label distributions. In our opinion the reasons for this skewness are twofold: First, the subjects are less likely to respond to or trigger an EMA while being in a high arousal (e.g., stressed) affective state. Secondly, according to the social desirability bias, see Edwards [1957], subjects are less likely to report on states less socially desired (like being in a bad mood). All in all, the data and label noise certainly has an adverse effect on the results.

Amount of labelled data: Labels gathered via EMAs are discrete and sparse. For the presented analysis we utilized 1083 valid EMAs. Training classifiers on such small amounts of (skewed) data is difficult and combating both over- and underfitting is challenging, even if different types of regularization, e.g., dropout or L2 regularization, are employed.

Considered Subject Cohort: In the presented study, students were targeted. As a result, the mean age of participants is 26 ± 2.5 and most (7 out of 11) were male. In addition, the study had "only" a duration of two weeks. These factors pose limitations on the presented results. Conducting a similar study with a different/more diverse (in an ideal case even representative) cohort of subjects monitored for a longer duration would be one way to mitigate this.

4.4.5.2 Inherent Pitfalls of Affect Recognition in the Wild

Reflecting on the insights gained during the study and the classification procedure, the following inherent pitfalls and lessons learned are formulated:

Curse of normality: Healthy users are unlikely to exhibit strong mood swings across the entire affective spectrum. Assuming a (skewed) Gaussian shape of the label distribution most labels will be reported around a mean value of "things are okay/normal". As a result, extrema in the affective spectrum are broadly underrepresented. In the label data presented, see Figure 4.8, this "normal state" is indicated by low arousal/positive valence and other states are underrepresented. Hence, classifying rare episodes where a user is in an extreme state is very challenging, due to the low number of data points. However, treating extreme cases as outliers and applying methods from outlier detection is a direction worthwhile investigating. This could also be used during data collection to trigger EMAs, once an outlier state is detected.

Awareness of affective states: According to Mattila et al. [2006], between 5 and 18% of the general population has difficulties with identifying and describing their emotions. Hence, label quality could be increased dramatically by providing mindfulness sessions for study participants. In addition, it might be interesting to explore

other labelling techniques than EMAs, where subjects are given more time to reflect about their affective state and then answer a set of questionnaires.

Representation of affect: Dimensional representations (e.g., valence and arousal) of affective states are intuitive. Based on our study it seems, see Figure 4.7, that most subjects do not utilize the entire spectrum. This might be due to personal biases (e.g., personality traits). One way to mitigate this bias could be to normalize the labels of each subject. However, this approach would require the label distribution to be known. This is not the case, especially in a real world application. An alternative approach would be to develop affective scales with a finer granularity tailored to certain personality types. For instance, if someone claims to be a rather positive person it might be beneficial to inquire finer granular levels of positivity (e.g., 'less than normal', 'normal', 'more than normal'), instead of asking about negative valence. Another idea would be to ask the user to compare events (e.g., "Are you currently more/less aroused compared to your last report?"). Both approaches would increase the variance in the label distribution, facilitating the uncovering of hidden correlations.

Human activity recognition vs. affect recognition: Both wearable-based human activity recognition and AR utilize similar inputs to create a user model. In the human activity recognition domain, however, the employed sensors, 3-axes acceleration for instance, offer direct measures of the performed activity (e.g., walking). In contrast, considering AR, the available sensors only offer indirect measures. In our opinion this contributes to the large performance gap between human activity recognition and AR systems.

Modalities: In our evaluation, we aimed at classifying the affective state of a person purely based on physiological data. Although subjects cannot actively influence their physiological responses, there are many confounding variables. Judging from our experiments, the classifiers had difficulties identifying these confounders. Based on literature, see Gjoreski et al. [2017], context data might be able to alleviate this. However, another direction could also be to add more informative biomarkers to the picture, e.g., cortisol levels, blood pressure, or the chemical sweat composition. Furthermore, Sano et al. [2015] showed that sleep quality is a powerful predictor for mood. Hence, this information could also help to improve the classification results. However, the sleep quality information should be acquired in a passive fashion.

Discrete vs continuous: In contrast to biophysical signals like electro-dermal activity which is available continuously, other information like sleep quality or cortisol measures are only available once a day or at discrete time points. Combining both types of information in a single model can be challenging. One approach could be to have different models for different scenarios (e.g., one specific mood classifier for a high and one for a low cortisol level). Alternatively, these discrete values could serve as one feature used by the classifiers.

4.5 Conclusion

At the beginning of this chapter a quick overview of current field studies was provided (see Section 4.1). In order to address the research question **RQ 3** (*What is the performance of machine learning systems that detect multiple affective states in unconstrained environments?*) a field study has been conducted. During this study, physiological, motion, and label data were collected and in Section 4.2 the study protocol was presented.

As pointed out previously, the labelling of affective states in the wild is challenging. In order to address this, research question *RQ 3a* (*What is an appropriate way to label affective states in everyday life reliably?*) has been posed. Related work suggests that smartphone-based ecological-momentary-assessment (EMA) tools can be employed to label affective states in the wild. Although these tools are often employed, little guidelines for the development of such apps were available. This was addressed in Section 4.3 by presenting paradigms and formulating guidelines. Using the EMA data gathered and experience gained during the study, a thorough analysis of the guidelines was performed. Most of the analyses were of a descriptive nature. Considering this, the evaluation is no strict proof of the formulated guidelines. However, the overall plausible results suggest that the guidelines were implemented successfully and can serve as starting points for other researchers

