Analyse von Methoden zur Emotionserkennung mit Wearables

An Analysis of Methods for Emotion Recognition via Wearables

Bachelorarbeit

Zur Erlangung des Grades
Bachelor of Science (B.Sc.)

eingereicht von
Beyza Cinar

Abgabedatum: 24.08.2022

Erstprüfer: Prof. Dr. Maria Maleshkova

Zweitprüfer: Florian Grensing

Universität Siegen
Lebenswissenschaftliche Fakultät
Lehrstuhl Medizinische Informatik mit Schwerpunkt mobile Gesundheitsinformationssysteme

Plagiatserklärung

Ich versichere, dass ich die schriftliche Ausarbeitung selbständig angefertigt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle Stellen, die dem Wortlaut oder dem Sinn nach (inkl. Übersetzungen) anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter genauer Angabe der Quelle (einschließlich des World Wide Web sowie anderer elektronischer Datensammlungen) deutlich als Entlehnung kenntlich gemacht. Dies gilt auch für angefügte Zeichnungen, bildliche Darstellungen, Skizzen und dergleichen. Ich erkläre mich damit einverstanden, dass die Arbeit mit Hilfe einer Antiplagiatssoftware auf enthaltene Plagiate überprüft wird und dazu auf dem Server des Plagiatserkennungsdienstes vorübergehend gespeichert wird. Zudem versichere ich, dass die elektronische Version mit der gedruckten Version inhaltlich übereinstimmt. Ich nehme zur Kenntnis, dass die nachgewiesene Unterlassung der Herkunftsangabe als versuchte Täuschung bzw. als Plagiat gewertet und mit Maßnahmen bis hin zur Zwangsexmatrikulation geahndet wird.

Ort, Datum	Unterschrift

Abstract

Die automatische Erkennung von Emotionen mit Hilfe biologischer Signale ist ein sehr vielversprechendes Forschungsgebiet in den Gesundheitswissenschaften. Vor allem die fortschreitende Entwicklung von Wearables und Cloud-Computing ermöglicht eine kontinuierliche Erfassung der Daten und die Erkennung der Emotionen, welche helfen frühzeitige Diagnosen von psychologischen Erkrankungen wie Depression festzustellen. Ebenfalls können Therapiemethoden entsprechend dem psychologischen Wohlbefinden individuell angepasst/gestaltet werden. Gängige Sensordaten hierbei sind der Blutvolumenpuls, die Herzrate, die Herzraten-Varibilität, die Hautleitfähigkeit und die Hauttemperatur. Nach einer Filterung und (statistischen) Merkmalsextraktionen der Signale werden öfteres maschinelle Lernverfahren zur Klassifizierung benutzt (Support Vector Machines, K-Nearest-Neighbor, Random-Forest...). Neuerdings gibt es auch Forschungen für Deep-Learning Methoden wie Convolutional-Neural-Networks. Für die Emotionsklassifizierung gibt es zwei konkrete Emotionenmodelle, das diskrete, in welchem vorbestimmte Emotionen analysiert und das dimensionale, in welchem Emotionen als Kombination (Vektoren) aus mehreren Komponenten (Dimensionen) dargestellt werden. Hierbei ist das zwei dimensionale Model am gängigsten, in welchem eine Achse die Intensität und die andere die Polung der Emotion darstellt. In dieser Arbeit wurde mittels bereits durchgeführter Studien, welche das zwei dimensionale Model benutzt haben, unter Betrachtung der verfügbaren Sensoren analysiert, ob Wearables eine gute Basis für die Ermittlung von Emotionen darbieten. Die Analyse zeigt, dass Wearables vielversprechend sind und genaue Ergebnisse liefern können, jedoch müssen Daten sehr gut für die Klassifizierungsmethode vorbereitet werden. Zudem ist eine große Datenmenge und homogen verteilte Gruppe an Probanden notwendig. Es wird festgestellt, dass die Genauigkeit stark von den Probanden abhängt und Emotionen sehr subjektiv bewertet werden. Des weiteren scheint das vorgestellte zwei-dimensional Model nicht ausreichend zu sein und es wird eine Erweiterung vorgeschlagen, um bessere Grenzen zwischen ähnlichen Emotionen zu ziehen. Letzlich kann durch den Vergleich verschiedener Arbeiten angenommen werden, dass es nicht das genau Richtige oder die Beste Klassifizierungsmethode/Algorithmus gibt und für jede Datenmenge die beste Methode “erkundet” werden sollte.

Contents

Glossary I

Acronyms I

1. Introduction 1

2. Emotions 4
   - 2.1. Definition . . . 4
   - 2.2. History of Emotion-Research . . . 5
   - 2.3. Circumplex Model of Emotion . . . 6
   - 2.4. Relation with ANS . . . 9

3. Physiological Data 11
   - 3.1. Blood-Volume-Pulse . . . 12
     - 3.1.1. Heart-Rate . . . 13
     - 3.1.2. Heart-Rate-Variability . . . 15
     - 3.1.3. Inter-Beat-Interval . . . 16
   - 3.2. Skin Conductance . . . 16
   - 3.3. Skin Temperature . . . 18

4. Comparison Between Works 20
   - 4.1. Methodology . . . 20
   - 4.2. Different Setups . . . 22
   - 4.3. Induced Emotions & Used Sensors . . . 24
   - 4.4. Used Features & Classifications . . . 24
   - 4.5. Results & Findings . . . 28

5. Discussion & Future Works 33
   - 5.1. Reflection & Discussion . . . 33
   - 5.2. Future Works . . . 38

6. Conclusion 39

A. Appendix 45

List of Figures

1. Methods for Emotion Recognition . . . 2
2. History of Emotion Recognition . . . 5
3. 2-Dimensional-Model: (a) [DKB20] (b) [ELH19] . . . 6
4. The Functions of SNS and PNS . . . 10
5. PPG-Signal (a) Compared with ECG [DKB20] (b) Typical Pattern [GG19] 13
6. HR-Signal for Different Emotions [ROM10] . . . 14
7. The Process of HRV Computation [Zha+18a] . . . 15
8. IBI-Signal Captured by the E4 . . . 16
9. Signal of SC (a) both components [DKB20] (b) only SCR [DS21] . . . 16
10. Pattern of SKT (a) [DKB20] (b) [Wan+20] . . . 18
11. Effect of Expanding Dimensions . . . 35
12. List of PPG-Signal's Extracted Features I. [HS18] . . . 45
13. List of PPG-Signal's Extracted Features II. [HS18] . . . 45
14. List of GSR-Signal's Extracted Features I. [HS18] . . . 45
15. List of GSR-Signal's Extracted Features II. [MGK18] . . . 46
16. List of SKT-Signal's Extracted Features [HS18] . . . 46
17. Different Fear Levels [Băl+19] . . . 46
18. Comparison of 12 Studies: Complete table . . . 47
19. Feature-Set of Ayata et al. [AYK18] . . . 48
20. Features of Zhao et al. [Zha+18b] . . . 48
21. Best Features of Zhao et al. [Zha+18b] . . . 48
22. Differences in Gender [Den+16] . . . 49

Glossary

Convolutional Neural Networks a deep learning method, which takes images as input data and learns their specific patterns regarding the output. 24

Deep Learning a sub-area of machine learning, which can handle more complex problems and data, since it has more hidden layers. As a difference to usual neural networks humans only ensure that the data for learning is available and do not intervene. 24

K Nearest Neighbor a supervised machine learning method, which classifies data and forms groups with the assumption that near data-points in the same space belong to the same group. The difference between points are calculated through a distance function and clusters are adjusted in an (k times) iterative process [Al +19]. 24

Random Forest a decision tree based algorithm combining multiple decision trees with different characteristics, which can be used for high dimensional data. Decision trees portion data into groups as homogeneous as possible and predict the value of a target variable by learning decision rules from the data [AYK16]. 24

Subject/User-dependent the outcome is dependent on the user's parameters, and cannot be used generally. Hence, for every person calibration is necessary, before the system can categorize. Here, training and testing are performed on the same individual [Al +19]. 12, 27, 28

Subject/User-independent the outcome is general and the system can categorize every person. Here, parameters are normalized first, for which the average is calculated and subtracted from the individual signals. Thus, no user specific information is contained in the data. Furthermore, training and testing are performed on totally different groups/persons [Al +19]. 11, 27, 28

Support Vector Machines a supervised machine learning method, which can handle non linearly separable data [Al +19]. The input is labeled and the computation of an optimized hyperplane is the output. As a result, data not related to each other are separated [Rag+17]. 24

Acronyms

ANS Automatic Nervous System. 3, 9–11, 14, 15, 17–19

BVP Blood-Volume-Pulse. 11, 13, 38

CNN Convolutional Neural Networks. 24, 26, 31, 32, 39

DL Deep Learning. 24, 27, 37

ECG Electrocardiography. 1, 12, 13

EDA Electrodermal-Activity. 1, 2, 11, 16–18, 24, 26, 28–30, 36, 39

EMD Empirical-Mode-Decomposition. 25, 30, 31, 33, 36, 39

GSR Galvanic-Skin-Response. 11, 16–18, 24, 25, 29, 32, 33, 39

HR Heart-Rate. 9–16, 24, 26, 28, 29, 32, 36

HRV Heart-Rate-Variability. 9, 11, 15, 16, 25, 30

IBI Inter-Beat-Interval. 11, 16, 26, 28, 29, 38

KNN K-Nearest-Neighbor. 24–26, 28

PNN Probabilistic Neural Networks. 24, 26, 27, 32, 39

PNS Parasympathetic Nervous System. 9, 10, 13, 15

PPG Photoplethysmogram. 1, 11–13, 24, 25, 29, 32, 38, 39

RF Random-Forest. 24–26, 28

SC Skin-Conductance. 9–11, 16–18, 24, 25, 28, 29

SKT Skin-Temperature. 2, 11, 18, 19, 24, 25, 36

SNS Sympathetic Nervous System. 9, 10, 14, 17, 18

SVM Support-Vector-Machines. 24–29, 32, 36, 39

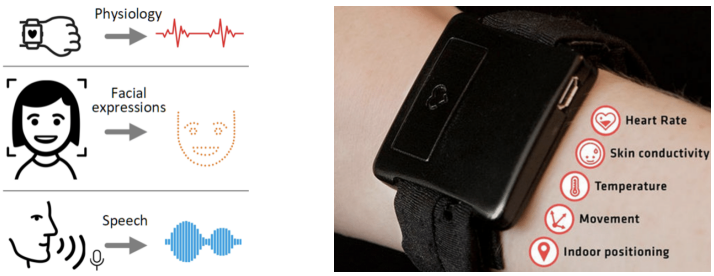
VR Virtual Reality. 22, 23, 29, 33

1. Introduction

In the past decade, understanding the emotions has played an important role in product recommendation as in music-/movie-streaming applications and also in designing human like robots for successful and comfortable human-machine-communication. Since then, researchers have explored various evocation, measurement and classification methods. Another key revolution is seen in the health care sector. It has gone through many reforms in the last years, and nowadays much is invested in digitization and knowledge-based-systems. Digital advances expand possibilities in detecting hidden diseases and can be used in almost every medical department. Consequently, clinical goals are changing from only curing the patient to preventing from diseases. Also life insurances promote disease preventing programs. Especially, elderly are monitored for a long time to ensure early detection and fast action. Furthermore, the human lifestyle and current trends have changed sharply, resulting in a way that more and more people want to be aware of and control their own health data. These progresses have been influential in the field of emotion recognition for medical use, since emotion recognition systems could diagnose mood-based diseases like depression, Alzheimer and Parkinson earlier. Besides, therapies could be adjusted individually, whilst recognizing the patient's fear level in phobia therapies would be possible. For instance, if the subject has a phobia of dogs or spiders, the animal could become calmer and go further away, or otherwise, nearer and more aggressive. Moreover, deaf people and children with autism could be helped since expressing emotions is a more difficult task for them. As a result, physicians would be aware of their patient's anxiety without burdening them during the examination or therapy. Thus, sensible people could be treated more suitable and individually.

Most of the research in this field approaches this challenge by studying speech and facial recognition. Nonetheless, these cannot be analyzed continuously, and a human might be able to control those parameters, resulting with inaccuracy. They can put on a poker face or still smile while being sad, the same goes for the control of voice. People often tend to say that they are fine with a calm voice, despite having a hard time. Hence, new perspectives are explored by considering physiological body signals, which cannot deceive easily because most of these parameters are beyond humans own control. Considering this problem, several physiological sensors have been tested like the Electroencephalogram (EEG), the Respiration-Rate (RSP), the Electrocardiography (ECG), the Electrodermal-Activity (EDA), the Photoplethysmogram (PPG), the Electromyography (EMG) and the Electrooculography (EOG) (most of them in laboratory or clinical setups) [DKB20].