The physiological, motion, and label data gathered during the study were utilized to address research question *RQ 3b* (*What is the performance of classifiers trained on labels generated with an ecological-momentary-assessment tool?*). For this purpose, classical feature-based and deep learning methods were employed (see Section 4.4). The presented approaches target the arousal, State-Trait Anxiety Inventory, and valence scales as a multi-class classification task. The stress classification was pursued in a binary fashion. In addition to the multi-class classification tasks, a multi-target classification task where all labels were predicted simultaneously using convolutional neural networks (CNNs) was formulated. In the subject dependent formulation, the performance of feature-based classifiers and different CNNs was analysed. The CNNs lead, compared to the classical method, to a minor improvement of the average F_1 score (1.8 percent points). These results indicate that despite state-of-the-art methods affect recognition in the wild is still very challenging. The challenges inherent to wearable-based affect recognition arise from different sides:

- The employed physiological and motion data are only indirect measures of affective states. Human physiology is influenced by many factors and the affective state is only one of them. Hence, noise introduced by device misplacement or motion as well as other confounding factors, like humidity, result in strong artefacts, posing limitations on the classification scores.
- The dataset considered in this work contained, after cleaning, 1083 valid self-reports collected from 11 healthy participants. In order to train affect recogni-

tion systems applicable in unconstrained environments much larger and ideally representative datasets are required. However, at the point of writing no such dataset meeting the criteria detailed in Section 1.1 was available.
- Kreibig [2010] pointed out that some affective states exhibit certain physiological patterns. The way individual subjects perceive (e.g., label) and react (in a physiological sense) to an affective stimulus can differ strongly. Hence, personalisation is a direction worth investigating.
- For a healthy subject, swings across the entire affective spectrum are unlikely to occur frequently. Hence, in order to detect the rare occurrences of exceptional cases methods from outlier detection might be applied successfully.

To achieve the goal of ubiquitous wearable-based affect recognition, these open points need to be addressed.

5 Résumé

5.1 Results and Contributions

As detailed in Chapter 1, a holistic user model requires the affective state of a user as an integral part. Wearables offer the ideal platform to monitor human activities and physiological parameters in an unobtrusive and continuous fashion. Hence, in the presented thesis the performance of wearable-based affect recognition (AR) systems relying solely on physiological and motion data have been investigated. For this purpose the presented thesis focused on different aspects, targeting three major research questions:

RQ 1: What is the current state-of-the-art in wearable-based affect recognition?

> This has been addressed in Chapter 2 and related literature has been reviewed carefully identifying the state-of-the-art in wearable-based AR. For this purpose the following aspects were covered: First, psychological and physiological constructs for affects (Section 2.1 and Section 2.2) were presented. Second, lab and field study protocols were detailed (Section 2.3). Third, the common classification pipeline has been outlined (Section 2.4). Chapter 2 is written in a tutorial style fashion, providing a newcomer to the field a deep and broad overview of the constructs and methods employed in wearable-based AR. Judging from this survey two major shortcomings were identified: First, the wearable-based AR community lacks a commonly used benchmarking dataset. Second, new machine learning approaches, such as deep learning methods, have found little application in wearable-based AR, yet.

RQ 2: How is benchmarking and direct comparison of different algorithmic approaches for wearable-based affect recognition feasible?

This has been addressed by recording a lab study dataset and making it publicly available [1]. During the lab study, presented in Chapter 3, physiological and motion data as well as self-reports of 15 healthy participants were collected. For the data collection, the *Empatica E4* (smartwatch) and the *RespiBan* (chest belt with additional sensor modalities) were employed. Following the classical (feature-based) pipeline of Bulling et al. [2014] the data was segmented, features were computed, and a classification has been performed. The classification targeted three different affective states, namely: Neutral, stress, and amusement. It was performed in a subject independent way using the leave-one-subject-out validation scheme. The three class classification task (stress vs. neutral vs. amusement) was solved with accuracies up to 80% (see Table 3.3). In the binary classification task (stress vs. non-stressed) classification accuracies up to 93% (see Table 3.4) were reached. These results are very promising, indicating that the distinction between different affective states based purely on physiological and motion data is feasible in a lab setting.

RQ 3: What is the performance of machine learning systems that detect multiple affective states in unconstrained environments?

In order to target this research question, a field study has been conducted, see Chapter 4. During the field study physiological and motion data were recorded using the *Empatica E4* smartwatch. The analyses presented in Chapter 4 are based on the data of 11 healthy subjects participating for at least 14 days each. Appropriate labelling of affective states in the wild is commonly done via ecological-momentary-assessments (EMAs). However, no guidelines and best practices were available for the development of smartphone-based EMA labelling tools. Hence, the following research question has been posed:

RQ 3a: *What is an appropriate way to label affective states in everyday life reliably?*

Overall the use of smartphone-based EMA tools seems to be appropriate to label affective states in the wild. Section 4.3 details and evaluates the guidelines which influenced the development of the smartphone app used as EMA labelling tool in the presented study. The formulated guidelines target the following points:

[1] Download Link: https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/

1	Sampling Rate and Scheduling	5	Context Information
2	Filing Time and Number of Items	6	Daily Data-Driven Screening
3	Manual Trigger of EMAs	7	Subjects' Commitment
4	Validity and Redundancy of EMAs		

i	Allow Hindsight Labelling	ii	Incorporate Reviewing Possibility

Guidelines 1 to 7 were evaluated using the EMA data gathered during the field study. Based on this analysis lessons learned were formulated. The overall plausible findings, see Section 4.3, suggest that these guidelines were successfully implemented into the presented field study. Guideline i and ii were formulated based on subjects' feedback after the completion of the study. We encourage the wearable-based AR community to use the guidelines listed above as a starting point for the development of EMA-based labelling tools and update them with their own findings.