Especially, EEG and EMG are supposed to identify whether the subject feels a positive or negative emotion as the distinction of sad and joyful. EDA and Skin-Temperature (SKT) would better detect how intense the person experiences the emotion, as the difference of being excited or calm, and further, cardiac activity as well as EEG would contain both information [ELH19]. However, most of these are not handy sensors and limited in comfort. Their preparation is elaborate, especially in time, and the subject cannot move freely without being conscious of the equipment. Consequently, they are obstructive and not suitable for everyday use, which is why they cannot be used outside controlled events [Rag+17]. These are also more expensive, since their accuracy and setup are of high medical level [Wan+20]. Particularly, recommendations and analysis estimated and calculated by own devices have become highly relevant in this matter. These could track physiological signals for a long time period and could alarm about serious conditions, allowing physicians to be updated with data in regular terms or only in emergency situations. These advances make individual adjusted monitoring possible.



(a) (b)

Figure 1: Emotion Recognition
(a) Methods [1] (b) Empatica E4 [2]

Wearables are a part of the field of ubiquitous computing, which deals with small computers augmenting daily life. It is a widely approached developing field nowadays and also studied for emotion recognition due to their affordable and comfortable characteristics. These can be used in easy setups and could track emotions during daily activity since the subject only needs to wear the device on their wrist like a regular watch. Furthermore, wearables can detect for as long as being worn and do not need complex preparations. Mark Weiser declares that *"the most profound technologies are those that disappear"* [Wei99], meaning that they are so intensely integrated in the lifestyle, that using them

[1] https://www.mdpi.com/2079-9292/11/3/496/htm, 27.07.2022
[2] https://www.researchgate.net/figure/Empatica-E4-wristband-physiological-signal-monitoring-14_fig9_322206805, 27.07.2022

gets a habit and is not considered as a task or burden. Hence, they are used unconsciously as an everyday tool without active consideration. The most important task of ubiquitous computing is that they operate in the background and not themselves, but their task and its output are relevant for the user [Wei99]. Another significant aspect of wearables is that they are always ready to be used. Therefore, the user and the physician should not take extra effort for the preparation and measurement of the sensors and their parameters, but only focus on the outcome (emotion) itself and it's reasons. As a result, wearables can expand the possibilities of emotion recognition applications, so that they can be used for measuring daily activity and the burden of measurement is taken away, giving more time in analyzing the actual emotion.

However, only few studies have directly dealt with emotion recognition via wearable sensors so far. The Empatica E4 (seen in figure 1) is a commonly used health wristband [Sch+19; Sag+20], but often other sensors' data are emerged. Furthermore, most of researches including wearables are focused on discrete emotions, whereas this study will explore the circumplex (dimensional) emotion model by Russel, which has a wider range of usage as later illustrated in section 2. In particular, no study, to our knowledge, has analyzed only in wearable integrated sensors, measuring emotions on a dimensional scope. As why, in this work approaches will be listed and compared, and it will be discussed if wearable sensors are sufficient. If possible, better methods will be recognized.

The structure of our work will be as follows: In section 2 emotions, the history of emotion recognition systems and the two-dimensional model will be defined, in which also the advantages of this approach are presented. Furthermore, the functions of the Automatic Nervous System (ANS) will be explained, as well as in which aspects they are related to emotions. Section 3 will introduce the most common wearable sensors, their data and relation with the ANS. In section 4, which is the focus of this work, 12 studies will be compared in detail to discuss the effect of various methods. Our aim is to state out whether current works and methods are satisfying and trustworthy in accuracy or need to be further researched and improved. Thus, we will analyze if wearables are qualified to measure emotions. We will answer if multiple sensors are more accurate than single ones and if more input-data and features increase the accuracy or not. Further, we will focus on the influences of the setup methods like the stimuli, the chosen participants and the chosen sensors. Consequently, we want to highlight outstanding methods, so that future works can build a system combining all introduced advantages and avoid represented mistakes. In the last section 5, described methods and possible projects will be discussed.

2. Emotions

2.1. Definition

Dzedzickis, Kaklauskas and Bucinskas explain that emotions are the human body's reaction to specific stimuli's activation. They are triggered through interacting with the environment, encountering a situation or particular circumstances of the person. Therefore, emotions do not last long, are very intense and the person acknowledges and senses them consciously [DKB20]. Additionally, Posner, Russel and Peterson explain that emotions inducing physiological changes are influenced by *"eliciting stimuli, memories of prior experiences, behavioral responses, and semantic knowledge"* [PRP05], demonstrating that they are subjective. These can be reactions to a joke or comedy clip, leading someone to laugh and be joyful or others to be bored. Whereas affect is a *"neurophysiological state"* indicated by emotions, with the main difference that it is not directly induced by any particular *"entity"* [Sag+20], but only a *"simple raw (nonreflective) primitive feeling"* [Sch+19]. Schmidt, Reiss, Dürichen and Van Laerhoven describe this difference with the example that a state of being angry arises quickly, but does not last long. However, exactly this emotion *"might lead to an irritable mood, which can last for a long time"* [Sch+19]. In contrast, feelings are individual reactions and depend on experiences with the event [DKB20]. For instance, someone can argue with a friend, be angry, but cannot hold a grudge, reflects the situation and thus, gets sad instantly. On the contrary, another person who does not have a particular relationship with the partner might be angry over a longer time, and may start to avoid this person. Lastly, mood is longer lasting and less intense. It affects the *"affective state"* and leads emotions to a *"positive or negative direction"* [DKB20]. Moreover, the person is often not conscious of their mood, since it happens in the background [Sch+19]. As a result, depressive mood arises sad and bored emotions, whereas anxiety leads to stress and could cause emotions like anger and fear. Another key point is that emotions could induce each other and that they could be felt simultaneously [PRP05]. Thus, affective states and emotions are dependent and should be considered together as related states, not as separate ones. Emotions can further be divided into primary and secondary. The first group consists of the six basic emotions [Al +19], which will be explained in the following subsection 2.2. Contrariwise, the second group are emotions arising through responses to primary emotions' experiences [BMT21c]. Knowing these distinctions helps to choose the most suitable setup for the targeted emotions. Hence, the chosen stimuli and time duration of measurement variate. Sadness and depression, for instance, may look very similar, as well as fear and anxiety,

but one is categorized as an emotion, the other one as a longer-lasting mood, respectively.

2.2. History of Emotion-Research

The research of emotion models had basically started long ago with roman philosophers like Cicero, who has classified into four basic emotions (*fear, pain, lust and pleasure*) [Sch+19]. However, Posner et al. present that the maintained discrete theory, in which emotions are labeled beforehand and categorized in primary basic emotions to approach separate affective states [Al +19], have been derived later from animal studies. They further report that researchers have claimed that specific neural pathways are connected with each basic emotion by observing the animal's behavior after stimulating the pathways [PRP05]. Nonetheless, animal studies could not include experiences and the animal's subjective thinking. Noticing these limitations, researchers had started studying on humans, allowing better interpretation and validation due to the participant's verbal responses, according to the views of Panksepp in 1998, which is cited by Posner et al.. Hence, inconsistencies between both approaches have revealed the importance of considering individual feelings [PRP05].

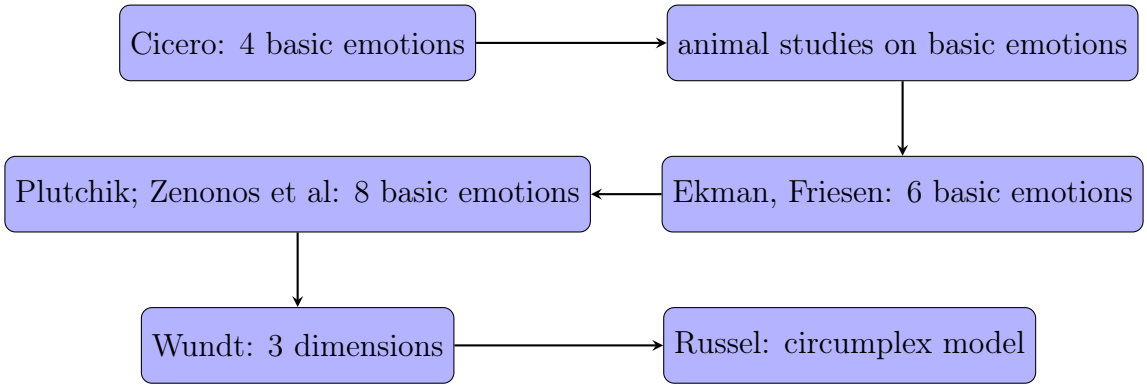


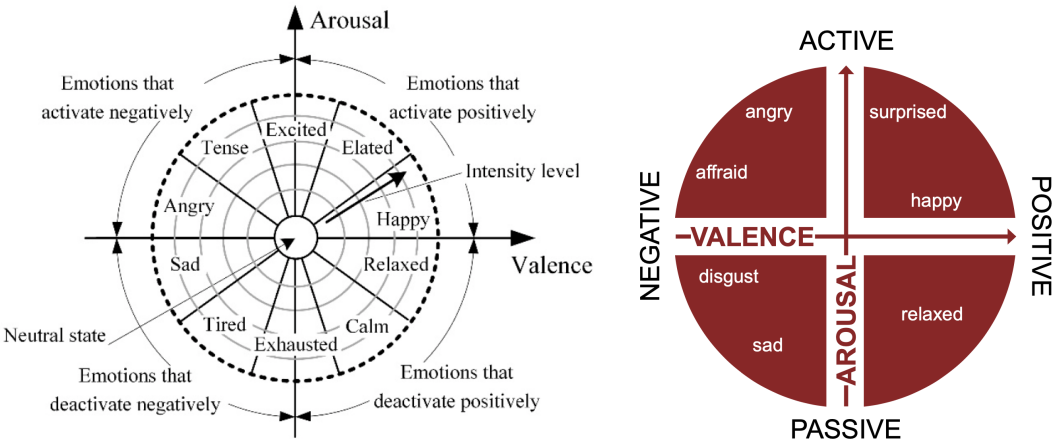
Figure 2: History of Emotion Recognition

Here, Ekman and Friesen's basic emotion model (1976), categorizing into six emotions (*joy, sadness, anger, fear, disgust, and surprise*) is the root of other discrete approaches [Sch+19; ELH19]. Following in 1980, Plutchik expanded the model to *"eight primary emotions: grief, amazement, terror, admiration, ecstasy, vigilance, rage, and loathing")* [Sch+19]. Later on in 2016, Zenonos et al. further categorized into *"8 different emotions and moods (excited, happy, calm, tired, bored, sad, stressed, and angry)"* [Sch+19]. Moreover, as introduced in section 1, facial expressions and peripheral physiological responses were researched by basic emotion theorists, assuming that patterns of autonomic

activation and facial innervation are specific to each basic emotion. However, these assumptions could not be proven and are criticized [PRP05]. In 1873, Wundt explored dimensional approaches and described an emotion as a three-dimensional point (with the axis of *"pleasure/displeasure"*, *"excitement/inhibition"* and *"tension/relaxation"*) [Sch+19; ELH19]. At the end of the 1970s, Russell highly suggested a two-dimensional model *(arousal/valence)*, which is the most known dimensional approach till date [Sch+19]. Additionally, infants' affective responses were studied, because researchers have claimed that the dimensional approach is not fit for infants, who cannot express themselves, since they do not have the necessary cognitive capacities. However, this model is highly compared to the animal studies since both are nonverbal and only limited to behavior [PRP05].

2.3. Circumplex Model of Emotion

As seen above, the two common approaches for emotion recognition nowadays are the discrete and the two-dimensional/circumplex model. In this section, we will focus on the dimensional approach, which portions the emotion into more *psychological dimensions*. Hence, the combination defines the final outcome [Al +19] as seen in figure 3.



(a) (b)

Figure 3: 2-Dimensional-Model: (a) [DKB20] (b) [ELH19]

In the two-dimensional model, the emotion consists of an arousal and a valence space, which are both *"two independent neurophysiological systems"* [PRP05]. Here, arousal means the intensity level (passive to active), and valence the distinction between positive and negative emotions [ELH19], and the output of both spaces' linear combination is

described as the emotion [PRP05]. So, emotions are projected into a two-dimensional space as a discrete point (vector) [Sch+19], in which the axes' values are discretely labeled [Sag+20]. Another aspect of these components is that arousal as well as *"single valence can have multiple levels"* [Sag+20], allowing the analysis of intensity levels of one emotion. Referring to the views of Posner et al., it is assumed that all affective states are the result of *"a complex interaction between cognitions"* [PRP05]. These assumptions illustrate that single emotions are more complex and consist of multiple aspects and components influencing them. As why, these influences should be considered and measured when estimating emotions.