In order to facilitate AR in the wild, machine learning classifiers need to be trained to perform a mapping between observables (e.g., physiological indicators) and affective labels (e.g., valence value). As this type of observables and labels were acquired during the field study the following research question has been targeted:

RQ 3b: *What is the performance of classifiers trained on labels generated with an ecological-momentary-assessment tool?*

To address this research question, the physiological and motion data gathered using the *Empatica E4* and labels collected via the EMA tool were used (see Section 4.4). Different binary and multi-class classification tasks were formulated, using both feature-based as well as deep learning methods. Using state-of-the-art methods and a subject-dependent leave-target-questionnaires-out validation scheme, the F_1 score averaged across the different classification tasks barely exceeded 45%. Hence, the poor performance of the presented approaches was discussed and structural issues for wearable, especially physiology- and motion-based, AR systems were identified. One of the key findings is, that extreme states in the affective spectrum of the participants occurred only rarely and that most affective states are somewhere around the label "things are okay/normal". Hence, detecting these outlier cases is difficult, as they are underrepresented in the training data. As a result, reformulating the classification task to an outlier detection task might be an interesting approach.

All in all this thesis made the following contributions:

1. An extensive analysis of the current state-of-the-art in wearable-based AR, focusing on stress and emotion, has been presented.
2. A lab study dataset (N=15) for WEarable Stress and Affect Detection WESAD has been recorded and made publicly available. Although published recently (October 2018), the dataset already exhibited a certain impact on the research community.
3. WESAD has been benchmarked using various classification algorithms and posing different AR problems.
4. Paradigms and guidelines for EMA-based field studies were formulated. These guidelines were evaluated using the EMA data collected during the AR field study.
5. A field study (N=11) has been conducted recording a large and realistic dataset. Using different classical and end-to-end trainable machine learning methods the dataset was analysed.
6. Based on these results pitfalls for wearable-based AR, relying solely on physiology and motion data, gathered in the field, were deduced.

The listed contributions are mainly of a practical nature. The extensive analysis of the state-of-the-art and the formulated guidelines, for instance, are beneficial for other researchers when planning an AR field or lab study. In addition, the published and benchmarked WESAD lab study dataset is of direct value for the entire research community. This is due to the fact that it incorporates data recorded from a large number of (redundant) modalities while the participants were subject to different affective stimuli. As a result, WESAD facilitates a direct comparison of different sensor locations, features, and algorithmic approaches to wearable-based AR. Moreover, the limitations and pitfalls identified in Chapter 4, are very relevant for future work and need to be overcome in order to realize real life wearable-based AR systems.

5.2 Future Work

The contributions made in this thesis can be viewed as a step towards unobtrusive and continuous wearable-based affect recognition. Some challenges, however, still need to be addressed to facilitate the goal of high precision affect recognition systems applicable in everyday life. Below possible next steps are outlined:

Algorithmic challenges: Humans perceive, react to, and evaluate affective stimuli in different ways. This calls for personalisation. However, up-to-date little use of personalisation methods is made in wearable-based affect recognition. One direction for personalisation could be the use of online learning where a population-based model is "fine-tuned" and gradually improved with the data of the current user. Following such an approach, the user could provide or correct labels, whenever he or she does not agree with the predictions of the model.

As detailed previously healthy subjects are unlikely to exhibit strong mood swings. Hence, enforcing some continuity constraints on the predictions could be beneficial. In addition considering the rare cases of extreme affective states, methods from anomaly detection, like proposed by Popoola et al. [2018], could also find application in wearable-based affect recognition.

Datasets: As of now there are only a few publicly available datasets (see Section 2.3.3). However, most of these datasets were recorded in lab settings. In addition, the data originates mainly from healthy (graduate) students or researchers. These two groups represent a rather specific cohort of subjects. This certainly introduces biases into the available data and labels. Overcoming this by recording large scale (representative) datasets and making them available to the community would be very beneficial. Furthermore, as wearable-based affect recognition could find application in clinical settings, see for instance Rubin et al. [2016] or Grünerbl et al. [2015], datasets recorded from persons with specific health conditions could be very valuable.

Labelling: In order to increase the amount of "interesting" labels, active labelling approaches could find application in wearable-based affect recognition field studies. For this purpose, ecological-momentary-assessment could be scheduled in an event-based fashion, e.g., when a dramatic increase of the heart rate occurs without a significant change in the motion data. This type of event-based labelling system could also serve as a precursor for online learning and personalisation methods.

Hardware: Recent progress in flexible electronics enabled the development of sensor patches (e.g., Vivalnk) and epidermal electronics. The potential of epidermal electronics measuring different electrophysiological signals, like electrocardiogram, electromyogram, and even electroencephalogram was demonstrated by Ameri et al. [2016] or Sadri et al. [2018]. At the point of writing, these devices found little application in wearable-based affect recognition. However, they have a great potential and their performance should be investigated in future affect recognition (field) studies.

Deployment on Embedded Devices: In this thesis wearable-based affect recognition systems were investigated. As a result, the sensor modalities and the algorithms used for classification were selected with (the computational capacity of) wearables in mind. The deployment of such models on actual hardware, was not scope of this work but is essential for a real-life applications.

Considering the large interest from academic, industrial, and consumer side in wearable-based affect recognition systems we are confident that these open research directions will be addressed soon.

Appendices

A Lists

A.1 List of Figures

A.2 List of Tables

B

Acknowledgements

Zum Schluss möchte ich mich bei allen bedanken, die zum Gelingen dieser Arbeit beigetragen haben:

- Zuerst möchte ich mich ganz herzlich bei Prof. Dr. Kristof Van Laerhoven für das Ermöglichen dieser Arbeit und die gute Betreuung bedanken.
- Ein herzlicher Dank geht an Herrn Prof. Dr.-Ing. Klaus David für die Zweitbetreuung.
- Ein ganz besonderer Dank gilt Attila Reiss, der mir mit Rat und Tat bei sämtlichen Fragen zur Seite stand und nie müde wurde mit mir über Probleme zu diskutieren.
- Ein ganz besonderer Dank gilt auch Robert Dürichen, der diese Arbeit mit vielen interessanten Ideen bereichert hat und maßgeblich an dem Zustandekommen meines Auslandaufenthaltes beteiligt war.
- Besonders möchte ich mich bei Thomas Plötz für die Ermöglichung des Aufenthaltes an der Georgia Tech und den sehr wertvollen Input bei der Feldstudienauswertung bedanken.
- To all the people I met in Atlanta - Thanks to y'all!
- Danke an Mirko Ruhs für das Finalisieren der EMA App und ein DICKES DANKE an die gesamte AEU (1 und 2) - schee wars mit euch ☺
- Ein herzliches Dankeschön geht an alle Studienteilnehmer - ihr wart toll!
- Danke auch an Jonas, David und Fabi für die Sportsessions und den Diss FAQ.
- Ein ganz herzlicher Dank geht an meine Familie, die mich auf meinem Weg immer begleitet.
- Zu guter Letzt geht ein ganz besonderer Dank an meine Frau Rahel, die mich bei meinem Promotionsvorhaben immer voll unterstützt hat.

This document was typeset using LaTeX.

C

Publication List

During the presented thesis the following first author contributions have been made:

- **Schmidt, P.**, Duerichen, R., Reiss, A., Van Laerhoven, K., Plötz, T., Multi-target Affect Detection in the Wild: An Exploratory Study, In *Proceedings of the 23rd International Symposium on Wearable Computers*, ISWC '19, http://doi.acm.org/10.1145/3341163.3347741.
- **Schmidt, P.**, Reiss, A., Duerichen, R., Van Laerhoven, K., Wearable-Based Affect Recognition - A Review, In Sensors 2019, 19(19), 4079, https://doi.org/10.3390/s19194079
- **Schmidt, P.**, Reiss, A., Duerichen, R., Van Laerhoven, K., Labelling Affective States "in the Wild": Practical Guidelines and Lessons Learned. In Adjunct *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, http://doi.acm.org/10.1145/3267305.3267551.
- **Schmidt, P.**, Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K., Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, http://doi.acm.org/10.1145/3242969.3242985.

In addition, the following co-author contributions have been made:

- Reiss, A., Indlekofer, I., **Schmidt, P.**, and Van Laerhoven, K., Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks, *Sensors*, https://www.mdpi.com/1424-8220/19/14/3079,
- Reiss, A., **Schmidt, P.**, Indlekofer, I., and Van Laerhoven, K. PPG-based Heart Rate Estimation with Time-Frequency Spectra: A Deep Learning Approach. In Adjunct *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18. http://doi.acm.org/10.1145/3267305.3274176,

- Dürichen, R., Verma, K., Yee, S., Rocznik, T., **Schmidt, P**, Bödecker, J. and Peters, C. Prediction of Electrocardiography Features Points Using Seismocardiography Data: A Machine Learning Approach. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ISWC'18, http://doi.acm.org/10.1145/3267242.3267283.
- Münzner, S., **Schmidt, P.**, Reiss, A., Hanselmann, M., Stiefelhagen, R. and Dürichen, R CNN-based Sensor Fusion Techniques for Multimodal Human Activity Recognition In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ISWC'17, http://doi.acm.org/10.1145/3123021.3123046.,

D

Glossary

AB	AdaBoost
AC	Affective computing
ACC	3-axes acceleration
ANOVA	Analysis of Variance
ANS	Autonomic nervous system
AR	Affect recognition
ASCERTAIN	Multimodal databASe for impliCit pERsonaliTy and Affect recognitIoN using commercial physiological sensors
BFI	Big Five Inventory
BN	Bayesian network
BP	Blood pressure
CNN	Convolutional neural network
convolutional AE	Convolutional auto-encoders
CV	Cross-validation
DBN	Deep belief network
DEAP	Database for Emotion Analysis using Physiological signals
DECAF	DECoding user physiological responses to AFfective multimedia content
DT	Decision-tree
ECG	Electrocardiogram
EDA	Electrodermal activity
EEG	Electroencephalogram
EMA	Ecological-momentary-assessment

EMG	Electromyogram
EOG	Electrooculography
ET	Randomized decision trees
FC	Fully-connected
FT	Function tree
GB	Gradient boosting
GMM	Gaussian Mixture Model
HAHV	High arousal/high valence
HALV	High arousal/low valence
HAR	Human activity recognition
HMI	Human machine interaction
HMM	Hidden Markow model
HR	Heart rate
HRV	Heart rate variability
IAPS	International Affective Picture System
IBI	Inter beat interval
kNN	K-nearest neighbour
LAHV	Low arousal/high valence
LALV	Low arousal/low valence
LDA	Linear discriminant analysis
LDF	Linear discriminant function
LOO	Leave-One-Out
LOSO	Leave-one-subject-out
LOTO	Leave-One-Trial-Out
LR	Logistic regression
LSTM	Long short-term memory
LTQO	Leave-target-questionnaires-out
MEG	Magnetoencephalogram
ML	Machine learning
MLP	Multilayer perceptron

MT-CNN	Multi-task CNN
NB	Naive Bayes
NN	Neural networks
PA	Passive aggressive classifier
PAM	Photo Affect Meter
PANAS	Positive and Negative Affect Schedule
PCA	Principal component analysis
PD	Pupil diameter
PHQ-9	Patient Health Questionnaire
PNS	Parasympathetic nervous system
PP	Percent points
PPG	Photoplethysmogram
PSD	Power spectral density
PSQI	Pittsburgh Sleep Quality Index
PSS	Perceived Stress Scale
QDA	Quadratic discriminant analysis
RELU	Rectifying linear unit
RESP	Respiration
RF	Random forest
RIP	Respiratory inductive plethysmograph
RR	Ridge regression
RSA	Respiratory sinus arrhythmia
RT	Regression tree
SAM	Self-Assessment Manikins
SCL	Skin conductance level
SCR	Skin conductance response
SNS	Sympathetic nervous system
SpO2	Arterial oxygen level
SRI	Stress Response Inventory
SSSQ	Short Stress State Questionnaire
ST-CNN	Single-task CNN
STAI	State-Trait Anxiety Inventory

SVM	Support vector machine
TEMP	Skin-temperature
TINN	Triangular interpolation index
TSST	Trier Social Stress Test
WESAD	Dataset for WEarable Stress and Affect Detection

E

Bibliography

M. Abadi, R. Subramanian, S. Kia, P. Avesani, I. Patras, and N. Sebe. DECAF: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3), 2015.