If having a two-dimensional space with one axis representing the arousal and the other one the valence level, four quadrants can be defined. A Low Arousal High Valence (LAHV), a Low Arousal Low Valence (LALV), a High Arousal Low Valence (HALV), and a High Arousal High Valence state (HAHV). These *"can be attributed with sad, relaxed, angry, and happy"* [Sch+19]. Some also add more emotions in one quadrant like angry, fear and being nervous. Notably, disgust is a confusing emotion, since some studies point it out as a LALV and some as a HALV state. The significant point in this model is that it simplifies emotion classification and is not limited to static values. The quadrants can be taken as a basis for a field of emotions, which are similar to each other in physiological aspects. In addition, quadrants can have a primary emotion and sub emotions [Dis+19]. It should also be noted that there are affected states which are so similar to each other, that based on arousal and valence, a distinction is very hard to impossible, since similar emotions have the same value in a quadrant. In the HALV quadrant emotions like fear and anger can be assumed, which are different emotions, sometimes felt simultaneously or affect each other. However, the system would define all in the same quadrant and here, depending on the viewed research, the vector-value of these emotions differ. For example in figure 3, anger is illustrated at different arousal levels in both figures. Consequently, it is not an easy task to separate similar feelings within the same quadrant correctly. A way to overcome this problematic could be using very impactful stimuli or adding dimensions. In extension, a model with 9 categories is also common, in which the neutral state is considered, for more detailed classification. But the more emotions and groups are considered the less accurate the system will possibly become [Al +19; AKJ21]. Moreover, with a dimensional perspective, it is possible to focus on one emotion's different levels like no fear, little fear, moderate fear, pretty much fear as seen in figure 17 [Băl+19].

It should be noted that one of the positive and most useful aspects of the dimensional approach is that more axes, hence, more emotions can be added as proposed above, since

emotions represent a vector in dimensional space. The model is very flexible and can be modified, adjusted and improved depending on the purpose. Therefore, additional spaces could support the distinction between similar emotions like anger and fear, which only have a slight difference in arousal and valence, but are completely different in terms of dominance (anger is a dominant emotion which is being controlled by the person itself, whereas fear arises spontaneously [Băl+19]). Another significant consideration is the extension of a time dimension, which would distinct sadness from depression and fear from anxiety, or a space for the likability.

The most critical disadvantage for an automatic emotion recognition system based on the dimensional approach is that the model needs to be trained depending on the user's self-evaluation of the perceived contribution of arousal and valence. However, the reports are very individual for each stimuli and person and influenced by various aspects like sociological stereotypes, culture and experience. Also, it is a difficult process to evaluate own emotions and feelings, especially if the emotion is split into different components. The same complication can be seen in case history for diagnosis, in which the patient is asked to define their level of pain. Some tend to exaggerate and some to understate. So, even in simple questionnaires, people do not always know how to rate themselves and how honest they want to be.

Contrariwise, the discrete labeling of emotions only requires the felt emotion, which is much easier to report. Nonetheless, it is limited to predefined emotions and cannot be expanded. Each emotion is restricted to a single word, so that emotions are all independent from the other. But, emotions are more complex to be summarized in only one word. Often multiple emotions are induced simultaneously or respectively arise each other, as explained above. It can even happen that emotions of two different quadrants arise, simultaneously. The discrete model does not consider these influences, hence, is more a one-sided model for each basic emotion. In addition to this critical perspective, Posner et al. declare the need of moving away from a static discrete model to dimensional models. They further claim that humans find it challenging to differentiate exactly between emotions and to be fully aware of which and how many emotions are felt in total [PRP05], inducing that for instance watching a horror movie, one often feels fear, but is amused and tensed/excited at the same time. The discrete model would not consider the involvement of multiple emotions or the distinction from a mood. Although not describing the feeling necessarily with words, the distinction between quadrants in the circumplex model can be sufficient to declare the affective state and the emotion. Additionally, the observation is shared that subjects often describe positive emotions in

relation to other positive ones [PRP05].
When comparing both methods, the discrete suits well, if only some particular emotions are considered by the system as in phobia therapies, in which one only wants to recognize if the emotion (fear) is present or not, thus, in binary evaluations. But in more detailed applications, which analyzes more emotions and more levels like the intensity of fear in a specific range, the (discrete) dimensional model suits better. Here, the developer is free to expand the model and set the number of dimensions and emotions. Furthermore, we claim that mood based diseases like depression are more likely to be detected and monitored by dimensional models, with the involvement of arousal, valence, time and dominance spaces. Since more aspects, influencing the mood and illustrating the evolution of the disease are considered and measured compared to the discrete model.

2.4. Relation with ANS

Emotions' physiological response are mostly controlled by the ANS, ruling over internal organs including heart activity, skin conductance, blood pressure and the digestion system[3]. Hence, Kreibig predicates that it manages the human organism and allows the extraction of human behavior, since the brain structure mostly controls the activity [Kre10].
The ANS consists of the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS), which interact with and regulate each other. Being excited as in a "fight and flight" situation and experiencing intense feelings, the SNS activates and the organism responses with increased parameters. As a result, more adrenaline, hormones and energy (glucose) are provided. Thus, high activity of the SNS regulates Skin-Conductance (SC), Heart-Rate (HR) and Heart-Rate-Variability (HRV) [Sch+19]. The HR and blood pressure increase through the *"constriction of blood vessels and bronchial dilation"*[4]. On the contrary, the PNS is triggered when normalizing the parameters as in the *"rest and digest"* state. The person feels relaxed, is calm and the bodies' signals become moderate again. Thus, a decrease in the related physiological signals can be observed [Sch+19; ELH19; DKB20]. This is why emotions' physiological responses are lead by the interaction of PNS and SNS and can be measured via sensors.
Despite knowing their own emotions, humans barely have any influence on the regulation of physiological signals, because it is almost impossible to control the ANS, reported by Schmidt et al. [Sch+19]. These signals are directly governed by the nervous- and

[3] https://psu.pb.unizin.org/psych425/chapter/744/, 10.06.2022
[4] https://wtcs.pressbooks.pub/pharmacology/chapter/4-2-ans-basics/, 10.06.2022

endocrine system as seen in figure 4, thus, by instinctive responses to stimuli and not by *"subjective thinking"* [Sch+19]. These observations induce that parameters measured by sensors can be an accurate source for emotion recognition.

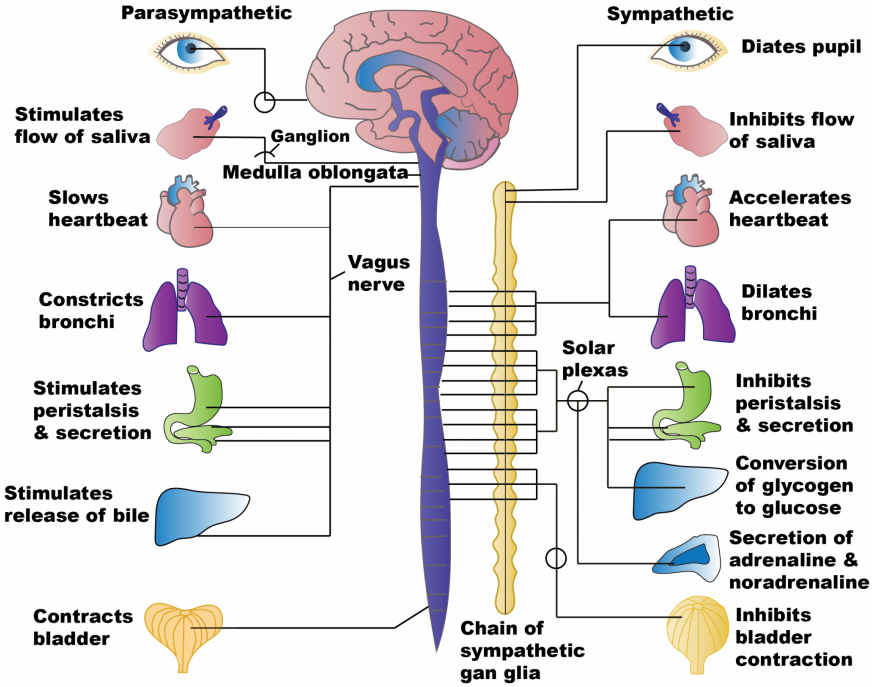


Figure 4: The Functions of SNS and PSN[4]

Lastly, it needs to be mentioned that the SNS and PNS can operate at the same time or simultaneously[3]. Hence, it is challenging to acknowledge which system gets triggered exactly by the targeted emotions. Here, the example is presented that the heart activity can increase due to activation of the ANS or sudden decrease of the PNS. Eventually, it is possible that only specific different organs are targeted at the same time by both systems as a normal HR but increased SC[3], which makes the measurement a difficult task.

3. Physiological Data

Nowadays, sensors are getting smaller and smaller and fit in wearable devices like health-wristbands or smartwatches, allowing the assessment of ANS's activation easily worn from the wrist. Especially the Galvanic-Skin-Response (GSR)/ EDA, the PPG and a skin-thermometer are seen as the most effective sensors. These can measure the Blood-Volume-Pulse (BVP), from which HR, HRV and Inter-Beat-Interval (IBI) are computed, the SC and the SKT, respectively. All parameters are out of humans own control and mostly regulated by the ANS as illustrated in subsection 2.4.

Sensordata	Increased	Decreased
BVP	excited	relaxed
HR	anger, anxiety, fear, embarrassment, crying sadness, pleasure, happiness, joy, surprise	disgust, imminent-threat fear, non-crying sadness, suspense
HRV	stress, frustration, contamination-related disgust, acute sadness, amusement, joy, potential state of mental stress	relaxed, happiness, visual anticipatory pleasure, psychiatric disorders
SC	all other	acute or non-crying sadness, pleasure, relief
SKT	relaxed	anxiety, stress, anger, embarrassment, humiliation, joy with anxiety, depression with hostility, sadness

Table 1: Physiological reactions to emotions according to the works of: [Kre10; DKB20; ELH19]

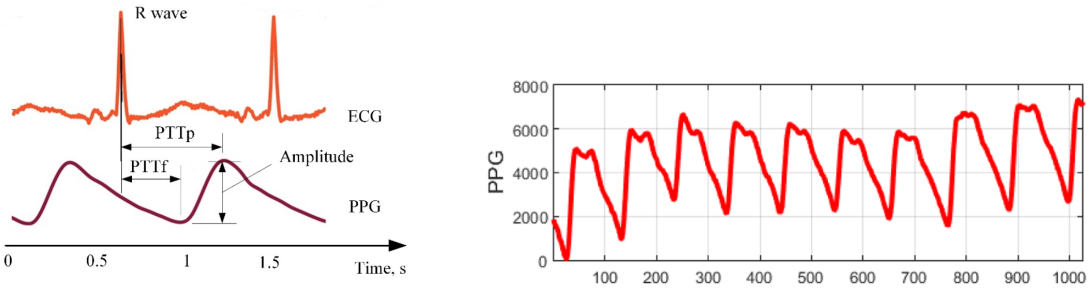
However, even if the measurement itself is a convenient task, an emotion recognition application is more complex due to the subject's high dependence. The responses can change with the individual and are influenced by various factors including age, gender, experiences, the state of health and the social environment [Al +19]. If a general system, which is Subject/User-independent, is targeted, the parameters need to be normalized, after acquiring the data in order to narrow down "the individual's impact" [Wan+20].

Therefore, the system does not have to calibrate for each user individually as in a Subject/User-dependent system, before making decisions about the affective state [Al +19]. A dependent model is more accurate in most studies, because it learns the behavior and parameters of the specific subject and adjusts to the measured values. Nonetheless, even the same person can react differently to same emotions. Furthermore, it is not practical for most cases, since it would take much time to evoke and capture relevant responses to stimuli for the calibration [Al +19]. in addition, filtering, scaling and noise reduction is required for accuracy improvement. Especially, sensors in wristbands are challenging due to high motion noises. They can depend on factors like skin color and diseases as blood circulation problems [Sch+19]. Also we have heard that the skin temperature can be influential as well, since some sensors like the PPG could not always measure well on cold skin. After data cleaning, the relevant features are computed, which often are statistical values in the time and/or frequency domain. These are then represented as feature-vectors, therewith ensuing that the data can be used as input for the learning algorithm. Later, the model is trained to classify and the estimated data-signals are split into training and testing data. Thus, before starting the measurement, necessary preparations need to be done, and the researchers should aim for a dependent or general system beforehand according to their purpose.