P. Adams, M. Rabbi, T. Rahman, M. Matthews, A. Voida, G. Gay, T. Choudhury, and S. Voida. Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, 2014.

F. Agrafioti, D. Hatzinakos, and A. K. Anderson. ECG pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, 3(1), 2012.

S. K. Ameri, R. Ho, H. Jang, Y. Wang, and N. Lu. Thinnest transparent epidermal sensor system based on graphene. In *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016.

Apple. www.apple.com/apple-watch-series-4/health/, 2019. Accessed: 2019-07-30.

B. Barker, H. Barker, and A. Wadsworth. Factor analysis of the items of the state-trait anxiety inventory. *Journal of Clinical Psychology*, 33(2), 1977.

C. Becker-Asano. *WASABI: Affect simulation for agents with believable interactivity.* 2008.

J. Behar, J. Oster, Q. Li, and G. D. Clifford. Ecg signal quality during arrhythmia and its application to false alarm reduction. *IEEE Transactions on Biomedical Engineering*, 60(6), 2013.

M. Benedek and C. Kaernbach. Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, 47(4), 2010.

N. Van Berkel, J. Goncalves, and V. Kostakos. Gamification of mobile experience sampling improves data quality and quantity. *IMWUT*, 2017.

Beyondverbal. http://beyondverbal.com/, 2019. Accessed: 2019-07-24.

S. Bhattacharya and N. D. Lane. From smart to deep: Robust activity recognition on smartwatches using deep learning. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 2016.

Biosppy. http://biosppy.readthedocs.io/en/stable/, 2017. Biosignal procssing in Python.

J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani. A non-EEG biosignals dataset for assessment and visualization of neurological status. In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2016.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 2006.

D. Biswas, N. Simões-Capela, C. Van Hoof, and N. Van Helleputte. Heart rate estimation from wrist-worn photoplethysmography: A review. *IEEE Sensors Journal*, 19(16), 2019.

Bosch, 2019. URL https://www.bosch-mobility-solutions.com/en/products-and-services/passenger-cars-and-light-commercial-vehicles/driver-assistance-systems/driver-drowsiness-detection/. Accessed: 2019-07-30.

G. Bower. Mood and memory. *Am Psychol*, 1981.

L. Breiman. Random forests. *Machine Learning*, 45(1), 2001.

E. Broek, V. Lisy, J. Janssen, J. Westerink, M. Schut, K. Tuinenbreijer, A. Fred, J. Filipe, and H. Gamboa. Affective man-machine interface: Unveiling human emotions through biosignals. In *Biomedical Engineering Systems and Technologies*, 2009.

K Budidha and P A Kyriacou. The human ear canal: investigation of its suitability for monitoring photoplethysmographs and arterial oxygen saturation. *Physiological Measurement*, 35(2), 2014.

A. Bulling, U. Blanke, and B. Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput Surv.*, 2014.

D. Buysse, C. Reynolds, T. Monk, S. Berman, and D. Kupfer. The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Research*, 28(2), 1989.

R. Calvo, I. Brown, and S. Scheding. *Effect of Experimental Factors on the Recognition of Affective Mental States through Physiological Measures*. 2009.

W. Cannon. Bodily changes in pain, hunger, fear and rage. 1929.

Ginevra Castellano, Loic Kessous, and George Caridakis. *Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech*. 2008.

G. Chanel, J. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8), 2009.

J. Choi, B. Ahmed, and R. Gutierrez-Osuna. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 16(2), 2012.

T. Christy, L. Kuncheva, and K. Williams. Selection of physiological input modalities for emotion recognition. Technical report, 2012.

G. Chrousos and P. Gold. The concepts of stress and stress system disorders: overview of physical and behavioral homeostasis. *Jama*, 267(9), 1992.

M. Cicero. *Cicero on the Emotions: Tusculan Disputations 3 and 4*. 2002.

S. Cohen, T. Kamarck, and R. Mermelstein. A global measure of perceived stress. *Journal of Health and Social Behavior*, 1983.

C. Darwin. *The Expression of the Emotions in Man and Animals 3rd edn. (Introduction, afterword and commentaries by Paul Ekman. Essay on the history of the illustrations by Phillip Prodger.)*. 1999. First published in 1872.

M. Dawson, A. Schell, and D. Filion. The electrodermal system. In *Handbook of Psychophysiology, Second Edition*. 2000.

Daylio. https://daylio.webflow.io/, 2019. Accessed: 2019-07-30.

E. Di Lascio, S. Gashi, and S. Santini. Laughter recognition using non-invasive wearable devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth'19, 2019.

E. Diener, D. Wirtz, W. Tov, C. Kim-Prieto, D. Choi, S. Oishi, and R. Biswas-Diener. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, 97(2), 2010.

S. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, 47(3), 2015.

S. Dobriek, R. Gajsek, F. Mihelic, N. Pavesic, and V. Struc. Towards efficient multimodal emotion recognition. *International Journal of Advanced Robotic Systems*, 10(1), 2013.

A Edwards. The social desirability variable in personality assessment and research. 1957.

T. Eerola and J. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1), 2011.

P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 1992.

P. Ekman and W. Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1), 1976.

P. Ekman and W. Friesen. Facial action coding system: A technique for measurement of facial movement. *Consulting Psychologists Press*, 1978.

M. Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current Cardiology Reviews*, 8(1), 2012.

Eyeris. https://www.eyeris.ai/, 2019. Accessed: 2019-07-24.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15(1), 2014.

D. Figo, P. Diniz, D. Ferreira, and J. Cardoso. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, 14(7), 2010.

Y. Freund and R. Schapire. A short introduction to boosting. volume 14, 1999.

B. Friedman. Feelings and the body: The jamesian perspective on autonomic specificity of emotion. *Biological Psychology*, 84(3), 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.*, 28(2), 2000.

P. García-Laencina, J. Sancho-Gómez, and A. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2), 2010.

A. Garland. Ex machina, 2015.

Garmin. https://buy.garmin.com/en-US/US/p/567813, 2019. Accessed: 2019-07-30.

S. Gashi, E. Di Lascio, and S. Santini. Using students' physiological synchrony to quantify the classroom emotional climate. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, 2018.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 2006.

D. Girardi, F. Lanubile, and N. Novielli. Emotion detection using noninvasive low cost sensors. In *Seventh International Conference on Affective Computing and Intelligent Interaction*, ACII '17, 2017.

M. Gjoreski, H. Gjoreski, and M. Gams. Continuous stress detection using a wrist device: In laboratory and real life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, 2016.

M. Gjoreski, M. Luătrek, M. Gams, and H. Gjoreski. Monitoring stress with a wrist device using context. *J. Biomed. Inform.*, 2017.

D. Goldstein and I. Kopin. Evolution of concepts of stress. *Stress*, 10(2), 2007.

A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 2016.

J. Gross and R. Levenson. *Emotion elicitation using films*. L. Erlbaum Associates, 1995.

A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, and P. Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J BIOMED HEALTH*, 2015.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3, 2003.

A. Haag, S. Goronzy, P. Schaich, and J. Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and Research Workshop on Affective Dialogue Systems*, 2004.

H. Hamdi, P. Richard, and P. Allain. Emotion assessment for affective computing based on physiological responses. In *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, 2012.

N. Hammerla, S. Halloran, and T. Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

T. Hanai and M. Ghassemi. Predicting latent narrative mood using audio and physiologic data. In *AAAI*, 2017.

J. Hannink, T. Kautz, C. F. Pasluosta, K. Gaßmann, and B. M. Eskofier. Sensor-based gait parameter extraction with deep convolutional neural networks. *IEEE J. Biomed. Health Inform.*, 2017.

M. Hassan, G. Alam, Z. Uddin, S. Huda, A. Almogren, and G. Fortino. Human emotion recognition using deep belief network architecture. *Information Fusion*, 51, 2019.

J. Healey and R. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2), 2005.

J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris. Out of the lab and into the fray: Towards modeling emotion in everyday life. In *Proceedings of the 8th International Conference on Pervasive Computing*, Pervasive'10, 2010.

J. S. Heinisch, C. Anderson, and K. David. Angry or climbing stairs? towards physiological emotion recognition in the wild. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2019.

W. Helton and K. Näswall. Short stress state questionnaire. *European Journal of Psychological Assessment*, 2015.

J. Hernandez, R. Morris, and R. W. Picard. Call center stress recognition with person-specific models. In *Affective Computing and Intelligent Interaction*, volume 6974, 2011.

J. Horne and O. Ostberg. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology*, 4(2), 1976.

K. Hovsepian, M. al'Absi, and S. Kumar. cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, 2015.

T. Huynh and B. Schiele. Analyzing features for activity recognition. In *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*, sOc-EUSAI '05, 2005.

W. James. What is an emotion? *Mind*, os-IX(34), 1884.

N. Jaques, S. Taylor, E. Nosakhare, A. Sano, and R. Picard. Multi-task learning for predicting health, stress, and happiness. In *NIPS Workshop on Machine Learning for Healthcare*, 2016.

S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014.

O. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999), 1999.

E. Kanjo, E. Younis, and C. Ang. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, 49, 2019.

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

Z. Khalili and M. Moradi. Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of eeg. In *2009 International Joint Conference on Neural Networks*, 2009.

D. Kim, Y. Seo, J. Cho, and C. Cho. Detection of subjects with higher self-reporting stress scores using heart rate variability patterns during the day. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008.

J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 (12), 2008.

K. Kim, S. Bang, and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3), 2004.

M. Kim, M. Kim, E. Oh, and S. Kim. A review on the computational methods for emotional state estimation from the human eeg. *Computational and Mathematical Methods in Medicine*, 2013.

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

S. Koelstra, C. Muhl, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 2012.

K. Koh, J. Park, C. Kim, and S. Cho. Development of the stress response inventory and its application in clinical practice. *Psychosomatic Medicine*, 63, 2001.

V. Kollia. Personalization Effect on Emotion Recognition from Physiological Data: An Investigation of Performance on Different Setups and Classifiers. *ArXiv e-prints*, 2016.

S. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3), 2010.

K. Kroenke, R. Spitzer, and J. Williams. The phq-9. *Journal of General Internal Medicine*, 16(9), 2001.

D. Kukolja, S. Popovic, M. Horvat, B. Kovac, and K. Cosic. Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. *International Journal of Human-Computer Studies*, 72, 2014.

P. Lang, M. Bradley, and B. Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*, 2, 1999.

B. Lee, J. Han, H. Baek, J. Shin, K. Park, and W. Yi. Improved elimination of motion artifacts from a photoplethysmographic signal using a kalman smoother with simultaneous accelerometry. *Physiological measurement*, 31(12), 2010.

E. Leon, G. Clarke, V. Callaghan, and F. Sepulveda. A user-independent real-time emotion recognition system for software agents in domestic environments. *Engineering Applications of Artificial Intelligence*, 20(3), 2007.