3.1. Blood-Volume-Pulse

The cycle of cardiac activity has two different phases, whose harmonized interaction regulates the heartbeat and the blood flow through the vessels to all organs[5]. These are the systole (state of activation), in which the heart contracts to pump the blood, and the diastole (state of relaxation) after the contraction. The heartbeat is triggered by electrical impulses, which can be measured via electrodes as in the ECG[5]. In addition, the cardiac activity can be identified with the reflection of light as in the PPG sensor [DKB20]. In the systolic phase, more light is reflected by the skin and absorbed back by the sensor, whereas in the diastolic phase the skin absorbs more light. Hence, more blood in the vessels result with more light absorption and the HR increases with the systolic and decreases with the diastolic phase, reflecting the interplay of contacting and relaxing as seen in figure 5.

[5] https://my.clevelandclinic.org/health/articles/17064-heart-beat, 10.06.2022



(a) (b)

Figure 5: PPG-Signal
(a) Compared with ECG [DKB20] (b) Typical Pattern [GG19]

Although being noisy in comparison, the PPG can still compete with the ECG's accuracy. Accordingly, a correlation of up to 88% is possible [ELH19] and studies of McCarthy and collaborators promise that PPG signals obtained from the Empatica E4 are sufficiently precise for cardiac activity assessment, cited by Ragot, Martin, Em, Pallamin and Diverrez [Rag+17]. Besides, it is reported that the Empatica E4 combines a red and a green light to remove motion related artifacts from the BVP signal [Sax+20] and as seen in figure 8, the E4 algorithm can detect correctly classified heartbeats (green points) from declassified (red crosses) ones. Goshvarpour and Goshvarpour report a study, which has compared the accuracy of HR measured by PPG and ECG and obtained an average cross-correlation of 99.17% [GG19].
The main difference to the ECG is that the PPG sensor measures blood volume changes, referred to as the BVP and not the pulse itself directly. Blood flows continuously through the vessels with a varying flow (volume), from which the heartbeat can be extracted [ELH19]. Hence, the BVP is dependent on the heart's activity, regulating the pulse-wave and the flow of blood [ELH19]. In figure 5, the signal pattern is illustrated in relation to the usual PQRST-wave. Furthermore, it can be noticed that in each cycle two peaks are present, referred as the main and secondary peak, representing the volume pulse wave and the pressure pulse wave, respectively [Wan+20]. Additionally, figure 12 and figure 13 illustrate possible features which can be extracted by the PPG sensor.

3.1.1. Heart-Rate

The HR describes the number of heartbeats in a minute and can be computed by the PPG through extracting the intervals between neighbor peaks [Sax+20]. Kreibig points out that the short term HR is regulated by the interaction between the PNS and

SNS. By contrast, the long term modulation is induced by the endocrine system, which produces noradrenaline, released in the blood flow [Kre10]. Noradrenaline contracts vessels, resulting with increased blood pressure. Both parameters can be influenced, but not regulated by the individual [Kre10]. For example, if someone is afraid of spiders, she/he can tell herself/himself that they should not fear and try to stay calm, but the heart still might beat faster and the emotion fear is experienced. Only if the person gets used to the spider and overcomes the phobia with time she/he might control.

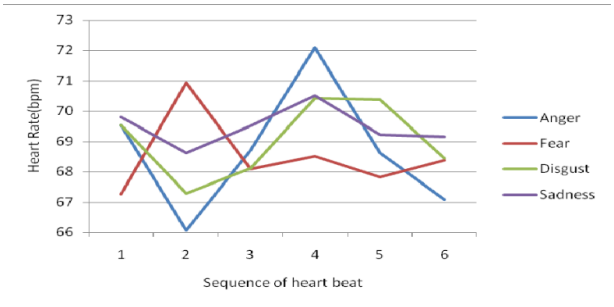


Figure 6: HR-Signal for Different Emotions [ROM10]

The HR can measure the valence composition, as Kreibig assumes. She supports her view with a study's insight illustrating how the participant's HR is suddenly faster while watching unpleasant films [Kre10]. Additionally, Bulagang, Mountstephens and Teo claim that HR represents the arousal level [BMT21b]. Thus, according to viewed researches, the HR holds both arousal and valence information [ELH19].
Furthermore, it is proven that the HR changes with the mood/emotion. Ekman et al. showed in 1983 that it increases with facing emotions like anger or anxiety and decreases "significantly with disgust", cited by Shu et al. [Shu+20]. In addition, Kreibig illustrates that the HR is increased for negative and positive emotions as well as for surprise. Also, she demonstrates that the HR is decreased in emotions like disgust, fear and types of sadness as well as suspense [Kre10]. Moreover, Rattanyu, Ohkura and Mizukawa share the observance that in an angry state the HR first decreases, then increases linearly before decreasing again. In comparison, when having fear the HR rises immediately as seen in figure 6 [ROM10]. Brittons' study further reveals that during a happy mood the HR is lower than in a neutral mood, cited by Shu et al. [Shu+20]. In interest is also the statement that the ANS's ability to regulate the HR decreases under negative circumstances like stress, tension and illness [Tak+21]. High physiological activity rises the HR [Tak+21], indicating that subjects with a history of diseases, likewise with specific conditions should not be chosen.

Lastly, Hui and Sherratt have discovered a correlation of HR variations with the ANS's activity by an accuracy of 87.4% [HS18].

3.1.2. Heart-Rate-Variability

The HRV represents time-related changes *"in each cycle of a heartbeat"* [DKB20] and detects irregularities and the time variation between them [Sag+20].

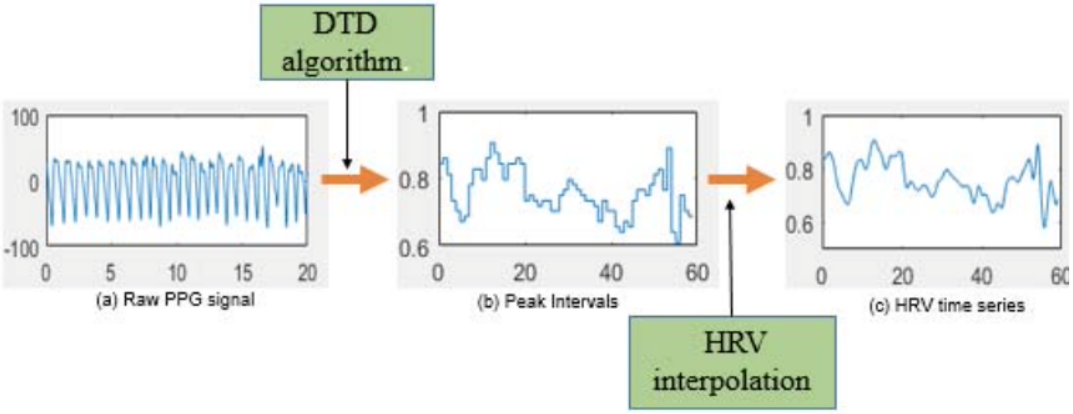


Figure 7: The Process of HRV Computation [Zha+18a]

Dzedzickis et al. report that out of regular pattern (increased HRV) are emotions like anger or moods as depression causing stress. (Additionally, it can be influenced by various other factors like health issues, genes, weight, age, gender and the use of tobacco, alcohol or caffeine.) By contrast, a regular beat (low HRV) demonstrates relaxed states like calmness or pleasure [DKB20]. Kreibig further observes that *"contamination-related disgust"* is the only negative emotion definitely followed by an increased HRV. In addition, she reports that acute sadness possibly could be characterized by increased HRV, but this assumption is not proven. Furthermore, the HRV rises with positive emotions like amusement and joy, whereas it drops with happiness and visual anticipatory pleasure. Hence, it is proposed that PNS's activity influences positive as well as some negative emotions [Kre10].

In addition, Egger, Ley and Hanke assume that a reduced HRV can relate to psychiatric illnesses as depression, anxiety and/or alcohol use disorders [ELH19], which present criteria for acquiring participants. Finally, referring to the statement of Kreibig, although features in the time domain are most often studied, the frequency domain should contain important data as well, especially, regarding the activity of ANS [Kre10].

3.1.3. Inter-Beat-Interval

The IBI signal is measured by computing the distance between two consecutive heartbeats in milliseconds [Sax+20; BMT21b]. Through the IBI the HRV and the HR can be estimated[6]. Unfortunately, no information regarding emotional data were found, but Bulagang et al. assert that it represents the valence state of emotions [BMT21b]. In the following section, a study exploring the IBI will be presented, as why a short introduction was seen necessary.

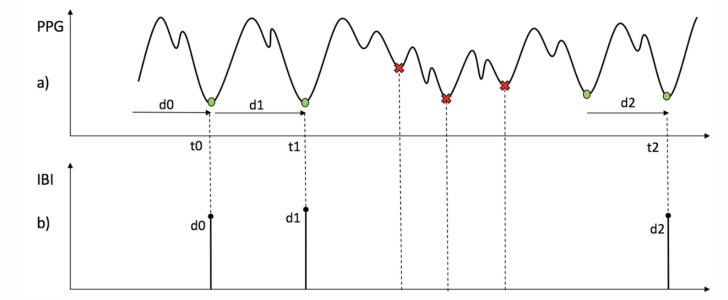
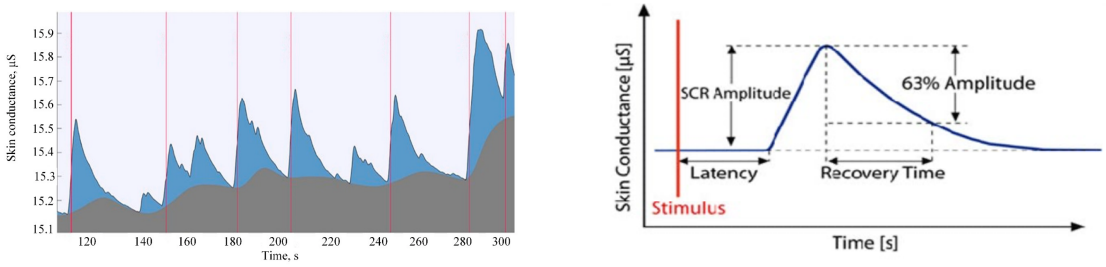


Figure 8: IBI-Signal Captured by the E4[6]

3.2. Skin Conductance

The human body has millions of sweat glands, whose activity can be measured with an EDA/ a GSR sensor. It detects the skin's electrical activation, changing with the variation of positive and negative ions' balance [DKB20] and the measured parameter is called the SC.



(a) (b)

Figure 9: Signal of SC
(a) both components [DKB20] (b) only SCR [DS21]

[6] https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal, 1.07.2022

Sweat glands activate with high physiological activity or intense psychological states as well as mood changes. Sweating increases the salt level, which increases the EDA. Thus, the skin surface gets moist, following changes in *"the electrical currents' flow property on the skin"* [DKB20; DS21]. It flows more readily when sweating, resulting in variations of the SC [DS21].

Row GSR data provides the tonic and phasic activity of the SC, which estimates the skin conductance level (SCL) and the skin conductance response (SCR), respectively [Sch+19; DKB20]. The tonic level determines the baseline SC, altering slowly over time and is individual for every person. It depends on the environment, temperature, skin hydration level and dryness [Sch+19; DKB20]. In comparison, phasic activity responses to the activation of the SNS and illustrates *"short term peaks"* of the EDA, which barely depend on the tonic level [DKB20; Sch+19; Dis+19]. Hence, the SCR can be measured by stimuli of really short duration [Kre10] and a short window size. So, it is claimed that it highly depends on the stimuli type, which has shown better results with acoustic stimuli [Al +19].

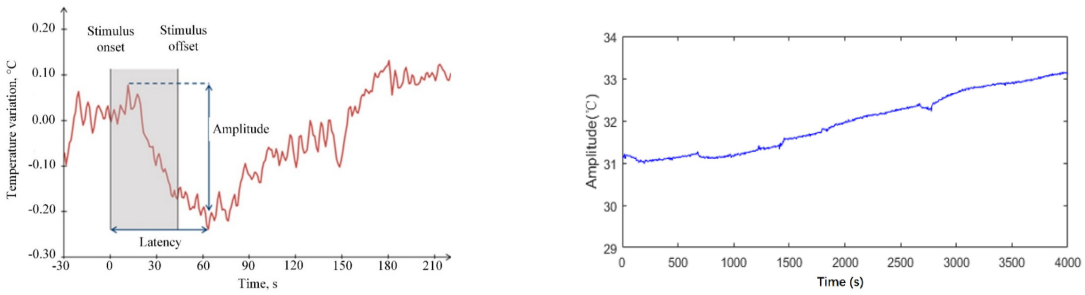
Furthermore, the components of the GSR define immediate increases (peaks), returning to baseline slowly. Thus, most information is captured by the amplitude (time) and frequency domain, usually through analyzing statistical features [DKB20]. Possible extracted features can be seen in figure 14 and in figure 15. The responses of the SC are split into stimulus (lower peaks) and spontaneous [Wan+20]. Wang et al. report that a typically sudden increase should last 1 to 3 seconds and returning to baseline would take longer with 2 to 10 seconds [Wan+20]. According to Zhao, Wang, Yu and Guo, it often happens that the EDA signal requires additional pre-processing. Especially, deep smoothing and signal separation could be necessary. Here, most often used are adaptive bandpass filters to remove artifacts [Zha+18b].

Moreover, Ayata, Yaslan and Kamasak claim that the resistance of SC decreases with increased sweat in aroused emotions like stress or surprise [AYK16]. Likewise, Dzedzickis et al. report that the signal's amplitude is related to high arousal states and stress, correlating with self-reported feelings of participants [DKB20]. Whereas Kreibig observes a decrease in sad states (acute or non-crying), pleasure and relief. All other emotions of her findings result an increased EDA [Kre10]. Another significant assumption is that the GSR could discover decision-making processes, since affective states contain attention-grabbing and demanding tasks, which follow a synchronized *"increase of the frequency and magnitude of GSR"* [DKB20]. In addition, Hui and Sherratt have reported a correlation of the SC's variations with the specificity of ANS by an accuracy of 95.8%

[HS18]. Hence, we notice that SC holds emotion-related information, especially regarding the intensity and the ANS's activity. Concluding, that the more intense the emotion, the faster the variations in SC can be detected. However, Desai and Shetty point out that the GSR is not capable detecting the exact sort of emotion, but only the existence [DS21]. Most researches claim that EDA variations are related to emotional arousal and can distinguish between relaxed and stressed states [HS18; Kre10; Sch+19].

3.3. Skin Temperature

SKT can be measured with an infrared thermometer and is controlled by the heart activity and sweat production [DKB20], but also environmental factors. Zhao et al. report that variations in the SKT mostly arise from localized changes in the blood flow, which are induced by "vascular resistance" or "arterial blood pressure". First, is regulated by smooth muscle tone, which is affected by the SNS. Second, is a model of cardiovascular regulation by the ANS [Zha+18b].



(a) (b)

Figure 10: Pattern of SKT (a) [DKB20] (b) [Wan+20]

The ANS controls the skin's hydration through regulating the vessels' activity [DKB20]. With the triggering of the SNS, the blood flow to the extremities can be restricted, resulting in changes in peripheral temperature [Sch+19]. Egger et al. present that during relaxed states dilated vessels can get warmer. In comparison, with the vessels' constriction during high aroused states, the individual might get colder, despite producing more sweat, which is more likely to be cold sweat in circumstances like anxiety and stress. Thus, the SKT can distinct stressed and relaxed states, similarly to the SC [ELH19]. It is said that (high aroused) negative emotions possibly lead to a decrease in finger temperature [DKB20]. A fall was also detected in a study when the subject was influenced by other's speech (feeling emotions like anger or anxiety), despite not

being involved himself/herself [DKB20]. Contrariwise, an increase was observed with the presence of less aroused negative emotions in comparison to less aroused positive emotions [DKB20]. Summing up, responses to stimuli can be measured by SKT in setups like watching movie clips, listening to music or discussions.

In the process of signal analyzing first, the measured SKT needs to be converted into an discrete electrical signal, then the arousal can be categorized into five states to identify which state should be low, medium and high aroused [ELH19]. Features like the minimum, maximum and average temperature can be extracted from longer measured signals, more possible features can be seen in figure 16. Consequently, it is noticed that one of the negative points of this method is the need of a window size with longer duration, since the temperature needs time to change. Hence, stimuli of short duration like pictures are not suited and effective [DKB20]. Referring to the views of Dzedzickis et al., SKT can detect arousal well, but is still not as sensitive to valence recognition [DKB20]. This statement is further claimed by Egger et al. [ELH19]. However, Hui and Sherratt assume that fingertip temperature helps to differentiate between pleasant and unpleasant emotions [HS18]. Moreover, Kreibig assumes a decrease in SKT when feeling variations of sadness [Kre10].

To conclude, there are different opinions about the dimension detection capabilities of SKT and it may vary from the measuring body part. Hui and Sherratt have captured fingertip temperature and as a result have discovered that SKT can measure the ANS specificity by an accuracy of 96.4% [HS18]. Here, the fingertip might be the most sensible, but is not suitable for wearable wristbands, as why we assume that SKT measured from the wrist can differ.

4. Comparison Between Works

4.1. Methodology

Only studies focusing on wearable sensors, introduced in 3 were chosen for this work. We have restricted our comparison to only dimensional emotion models after comparing both in section 2, since it was recognized to be more flexible and advantageous for a general analysis. Further, we have excluded works which do not testify significant results or have too few participants and have eliminated studies which have fused additional data/used clinical sensors. Moreover, we have tried to choose more studies of the same research groups to see improvements within the same setup and research. We have also tried to show different results regarding the used emotion model and did not want to focus on the four quadrants approach only. 12 works are arranged in table 2, listed after their release date, so a difference in accuracy within the years can be noticed. In this section we will compare these in the following order: First, we will sum up their setups and note most significant points, further, the induced emotions and the used sensor-data will be presented. Moreover, we will compare the classification and feature extraction methods, and finally present the results of each study. So works will be compared from the worst to the best accuracy. We aim to highlight better methods and want to point out the significance of each viewed aspect with this analysis. So that we can make suggestions for further researches if needed. A more detailed table of our comparison can be seen in figure 18.

[7] Abbreviations:
a= arousal; v= valence; s-d= subject-dependent; s-i= subject-independent;
EDG= Electrodermography;
img-trans.= image-transformation; f.e.m= feature extraction methods
DEAP: open database with 32 subjects and 40 music video clips [GG19]

Document	Setup	Data	Emotions	Methods	Features	Results
[AYK16], 2016	DEAP	SC	4 class	RF	14	81-85% a, 82-89% v
[Rag+17], 2017	19 subjects, 45 pictures	SC, HR (E4)	9 class	SVM	9	70% a, 66% v (s-i)
[AYK18], 2018	DEAP	SC, BVP	4 class?	RF	22	72.06% a, 71.05% v
[Zha+18b], 2018	15 subjects, self-chosen movie clips	SC, BVP, HRV, SKT (E4)	4 class	SVM	28	76% (s-d)
[GG19], 2019	DEAP	SC, BVP	4 class	PNN	3 f.e.m.	88.57% a, 86.8% v (s-i)
[Al +19], 2019	DEAP	SC	4 class	CNN	raw data	85% s-d, 82% s-i
[AKJ21], 2021	16 debate sessions with pairs	SC, BVP, HRV, SKT (E4)	4 class	fine KNN	N/A	87.80%
[AKJ21], 2021	16 debate sessions with pairs	SC, BVP, HRV, SKT (E4)	9 class	fine KNN	N/A	75.80%
[BMT21c], 2021	20 subjects, 16 videos (VR)	HR (E4)	4 class	SVM	3	46.7% (s-i)
[BMT21a], 2021	10 subjects, videos (VR)	EDG/SC, HR (E4)	4 class	SVM	raw data	66% (s-i)
[BMT21b], 2021	24 subjects, 16 videos (VR)	HR, IBI (E4)	4 class	SVM	N/A	67.4% (s-d)
[Nas+21], 2021	80 subjects, self-report	HR (MB 2)	valence	CNN (SVM)	5 img-transf.	$> 91\%$

Table 2: Comparison of 12 Studies[7]

4.2. Different Setups

Ayata et al., Al Machot, Elmachot, Ali, Al Machot and Kyamakya as well as Goshvarpour and Goshvarpour use the DEAP data-set [AYK18; AYK16; Al +19; GG19], which measures various body signals of 32 subjects (16 females, 16 males) between 19 and 37 years old [GG19]. 40 one-minute music video clips were used to evoke the emotions and the participants have rated valence and arousal. Here, 63 seconds of signals were acquired for every video in total [AYK16; GG19].

Furthermore, Zhao et al. measure on 15 participants (9 males and 6 females), who are between 22 and 28 years old and have none diseases, that could influence the parameters [Zha+18b]. Additionally, the subjects are asked to avoid food/beverage, which could affect the measurement. As a difference to DEAP, a video bank of 20 five-minute movie clips is offered, in which the subjects can choose which video to watch. There is a one-minute break after each video to report the emotion via a questionnaire. The data-set consists of different genres including comedy, documentary, horror and war. Thereby, the clips are split into four different variations of targeted emotions: happy (4 videos), sad (4 videos), fear (2 videos) and anger (2 videos), with the aim to only evoke one emotion [Zha+18b].

Bulagang et al. focus on 16 360° video clips in a Virtual Reality (VR) environment as stimuli, which are 6 minutes and 5 seconds long in total, including rest periods of 10 seconds to recover from the emotion [BMT21c; BMT21a; BMT21b]. 3 works with the same setup, but different sensor-data and number of participants are done by the same researchers. In the first one 20 healthy subjects (12 males, 8 females), who do not have any heart disease history, between 20 and 28 years, currently working and/or studying, participate [BMT21c]. In the second one 10 healthy subjects [BMT21b] are present, and lastly 24 subjects participate [BMT21b]. As seen, only in the first study details about the subjects are revealed.

Only the study of Ragot et al. presents a set of 45 pictures to induce emotions, which are randomly shown on a display [Rag+17]. They measure parameters on 19 participants (12 women, 7 men) with an average age of 33,89, a minimum age of 23.49 years and a maximum age of 52.46 years. All are said not to have diseases, which could influence the parameters and have not taken any somatic drugs. Valence and arousal are balanced for each picture rated from three arousal and three valence levels, then five pictures of each category are presented with the condition that two of the same subcategories are not shown consecutively. Thereafter, the emotions are reported via Self-Assessment-Manikin (SAM) within 15 seconds and a black fixation cross on a white screen is displayed in

order to bring the emotion to baseline values [Rag+17].
Alskafi, Khandoker and Jelinek try a different approach beyond a controlled setup. They decide for the "K-EmoCon" data-set, in which subjects debate in pairs for 10 minutes, which should naturally induce specific emotions [AKJ21]. The participants consist of male and female students between 19 and 36 years, but the exact numbers are not announced. In total, they record 16 debate sessions [AKJ21].
Lastly, Nasrat et al. research emotion stimuli in natural life with 80 healthy subjects consisting of male and female between 19 and 40 years [Nas+21]. The participants live normally in their daily surroundings and report the emotional experiences at random times for 7 days by rating the state of arousal and valence on a 7-point scale with the Experience Sample Method (ESM) [Nas+21].
To conclude, the number of participants varies from 10 to 80 and their combination of age and gender also differ. We notice that almost only students are acquired. The maximum age within the studies is 40/50 and the majority seems to be around 20 years. Hence, we do not think that the group is evenly distributed and that elderly are considered. Furthermore, not every study reveals the same amount of information about their participants which does not allow a direct comparison. But, when targeting a general recognition system, it is critical to have a biased group of gender and age, since the likelihood for overfitting increases. Especially, with a small number of participants, a slight difference in the combination could possibly affect the results, as why the studies of Ragot et al., Bulagang et al. and Zhao et al. [Rag+17; BMT21c; Zha+18b] are seen critically. In addition, it is noticeable that videos are the most popular method for emotion evocation and are seen more powerful, because the subject is confronted with visual and audio stimuli at the same time. But while sitting and watching different clips for a longer time, without much possibility for relaxation and movement, the subject could be out of focus and feel forced. Therefore, the VR environment seems very promising, allowing the subject to experience the virtual environment more intensely. Contrariwise, pictures possibly depend more on the individual's experiences and are seen critically. Natural environments out of control can become more complex and elaborate because a longer observation is needed to capture every desired emotion, and there is a strong dependence on the subjects. Nonetheless, it can be assumed that the ratings are more reliable, since the participant does not have any pressure. The most critical problem in all these setups is the self-rating of emotions by participants as explained in 2. But there is no other known method to get a response from the user.

4.3. Induced Emotions & Used Sensors

Eight studies in this analysis focus on the four-quadrant model [AYK16; Al +19; Zha+18b; AKJ21; BMT21c; BMT21a; BMT21b; GG19], whereas two approach the nine-class one [Rag+17; AKJ21]. Ayata et al. focus only on separate arousal and valence levels with their second experiment and do not exactly emerge them. But considering their first study, we suppose that they have used the four-class one. However, this cannot be assumed for sure. [AYK18; AYK16]. Furthermore, Nasrat et al. only concentrate on binary valence level and do not take into account the arousal space. They aim to find out if the participant feels negative or positive emotions, but the intensity is not relevant [Nas+21].