R. Levenson, P. Ekman, and W. Friesen. Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, 27(4), 1990.

Q Li and G D Clifford. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiological Measurement*, 33(9), 2012.

Chong L. Lim, Chris Rennie, Robert J. Barry, Homayoun Bahramali, Ilario Lazzaro, Barry Manor, and Evian Gordon. Decomposing skin conductance into tonic and phasic components. *International Journal of Psychophysiology*, 25(2), 1997.

M. Lin, Q. Chen, and S. Yan. Network in network. 2013.

W. Lin, D. Wu, C. Li, H. Zhang, and Y. Zhang. *Comparison of Heart Rate Variability from PPG with That from ECG*. Springer International Publishing, 2014.

C. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Appl. Signal Process.*, 2004.

B. Liu. Many facets of sentiment analysis. In *A Practical Guide to Sentiment Analysis*. 2017.

C. Liu, P. Rani, and N. Sarkar. An empirical study of machine learning techniques for affect recognition in human-robot interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005.

H. Lu, D. Frauendorfer, and T. Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp'12, 2012.

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

D. Lykken and P. Venables. Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology*, 8(5), 1971.

S. Mahdiani, V. Jeyhani, M. Peltokangas, and A. Vehkaoja. Is 50 hz high enough ecg sampling frequency for accurate hrv analysis? In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

M. Malik. Task force of the european society of cardiology and the north american society of pacing and electrophysiology. heart rate variability. standards of measurement, physiological interpretation, and clinical use. *Eur Heart J.*, 17, 1996.

H. Martinez, Y. Bengio, and G. Yannakakis. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2), 2013.

J. Mason. A review of psychoendocrine research on the sympathetic-adrenal medullary system. *Psychosomatic Medicine*, 30(5), 1968.

A. Mattila, J. Salminen, T. Nummi, and M. Joukamaa. Age is strongly associated with alexithymia in the general population. *J Psychosom Res*, 2006.

L. McCorry. Physiology of the autonomic nervous system. *American Journal of Pharmaceutical Education*, 71(4), 2007.

B. McEwen and E. Stellar. Stress and the individual: mechanisms leading to disease. *Archives of Internal Medicine*, 153(18), 1993.

A. Mehrotra, F. Tsapeli, R. Hendley, and M. Musolesi. Mytraces: Investigating correlation and causation between users' emotional states and mobile phone interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3), 2017.

J. Mikels, B. Fredrickson, G. Larkin, C. Lindberg, S. Maglio, and P. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4), 2005.

S. Mirsamadi, E. Barsoum, and C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Jon D Morris. Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *J Advert Res*, 1995.

M. Morris and F. Guilak. Mobile heart health: Project highlight. *IEEE Pervasive Computing*, 8(2), 2009.

O. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. Ferrandez. Stress detection using wearable physiological and sociometric sensors. *International Journal of Neural Systems*, 27(02), 2017.

A. Muaremi, B. Arnrich, and G. Tröster. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience*, 3(2), 2013.

S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen. Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ISWC '17, 2017.

F. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016.

J. Pan and W. J. Tompkins. A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3), 1985.

J. Parkka, M. Ermes, K. Antila, M. van Gils, A. Manttari, and H. Nieminen. Estimating intensity of physical activity: A comparison of wearable accelerometer and gyro sensors and 3 sensor locations. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, E. Duchesnay, et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12, 2011.

R. Picard. Affective computing. Technical report, 1995.

R. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 2001.

K. Plarre, A. Raij, S. Hossain, M. Ali, A. Scott, et al. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *10th International Conference on Information Processing in Sensor Networks (IPSN)*, IPSN 2011, 2011.

R. Plutchik. *Emotion: A psychoevolutionary synthesis.* 1980.

J. Pollak, P. Adams, and G. Gay. Pam: A photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, 2011.

G. A. Popoola, C. A. Graves, and P. Ford-Booker. Using unsupervised anomaly detection to analyze physiological signals for emotion recognition. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2018.

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

P. Rainville, A. Bechara, N. Naqvi, and A. Damasio. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, 61(1), 2006.

M. R. Ram, K. V. Madhav, E. H. Krishna, N. R. Komalla, and K. A. Reddy. A novel approach for motion artifact reduction in ppg signals based on as-lms adaptive filter. *IEEE Transactions on Instrumentation and Measurement*, 61(5), 2012.

J. Ramos, J. Hong, and A. Dey. Stress recognition: A step outside the lab. In *Proceedings of the International Conference on Physiological Computing Systems*, 2014.

P. Rathod, K. George, and N. Shinde. Bio-signal based emotion detection device. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2016.

A. Reiss and D. Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109, June 2012.

A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14), 2019.

Y. Rogers and P. Marshall. Research in the wild. *Synthesis Lectures on Human-Centered Informatics*, 10(3), 2017.

R. Rosmond and P. Björntorp. Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity. *Metabolism*, 47(10), 1998.

J. Rubin, H. Eldardiry, R. Abreu, S. Ahern, H. Du, A. Pattekar, and D. Bobrow. Towards a mobile and wearable system for predicting panic attacks. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, 2015.

J. Rubin, R. Abreu, and S. Ahern. Time, frequency & complexity analysis for recognizing panic states from physiologic time-series. *PervasiveHealth*, 2016.

D. Russell. Ucla loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment*, 66(1), 1996.

J. Russell. *Affective space is bipolar*. American Psychological Association, 1979.

J. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 2003.

B. Sadri, D. Goswami, M. Sala de Medeiros, A. Pal, and R. Martinez. Wearable and implantable epidermal paper-based electronics. *ACS applied materials & interfaces*, 10(37), 2018.

S. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. Chon. A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor. *Sensors*, 16(1), 2016.

A. Samson, S. Kreibig, and J. Gross. Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and Emotion*, 30 (5), 2016.

P. Sanches, K. Höök, E. Vaara, C. Weymann, M. Bylund, P. Ferreira, N. Peira, and M. Sjölinder. Mind the body!: Designing a mobile stress management application encouraging personal reflection. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, 2010.