Some studies only use single sensors, others rather research on the effectiveness of mutliple sensors. We see that the PPG sensor is the most popular, represented in a total of 10 studies, followed by the EDA in 9 studies. SKT is only analyzed in 3 works, of which two are from the same environment. However, the SC is measured more often, followed by HR as seen in table 2.

4.4. Used Features & Classifications

Most studies use supervised learning methods, since the data is labeled (input as well as the output is known). Support Vector Machines (SVM) are most often utilized in a total of 5 studies [Rag+17; Wan+20; BMT21c; BMT21a; BMT21b]. Followed by Convolutional Neural Networks (CNN), a Deep Learning (DL) method, in 2 studies [Al +19; Nas+21]. K Nearest Neighbor (KNN) as well as Random Forest (RF) are each used in two studies: [AYK18; AYK16; AKJ21; AKJ21], respectively. Lastly, Probabilistic Neural Networks (PNN) are only explored by Goshvarpour and Goshvarpour's work [GG19].

Ayata et al. explore 4 different approaches (RF, KNN, Decision Tree and SVM) to choose the most accurate learning algorithm, which ends up to be Random-Forest (RF) [AYK16; AYK18]. They split each sensor's data into sub-signals, then compute features in the time domain for each sensor's signal. Each sub-signal's features are concatenated into one feature-vector, used as the input for each emotion's learning process. In addition, they test four feature-sets (10, 14, 18, 22), which can be seen in more detail in figure 19, and feature-set-14 and feature-set-10 seem to be the most accurate for GSR and PPG, respectively [AYK18]. Also, the window size duration between 1 and 60 seconds are experimented. The best results are achieved with 3 and 8 seconds for GSR and PPG,

respectively. Additionally, they analyze signal processing with and without convolution, resulting that convolution increases the accuracy, especially in GSR's valence. Finally, the system is tested with the 10-fold-cross-validation [AYK16]. Same procedures are done for the GSR sensor only, in their study before. However, Empirical-Mode-Decomposition (EMD) is used before feature extraction to increase the accuracy [AYK16]. Eventually, discrete wavelet transformation are compared with the accuracy of time domain features. This approach is advantageous for non-stationary signals and can handle noise better than the usual frequency domain features [AYK16]. However, time-based features result to be more reliable [AYK16].

Similarly, Zhao et al. explore four algorithms (RF, Neural Network, Support-Vector-Machines (SVM) and Naïve Bayes) and use the leave-one-out-cross-validation for each classification [Zha+18b]. Overall, SVM give the best results. A window size of 16 seconds is applied before filtering and normalizing the signal to eliminate noise. They study feature extraction and feature-vectors' influences, therefore extract time, frequency and nonlinear features. Then, the extracted features are filtered via the sequence forward floating selection (SFFS) and Information Gain (IG) method to select the most relevant ones relating emotions [Zha+18b]. In total, 28 features are calculated from all sensors (6 for PPG, 12 for HRV, 6 for SC and 4 for SKT) and for all obtained instances the SFFS+SVM is applied [Zha+18b]. The extracted features can be seen in figure 20 and the result of their filtering in figure 21. Lastly, for system evaluation the 10-fold-cross-validation is applied [Zha+18b].

Likewise, Alskafi et al. analyze different classification methods (decision tree, SVM, K-Nearest-Neighbor (KNN), Kernel Naive Bayes as well as ensembles classifiers) and decide for fine KNN, which has the highest accuracy for the four and even nine-class emotion classification [AKJ21]. They implement 5-second segments for each data, but no detailed inside about extracted features are given.

Ragot et al. extract nine features in total (HR, AVNN, SDNN, rMSSD, pNN50, LF, HF, RD, AVSCL) and apply SVM to classify the emotions [Rag+17]. Furthermore, cross-validation is used and data is divided into training (80%) and testing (20%) to improve the accuracy. In addition, they train two separate models: one for valence and one for arousal [Rag+17].

Also, Bulagang et al. implement SVM in all of their works after testing three different classifiers (SVM, KNN, RF) in their first study [BMT21c]. The average BPM (beats per minute), maximum, and minimum for individual data signals are extracted. Furthermore, subject dependent and independent systems are approached in this study and they have

used the 10-fold-cross-validation [BMT21c]. In their second work input data consists of raw data without any extracted features [BMT21a] and lastly, in their third, in which the IBI and the HR are fused, no specific information about the input data is given [BMT21b], consequently, raw data can be assumed. Hence, neither Ragot et al., neither the two other studies of Bulagang et al. search after the best classification methods in relation to their data-set. But, only because SVM were the best method for a data-set with only HR measurements does not imply giving the best solution for the other cases. On the contrary, Al Machot et al. explore Convolutional Neural Networks (CNN), in which raw data is given after the input-data is transformed into matrices [Al +19]. Additionally, the system is tested with two different databases: DEAP and MAHNOB. This study aims to study the relevance of Deep-Learning-methods in emotion's signal analyzing, to reveal if learning from images are more accurate than extracting features by hand. Their system has three convolutional layers, three subsampling layers in between and one output layer [Al +19]. Also here, the 10-fold-cross-validation is used, hence 10 subjects from each database are selected for training. Additionally, SVM, KNN, Naive Bayes, RF are compared to evaluate whether the proposed system has the best performance. As a result, CNN indeed perform the best, followed by KNN and then RF [Al +19]. For data pre-processing, *"raw data of EDA are scaled such that the distribution is centered around 0, with a standard deviation of 1"*, then data are normalized and labeled with valence and arousal [Al +19].

Likewise, Nasrat et al. develop CNN, and for accuracy improvement they combine the network with SVM, whilst the final output layers of the CNN are SVM's input [Nas+21]. After the HR signal's time series are normalized and labeled, they are transformed into five different images, which act as input data for the CNN. These are explained in detail in [Nas+21], namely: Spectrogram (STFT), Scalogram (WT), Wigner Ville distribution images, Gramian angular fields (GAF) and Markov transition fields (MTF) images. As a result, MTF had the best accuracy and STFT the lowest. But all images are used for the network. Thereafter, the probabilistic outputs of CNN train the SVM, used as arrays, to classify them into binary valence labels [Nas+21]. Here, the SVM utilizes "a fast linear solver as the kernel function of the separating hyperplane" [Nas+21]. In total, the network consists of three convolutional layers, two fully connected layers (with a dropout layer in between) and the output layer. Finally, they apply five-fold-cross-validation for training and testing [Nas+21].

Lastly, Goshvarpour and Goshvarpour try a new approach with PNN. They assume that physiological reactions to emotions do not necessarily have a typical pattern, rather be

more chaotic. It is claimed that the parameters are not random, since they depend on the individual and the emotion [GG19]. This approach categorizes data into two levels of arousal and two of valence, whilst the normalized non-linear data are given as input. Then the first layer computes the similarity between the feature and the training vector [GG19]. In the second layer, the contributions for each valence and arousal categories are summed up. Later, a probability vector is attained and a transfer function selects the maximum probability. Finally, the network produces two classes (0 and 1). This study extracts three non-linear based features of the signals (ApEn, LE and Poincare) [GG19]. They describe that the first ones presents *"global information about the signal's trajectory in the reconstructed state space, while Poincare's indices characterize the trajectories shape in detail"* [GG19]. Finally, the PNN is trained by adjusting the sigma value. Here, different parameters are tested and the best ends up to be 0.1 for sigma (for both Subject/User-dependent and Subject/User-independent) [GG19]. The features and methods are further explained in [GG19].

To sum up, most studies test various classification methods and choose the one with best accuracy. We cannot point out the best method, since the results of the comparisons show that the classification method's output depends on the given data-set and number of classes. Eventually, the chosen emotion model could affect them, too. Due to the tests and comparisons of Ayata et al., it is obvious that the window size duration and number of extracted features influence the effectivity of machine learning algorithms. Hence, the chosen methods of most studies are reliable, because they are used after evaluating more models. It is also noticed that Ayata et al. and Zhao et al. explore different feature-sets and verify that more data does not mean more relevant information, whereas too less data is also risky, since relevant information could be missed. Especially, some features extract the arousal and some the valence level better (as seen in the table 21). Eventually, the feature sets also depend on the used sensors [AYK18]. Therefore, it would be significant to know the influences of feature selections in the other studies, but none have explored the relevance of features as detailed as the two works mentioned above. Furthermore, machine learning methods like SVM are the most applied ones, but there are various recent approaches with neural networks, too. On the one side, SVM are easy to use, efficient in terms of memory and the kernel can be linear or non-linear. They are effective for higher dimensions, however not very suitable for large data-sets, since the training time can become very long. On the other side, neural networks including Deep Learning (DL) methods are very powerful, since they do not need extracted features and one does not have to worry if too little or much input vectors are given. However,

they should be openly revealed by the authors to be more understandable and can be used or improved further. They often act like a black box, hence, the reasons for the outcome are not easy to follow. Additionally, they cannot be visualized to understand the decision-process. If inputs with high noises, wrong measurements and outliers are used, it can affect the output enormously, and when having only few subjects they can tend to overfit. (Hence, Wang et al.'s study shows that SVM suits better for their data-set than neural networks [Zha+18b].) However, neural networks can provide better results than statistical methods, since all relevant features are extracted and connected through the whole system and they are suitable for large data-sets and many data dimensions. Additionally, with the saying of the researchers, it can be noticed that multiple methods can increase stability [AYK16; Al +19]. Lastly, ten- fold-cross-validation is applied the most.

4.5. Results & Findings

Bulagang et al. reach accuracies of 46.7%, 42.9%, 43.3% for their tested three classifiers: SVM, KNN, RF, respectively in an Subject/User-independent approach [BMT21c]. However, an accuracy up to 100% can be achieved with Subject/User-dependent models. We notice here, that the lowest dependent results are 45.4% with SVM and these are near the overall accuracies of the independent model before. They observe that with only HR data, in which three features are extracted, negative valence is more accurate (LAHV with 40%, HANV with 96%) [BMT21c]. In their next study once again the subject-independent approach is challenged and an accuracy of 66% is achieved [BMT21a]. As a result, despite using the same setup and having less participants, the accuracy is much higher when combining two sensors, although only raw data is used. Notably, single data reveals less information regarding the emotion response. Here, it would be interesting to see if the accuracy increases through feature extraction, which was defined as a key factor for better results in the subsection above. It also must be revealed that this study is a preliminary work and a future study may be more improved and accurate. Nevertheless, we wanted to include it in our comparison to highlight the importance of feature extraction and of the EDA. Moreover, Bulagang et al. also explore a model only considering subject-dependent accuracy with 67,4% [BMT21b]. This study claims that the IBI is effective for valence detection, but it cannot be said if the accuracy is better than with only measuring the HR, since here, information about user-independent results is not revealed and in the other study an average accuracy of subject-dependent is not computed. Nonetheless, their results are not satisfying compared to their study combining SC and HR, which

has almost the same results for user-independent measurements, despite the assumption that SC is one-dimensional and only sensitive to arousal. So, alone with the results of this study, no assertion can be made about the relevance of the IBI. Despite having more participants none improvement can be seen. Additionally, possibly none feature is extracted in this study, hence the importance of the input data is once again noticed [BMT21b].

In Ragot et al's research accuracies of 66% for valence and 70% for arousal based on subject-independent approach are obtained fusing both sensor's data [Rag+17]. (However, not all data of participants can be used as inputs, because 2.3% of responses were missing.) This study proves that wearables are as accurate as laboratory sensors, inducing that their parameters are clinically reliable [Rag+17]. Furthermore, high correlations between cardiac activity and middle correlation between EDA and emotions are observed [Rag+17]. Ragot at el. have better accuracy as Bulagang et al. with the fusion of SC and HR [BMT21a]. Here, a direct comparison is possible since both use user-independent models, the same sensors and SVM. Bulagang et al. use less subjects and a different stimuli. And eventually, Ragot et al. classify 9 classes of emotions, which indeed should have less accuracy, since more classes are considered as defined in section 2. It is obvious that in Ragot et al. better results are obtained, regardless of using pictures as stimuli, which were seen skeptical before. Therefore, it is surprising that the VR environment could not induce higher results, despite seeming very promising. It can be further claimed again, that feature extraction is of high relevance and only raw data does not feed the classification algorithm enough. Bulagang et al. have poor data preparing, thus, we suggest that the VR method needs to be further investigated with better methods.

In addition, Ayata et al. achieve best accuracy with the feature-set 22 with 72.06% and 71.05% for arousal and valence, respectively, fusing both sensor's data [AYK18]. It is not uncovered if the result is subject-dependent or not, hence more comparison with other results and an evaluation of their methods is not directly possible. They try to develop a system for individual music recommendation, inducing that a dependent approach is sensible. Comparing single and multiple sensor results, it is observed that two are more reliable and increase the accuracy. Thereby, for the GSR it was 71.53% and 71.04% for arousal and valence, respectively and for PPG 70.92% and 70.76% for arousal and valence, respectively before [AYK18]. Despite not changing much in GSR, there is an increase in arousal with both sensors. Using both sensors' data, the most suitable feature-set has changed, which could mean that the extracted features are very relevant for the outcome. Zhao et al. have an an overall subject-dependent (cross-subject) accuracy of 75.56%,

and we want to mention that they calculate accuracy and precision for every class [Zha+18b]. Overall accuracy for arousal and valence is 78.89% and 75.56%, respectively [Zha+18b]. They comment that *"the performance of the model across participants and subject-independent method is worse than that for a single participant model"* [Zha+18b]. However, the meaning after this claim is not clear, since a subject-independent accuracy is not presented and is not the same as a cross-subject accuracy, which in the end is only the average performance of all subjects' individual results. Nonetheless, their results regarding feature selection shows that the accuracy improves with the iterative adding of features. The arousal has its best accuracy with 14 and the valence with 18 features, while the four-class model has its peak with 16 features [Zha+18b]. Furthermore, since most of the features are extracted from the HRV in the arousal space, they assume that the HRV is more sensitive to arousal. Surprisingly, for valence EDA-related features are more effective [Zha+18b]. Here, the assumptions that the EDA sensor is a one-dimensional sensor to arousal are turned down. Their results prove that multiple sensors are more effective. Additionally, also here it is pointed out that features need to be analyzed thoroughly since like in this study, the number varies regarding arousal, valence and 4-class approaches.

Alskafi et al. who have applied the 9-class model, achieve results of 80,9% (fine KNN) for arousal and 81,1% (weighted KNN) for valence, which ends up being 75,8% on average with the fine KNN method [AKJ21]. We notice that this result is much higher than the one of Ragot et al.. These results could support their method of evoking emotions naturally through debates and conversations. Likewise in the study of Zhao et al. it is observed that more sensor data leads to better results. However, it is not announced if the result is subject-dependent or not, which does not allow direct comparison with the other works. Even for subject-dependent results, their system have higher accuracy than Bulagang et al., Ayata et al. and Zhao et al, despite approaching 9 classes.

Furthermore, Ayata et al. prove that EMD before feature extraction will improve the overall accuracy and that time domain features are more effective [AYK16]. With the EMD, the accuracy increase from 71.93% to 85.07% for arousal, 71.04% to 82.81% for valence. But within the same study it is claimed that the accuracy increase from 71.53% to 81.81% for arousal and 71.04% to 89.29% for valence. [AYK16]. The results are not extremely reliable and it can only be claimed that at least an accuracy of 81% for arousal and 82% for valence can possibly be achieved. Additionally, it is unknown if the model is subject-independent or not, and it also cannot be analyzed which dimension is more accurately classified due to the different presented numbers. Only the increase with the

use of the EMD can be pointed out here.

Machot et al. achieve a total accuracy of 81% for the MAHNOB and 85% for the DEAP data-set for subject-dependent and 78% and 82% for subject-independent, respectively [Al +19]. Since the subject independent results are pretty high even for the 4-class model, it can be supposed that CNN can indeed analyze emotional patterns in sensor data accurately. Eventually, accuracy, precision, recall and F-measure are calculated in this study. The DEAP data-set results in 0.85 in all user-dependent metrics and with independent: 0.82, 0.83, 0.82, 0.83, respectively, when using CNN [Al +19]. Hence, CNN performs best among all tested classifiers in every metric. Furthermore, MAHNOB[8] and DEAP are compared in this study [Al +19]. Comparing the results, DEAP achieve higher accuracy with every metric, inducing that their subject composition (gender, age) and stimuli corpus must be better, since for both the same classification methods are used. Thus, this work shows that CNN are indeed promising.

Moreover, Alskafi et al. attain 87.2% (fine KNN) accuracy for arousal and 89.5% (finegaussian SVM) for valence, which results to be 87.2% on average with the fine KNN method [AKJ21]. They state that 2-class models for separate arousal and valence (4-quadrant) performs better than more classes, nonetheless, the 3 class one shows *"higher balance between classes in terms of correctly classified instances. Same as observed between joined models"* [AKJ21]. Furthermore, despite not knowing if the results are subject-dependent or not and despite using way less subjects, their accuracy is better than with the CNN method of Al Machot et al.. The main difference between both studies are the available sensor-data, the classification method and the stimuli inducement, which implies that measuring multiple sensors increases the accuracy and natural setups could be more reliable than videos. Eventually, it is noticeable that the accuracy of valence is higher than the arousal.

Goshvarpour and Goshvarpour obtain maximum accuracy of 88.57% and 86.8% for arousal and valence in subject independent mode, respectively [GG19]. Generally, results of the arousal dimension are the highest as in every other study until now, regardless of being user-independent. They demonstrate that emotion recognition is highly dependent on the participant and it is easier to detect arousal [GG19]. For both sensors, highest accuracy of 100% can be achieved for the most participants in arousal space and less in valence. Furthermore, the most irregularity is observed during the high valence state and the least during low valence [GG19]. Consequently, signals' irregularity is supposed to be

[8] open databse with 30 participant (17 men and 13 women) between 19 and 40 years old without any diseases. Here, 20 video clips were presented and the emotion was reported via SAM (arousal and valence from 1 to 9) [Al +19]

influenced by the valence dimension [GG19]. The results also show that the fusion has more potential for emotion recognition than using each signal separately. Another key point is that emotions indeed have irregular patterns, which can be analyzed better with PNN [GG19]. Hence, this approach is very promising and should be further investigated with more sensor-data and more classes. Additionally, it is observed that the PPG obtain better classification results compared to the GSR. Finally, in total, better accuracy is achieved for the arousal space.

Lastly, Nasrat et al. attain an accuracy of more than 91% with the classification-combination method, showing an improvement of the binary classification of emotional valence by more than 19% compared to using CNNs on their own [Nas+21]. Furthermore, it is proven that the combination of CNN with SVM can achieve pretty high and satisfying results, which can be used as a basis for further research. After emerging all image transformations very high results of 97.7% are reported, in spite of the fact that the lowest accuracy and F1-Score are 64.37% and 76.95% with STFT, and the highest are 72.32% and 83.34% with MTF, respectively [Nas+21]. It is not revealed if the system is dependent on the subject or not, however, the results imply that HR can be very accurate in measuring valence. In addition, it is seen that self-reports at random times when emotions naturally arise, are more accurate than controlled setups. The study have better results than other subject-dependent or independent models and also stands out beside other studies' valence accuracy. However, since only two classes are classified a comparison is not that effective. Nevertheless, the results are very promising and also reveal that a model can be better trained with a large data-set and more subjects [Nas+21]. It would be significant to expand the study of Al Machot et al. with the advantages of Nasrat et al. to allow better comparison. Eventually, multiple CNN could increase the accuracy even more and also more image transformations could be of advantage in Machot et al's work. In both studies, it would be interesting to see how the accuracy changes with including more sensor data. However, we need an extension of dimensions and classes in the work of Nasrat et al. [Nas+21] to be able to compare and to imply the better methods.

5. Discussion & Future Works

5.1. Reflection & Discussion

Our comparison and analysis result with promising outcomes for the viewed purpose of estimating emotions via wearables. However, not all studies are individually satisfying in every aspect. Hence, new studies focusing on each study's advantages and fusing them in one system would be beneficial to attain maximum accuracy.

First of all, we have noticed in which aspect the emotion evocation methods influence the result. Here, natural stimuli act better, assuming that the subject is not stressed and forced, so that the emotion is induced without putting pressure. In addition, participants have reported the emotion more accurately as seen in the works of Alskafi et al. and Nasrat et al. [AKJ21; Nas+21]. The VR environment seems promising, considering that it brings visual and audio stimuli together in a virtual world experience, but this could not be proven with results by the analyzed studies of Bulagang et al. [BMT21c; BMT21a; BMT21b]. More researches into these stimuli are welcome for a better evaluation, since we claim that the problems in those studies mostly lied in the feature extraction methods, not in the stimuli itself. Furthermore, it would be meaningful to know if self-chosen video clips are more accurate, due to the high dependency of emotions on the individual's experiences. But direct comparison was not seen as sensible with the presented works, because in Zhao et al.'s study [Zha+18b] only 15 participants are present and twice as much in the other studies. Hence, despite approaching similar methods as in the studies of Ayata et al., the better stimuli cannot be pointed out clearly. Zhao et al. indeed achieve better accuracy than in Ayata et al. with the fused sensor data [AYK18]. However, in the study of Ayata et al., in which only GSR data is estimated [AYK16], it is less accurate, possibly highlighting the impact of applying methods like EMD. It is not directly possible to state out which stimuli or number of participants is the best, since even when using the same method, it is possible that one study has a better selection of videos/pictures/music/prepared discussion themes,[...] and we do not have a detailed insight into every study, which restricts the comparison. Here, for example one song could induce sadness for one person, if it was listened to with someone who is not around anymore, and the other person might be very happy, because it was their wedding song. Moreover, we think that the number and combination of participants could play a great role, but it seems that not every study has paid much attention to the selection of their subjects. It is especially seen that small groups are preferred and most of them are in the same age group, since they acquire college students. We further notice that some

studies offer money for subjects to motivate/encourage them like in the study of Ragot et al. [Rag+17]. Of course, even with only 10 to 30 subjects, much data can be acquired, especially when measuring multiple sensor-data and for a longer time. However, this also means that a big data source is only from one person, which is not effective for training a subject-independent system. Regarding this matter, Zhao et al. compare two different databases with the same methods (more information about the databases is given in section 4) resulting with varying accuracy, inducing that the chosen videos or the selection of subjects are effective, since in both the same stimuli were used and the only difference lies in the subjects. Here, individual preferences, the culture and the education could play an important role. Some may enjoy history and war movies, while others could become emotional or angry. The same can be seen in comedy movies: some may enjoy and laugh, others might not really understand the jokes or even feel offended. Furthermore, as described in section 2, we assume that some subjects could rate their own emotions better than others. Knowing this problem, Deng, Chang, Yang, Huo and Zhou focus on heartbeat differences between men and women feeling the same emotion [Den+16]. As a conclusion, emotions like anger and joy are felt more intense by men. But, women usually tend to report stronger responses, especially related to sadness. The heartbeat shows that men experience the most emotions more intense, but report less aroused states. Consequently, the researchers suppose that social stereotypes force participants to think about how they need to feel or to hide/exaggerate their emotions [Den+16]. The results of this study can be seen in figure 22. Leading to a further relevant point that, again, self-reports are not very accurate and trustworthy. Thus, depending highly on the individual. Furthermore, rating emotions on different dimensions is a pretty complex task, resulting that people may not be capable to exactly know and report the felt affective state, especially if multiple emotions are involved.

Another research group highlights that the felt emotion varies with the participant's age by letting them report their emotions after watching video clips [Fer+18]. Older adults in this study experience negative emotions stronger than younger adults. Most differences are noticed in clips inducing disgust and fear. Furthermore, higher arousal is reported by elder people for sadness, anger and tenderness clips and by young adults for amusement clips. Another important point is that older adults can recover more easily from the emotions. Therefore, from this study one can learn the importance of controlling the baseline and setting rest periods of sufficient duration [Fer+18]. We have also seen in section 3 that diseases can influence physiological parameters, as why the outcome of the emotion can differ from person to person. Here, mostly elderly are likely

to have health issues. As explained before, it is relevant to choose subjects without health issues and past diseases. Moreover, it seems sensible asking them to obviate specific food/drugs/beverage as in Zhao et al.'s work [Zha+18b]. We think that it is a naive and not reliable approach to build a classification system with only 10-20 subjects, especially, if one gender is more presented. Consequently, we see that the selection of participants is an important task, but in our comparison it is noticed that most studies only use students as participants. The same observance is reported by Schmidt et al. who say that it is more convenient to acquire research stuff or students [Sch+19]. Hence, the groups are unevenly distributed. As why, we assume that more studies involving elderly and more age groups are necessary, particularly, for health-related applications, since parameters can act differently with the age.