A. Sano and R. Picard. Stress recognition using wearable sensors and mobile phones. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013.

A. Sano, A. Yu, A. McHill, A. Phillips, and R. Picard. Prediction of happy-sad mood from daily behaviors and previous sleep history. In *Conf Proc IEEE Eng Med Biol Soc*, 2015.

L. Santamaria- Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar. Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos). *IEEE Access*, 7, 2019.

H. Sarker, M. Tyburski, M. Rahman, K. Hovsepian, and S. Kumar. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 2016.

U. Schimmack and R. Reisenzein. Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, 2(4), 2002.

P. Schmidt, A. Reiss, R. Dürichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, 2018a. doi: 10.1145/3242969.3242985. URL http://doi.acm.org/10.1145/3242969.3242985.

P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven. Labelling affective states "in the wild": Practical guidelines and lessons learned. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, 2018b. doi: 10.1145/3267305.3267551. URL http://doi.acm.org/10.1145/3267305.3267551.

P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz. Multi-target affect detection in the wild: An exploratory study. In *Proceedings of the 23rd International Symposium on Wearable Computers*, ISWC '19, 2019a. doi: 10.1145/3341163.3347741. URL http://doi.acm.org/10.1145/3341163.3347741.

P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven. Wearable-based Affect Recognition - a Review. *Sensors*, 19(19), 2019b. doi: 10.3390/s19194079. URL https://www.mdpi.com/1424-8220/19/19/4079.

H. Selye. Stress without distress. In *Psychopathology of Human Adaptation*. 1974.

C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2), 2010.

K. Shear, T. Brown, D. Barlow, R. Money, D. Sholomskas, S. Woods, J. Gorman, and L. Papp. Multicenter collaborative panic disorder severity scale. *American Journal of Psychiatry*, 154(11), 1997.

H. Sietz. I, robot, 2004.

SoftBankRobotics, 2017. URL https://www.generationrobots.com/pepper/Pepper%20Datasheet%201.8a%2020170116%20EMEA.pdf. Accessed: 2019-07-24.

M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1), 2012a.

M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2), 2012b.

S. Spielberg. A.i., 2001.

C. Spielberger, R. Gorsuch, and R. Lushene. Manual for the state-trait anxiety inventory. 1983.

R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 1935.

R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, PP(99), 2017.

F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss. Activity-aware mental stress detection using physiological sensors. In *Mobile Computing, Applications, and Services*, 2012.

S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis. A multimodal dataset for various forms of distracted driving. *Scientific data*, 4, 2017.

T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida. Wearable photoplethysmographic sensors - past and present. *Electronics*, 3(2), 2014.

B. Taylor, A. Dey, D. Siewiorek, and A. Smailagic. Using physiological sensors to detect levels of user frustration induced by system delays. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, 2015a.

N. A. Taylor and C. A. Machado-Moreira. Regional variations in transepidermal water loss, eccrine sweat gland density, sweat secretion rates and electrolyte composition in resting and exercising humans. *Extreme Physiology & Medicine*, 2(4), 2013.

S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard. Automatic identification of artifacts in electrodermal activity data. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 2015b.

S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard. Personalized multitask learning for predicting tomorrows mood, stress, and health. *IEEE Trans. Affect. Comput.*, 2017.

R. Thayer. *The biopsychology of mood and arousal.* 1990.

P. Tzirakis, G. Trigeorgis, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *CoRR*, 2017.

G. Valenza, A. Lanata, and E. Scilingo. The role of nonlinear dynamics in affective valence and arousal recognition. *IEEE Transactions on Affective Computing*, 3 (2), 2012.

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

A. van Boxtel. Optimal signal bandwidth for the recording of surface EMG activity of facial, jaw, oral, and neck muscles. *Psychophysiology*, 38(1), 2001.

Vayyar. https://vayyar.com/automotive, 2019. Accessed: 2019-07-24.

Vivalnk. http://vivalnk.com/, 2017. Accessed: 2018-01-09.

Vokaturi. https://vokaturi.com/, 2019. Accessed: 2019-07-24.

J. Wagner, J. Kim, and E. André. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, 2014.

D. Watson, L. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54(6), 1988.

J. Wijsman, B. Grundlehner, and H. Hermens. Trapezius muscle emg as predictor of mental stress. In *Wireless Health 2010*, WH '10, 2010.

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

N. Willigenburg, A. Daffertshofer, I. Kingma, and J. van Dieen. Removing ecg contamination from emg recordings: A comparison of ica-based and other filtering procedures. *Journal of Electromyography and Kinesiology*, 22(3), 2012.

W. Wundt. *Vorlesung über die Menschen- und Tierseele*. 1863.

H. Yasufuku, T. Terada, and M. Tsukamoto. A lifelog system for detecting psychological stress with glass-equipped temperature sensors. In *Proceedings of the 7th Augmented Human International Conference 2016*, AH '16, 2016.

A. Zenonos, A. Khan, and M. Sooriyabandara. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In *PerCom Workshops*, 2016.

J. Zhai and A. Barreto. Stress detection in computer users through non-invasive monitoring of physiological signals. *Blood*, 5(0), 2008.

B. Zhao, Z. Wang, Z. Yu, and B. Guo. Emotionsense: Emotion recognition based on wearable wristband. In *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2018.

Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J. Leon Zhao. Exploiting multi-channels deep convolutional neural networks for multivariate time series classification. *Front. Comput. Sci.*, 10(1), 2016.

Z. Zhu, H. Satizabal, U. Blanke, A. Perez-Uribe, and G. Tröster. Naturalistic recognition of activities and mood using wearable electronics. *IEEE Transactions on Affective Computing*, 7(3), 2016.