Another key aspect we want to point out is related to the emotion model. When comparing two-dimensional models of different studies, it is noticed that the definition of emotions in vector space differ. It is hard to impossible to tell which emotion exactly is composed of which arousal and valence level with current studies. This realization demonstrates again that self-reports are very subjective and it is critical to build an overall user-independent system with few participants' data. Therefore, it is proposed that at least a new axis could differentiate between similar emotions in a new dimension as explained in section 2 and illustrated in figure 11.

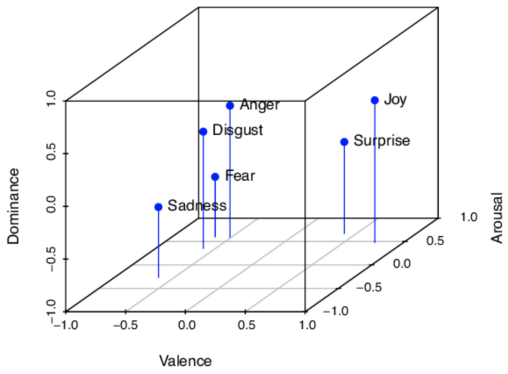


Figure 11: Effect of Expanding Dimensions[9]

Furthermore, it is relevant to mention that not all sensors can measure valence and arousal spaces with the same sensibility. But, due to different accuracies of studies, it cannot be outlined which dimension is measured best by which sensor. Mostly, it is observed that the detection of valence is still not as accurate as the arousal as seen in

[9] https://www.researchgate.net/figure/The-three-dimensional-space-spanned-by-the-VAD-dimensions-For-a-more-intuitive_fig1_313056245, 04.07.2022

section 4. Dzedzickis et al. present that clinical-level-sensors like the EEG or EMG/EOG can evaluate valence better [DKB20]. Whereas it is said that EDA is sensible to arousal, HR is two-dimensional and can detect arousal and valence [Sch+19]. However, for skin temperature the assumptions are contradictory as seen in section 3.3. Such assumptions are hard to prove, since the results of sensor analysis depend highly on the extracted features and evoked stimuli. We want to mention that studies have shown, as pointed out before, that algorithms and the extraction of features improve the quality of valence, as in the work of Nasrat et al. [Nas+21] and Ayata et al. with using the EMD method [AYK16]. In addition, wearable sensors have almost the same accuracies as clinical sensors, which is proven by Ragot et al.'s work, in which laboratory sensors (Biopac MP150) are compared with wearables (E4) for cardiac and electrodermal activity [Rag+17]. Furthermore, Wang et al. have improved the "EmotionSense" [Zha+18b] by fusing data of ACC (acceleration) to measure the activity, which shows results up to an accuracy of 74.3% [Wan+20]. Hence, it is advantageous to use every available sensor and fuse the data because the information of all can lead to better accuracies in valence and arousal, as analyzed in section 4. However, we have also seen in chapter 3 that different sensors need different window sizes and not all stimuli methods need to results in the same way with every sensor. SKT for example was said not to be suited for short stimuli evocation methods like pictures.

In data science is is a skeptical approach to feed the system with every possible data without knowing it's impact on the result, since irrelevant data can lead to very insignificant findings. However, in our viewed purpose it is already proven that viewed sensors hold information on emotional states as shown in section 4.

Coming to the classification methods and extracted features, we want to convey that filtering, normalizing and a thorough exploration of feature-sets are of high relevance. The optical window size duration needs to be individually experimented, which varies from the considered emotion. The studies have illustrated that the classification methods depend on the data-set and hence, the best methods needs to be explored. Consequently, we are not able to highlight the best proceedings as originally intended, but for large data-sets neural networks combined with SVM seemed the most promising so far, like in Nasrat et al.'s system [Nas+21]. At the beginning, we have approached neural networks more critically since they can lead to better accuracies, but also are more out of control and can be highly influenced by outliers. They are more complex and require more time for training and preparation. Furthermore, with other methods the origin of the decision can be better understood as presented by Rattanyu et al. [ROM10]. Alternatively,

Machot et al. report that limits of supervised machine learning systems, in which feature extraction is very important and having one more feature or less has a great impact on the result, can be overcome with DL methods. They claim that researches focus more on basic features and relevant information is sometimes not extracted [Al +19]. Concluding, again, that we cannot highlight the best solution, but only can suggest to compare more methods within the study and chose the best for the purposed data-set. Finally, only using raw data is not considered a good approach.

Our most critical point regarding the studies is that not everyone reveals the same amount of information, resulting in a suspicious impression. Hence, we as the readers, have to make our own assumptions. Sometimes the number of gender is not revealed, the number of features and most importantly, if a subject dependent or independent approach was aimed. Here, it is important to acknowledge that subject independent does not mean an average of all subject dependents or an average accuracy when the signals are normalized. Rather, they need to be normalized and trained and tested on separate groups [Al +19]. In addition, Machot et al. support the view that it is far from practical reality to collect data each time for each subject. Thus, they state out that independent systems should be targeted [Al +19]. Another point is that not every researcher in our comparison has made use of metrics and frequently only has calculated the accuracy. Sometimes a confusion matrix is revealed, but metrics like recall, precision and F-measure are rarely used. Nonetheless, as Alskafi et al. mention, in emotion recognition, it is significant to know the number of correctly classified emotions, and in our interest the declassified are important as well if we want to build phobia therapies. Therefore, besides accuracy the results of other metrics are important to know [AKJ21].

Lastly, nobody has approached the privacy aspect of such applications. As mentioned in section 1, the lifestyle has changed in the last decade, and most humans are conscious about the rights of their own data. We know that we are tracked through various websites and applications, and that our data is almost never secure. We even need to accept cookies and allow websites to make use of our data. However, people tend to act different about their health information. It is observed that most want to protect their data and were pretty long skeptical to the health care sector's digitization. Especially, diseases regarding emotional states are sensible data and emotional information could be of interest for advertising, hence, need to be secure. Therefore, the system should be aimed to be as secure as possible. But since data are captured through wearables and often connected to apps/ the web/ the cloud, a secure protection cannot be guaranteed, which makes the use and acceptance of these systems pretty complicated in health care.

Lastly, we suggest anonymous questionnaires to be aware of the actual perspective of the different age groups.

5.2. Future Works

For further work, we suggest taking advantage of all possible sensor data. We want to point out that according to some reviews like the one of Dzedzickis et al. and the study of Egger et al., the PPG is capable of calculating the RSP from the BVP, containing information about the breathing behavior and thus, can possibly distinguish between negative and positive emotions and their intensity [DKB20; ELH19]. Therefore, we suggest correlation studies between the PPG extracted RSP and conventional methods. We have found studies focusing on a pulse-oximeter, but no reliable statements could be extracted since they were contradictory [Hak+18; Wen+14]. Hence, more research is advantageous, to see if SpO2 contains emotional data or not. Moreover, blood pressure could hold information about the affective state, which can possibly also be measured by wearables. Summing up, features as IBI, RSP, blood-pressure and SpO2 should be further researched in relation to emotion recognition. Eventually, acceleration can be measured as in the "EmotionSense" of Wang et al. [Wan+20]. We notice that gender and age differences should be researched more and studies need to focus on their containment of participants. As proposed before in section 2, dimension extensions should be investigated, since the two-dimensional approach has it's limitations. Lastly, we want to mention that the E4 (which is used in most studies including seven of our compared works) is very pricey and acts as a medical device. Therefore, we suggest that more studies focus on usual recent smart/fitness-watches, which are used in everyday life by more people. However, if too many different wearables are used and tested by different studies, comparison between works becomes even harder. Maybe watches with the most accurate sensors can be chosen as basic wearable devices, until better ones are released.

6. Conclusion

This work has shown that emotion recognition via wearables would bring new possibilities to health care and new diagnosis/support applications could be designed. The analysis reveals that current trends are promising (best accuracy in the 4-class model with 87.2% (fineKNN) and 88.57% and 86.8% for arousal and valence, respectively with (PNN) and through different algorithms like the EMD and specific feature extractions, accurate valence and arousal detection is possible. Wearable sensors like the EMD, PPG, the GSR/ EDA and skin temperature are very reliable and almost of clinical accuracy. Furthermore, it is seen that more sensors result in better accuracy, since each sensor has their own advantages regarding the spaces. Some give better results in one space, some in the other, or in both dimensions, thus, the fusion increases overall accuracy. Regarding the classification methods, neural networks have shown the most promising accuracies (CNN, PNN), especially when emerging with SVM. Another significant point is that more studies need to approach a subject-independent-system and we think that the influences of the subject group's composition should be more researched. In addition, it would be interesting to see if gender-related systems are more accurate. Moreover, RSP, SpO2 and ACC should be further investigated, since not much information could be found. We also suggest that the dimensional-model needs to be expanded and tested for accuracy with additional spaces like dominance or time. We have seen that the dimensional-model is much more flexible and reliable than the discrete one, since the emotions are not fixed discretely. Lastly, it is noted that natural environments can induce stronger emotions or the emotion report is more accurate, and it is more trustworthy to monitor the person for a longer time.

References

[AKJ21] Feryal A. Alskafi, Ahsan H. Khandoker, and Herbert F. Jelinek. "A Comparative Study of Arousal and Valence Dimensional Variations for Emotion Recognition Using Peripheral Physiological Signals Acquired from Wearable Sensors". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine &amp Biology Society (EMBC)*. IEEE, Nov. 2021. DOI: 10.1109/embc46164.2021.9630759. URL: https://doi.org/10.1109/embc46164.2021.9630759.

[Al +19] Fadi Al Machot et al. "A Deep-Learning Model for Subject-Independent Human Emotion Recognition Using Electrodermal Activity Sensors". In: *Sensors* 19.7 (2019). ISSN: 1424-8220. DOI: 10.3390/s19071659. URL: https://www.mdpi.com/1424-8220/19/7/1659.

[AYK16] Değer Ayata, Yusuf Yaslan, and Mustafa Kamaşak. "Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches". In: *2016 Medical Technologies National Congress (TIPTEKNO)*. 2016, pp. 1–4. DOI: 10.1109/TIPTEKNO.2016.7863130.

[AYK18] Deger Ayata, Yusuf Yaslan, and Mustafa E. Kamasak. "Emotion Based Music Recommendation System Using Wearable Physiological Sensors". In: *IEEE Transactions on Consumer Electronics* 64.2 (2018), pp. 196–203. DOI: 10.1109/TCE.2018.2844736.

[Băl+19] Oana Bălan et al. "Fear Level Classification Based on Emotional Dimensions and Machine Learning Techniques". In: *Sensors* 19.7 (2019). ISSN: 1424-8220. DOI: 10.3390/s19071738. URL: https://www.mdpi.com/1424-8220/19/7/1738.

[BMT21a] A Bulagang, James Mountstephens, and Jenna Teo. "Exploring Standalone Electrodermography for Multiclass VR Emotion Prediction using KNN". In: *Journal of Physics: Conference Series* 1878 (May 2021), p. 012061. DOI: 10.1088/1742-6596/1878/1/012061.

[BMT21b] Aaron Bulagang, James Mountstephens, and Jason Teo. "A Novel Approach for Emotion Classification in Virtual Reality using Heart Rate (HR) and Inter-beat Interval (IBI)". In: Nov. 2021, pp. 247–252. DOI: 10.1109/ICOCO53166.2021.9673506.

[BMT21c] Aaron Bulagang, James Mountstephens, and Jason Teo. "Multiclass emotion prediction using heart rate and virtual reality stimuli". In: *Journal of Big Data* 8 (Jan. 2021). DOI: 10.1186/s40537-020-00401-x.

[Den+16] Yaling Deng et al. "Gender Differences in Emotional Response: Inconsistency between Experience and Expressivity". In: *PLoS ONE* 11 (2016).

[Dis+19] Theekshana Dissanayake et al. "An Ensemble Learning Approach for Electrocardiogram Sensor Based Human Emotion Recognition". In: *Sensors* 19 (Oct. 2019), p. 4495. DOI: 10.3390/s19204495.

[DKB20] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. "Human Emotion Recognition: Review of Sensors and Methods". In: *Sensors* 20.3 (2020). ISSN: 1424-8220. DOI: 10.3390/s20030592. URL: https://www.mdpi.com/1424-8220/20/3/592.

[DS21] Usha Desai and Akshaya D. Shetty. "Electrodermal Activity (EDA) for Treatment of Neurological and Psychiatric Disorder Patients: A Review". In: *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. Vol. 1. 2021, pp. 1424–1430. DOI: 10.1109/ICACCS51430.2021.9441808.

[ELH19] Maria Egger, Matthias Ley, and Sten Hanke. "Emotion Recognition from Physiological Signal Analysis: A Review". In: *Electronic Notes in Theoretical Computer Science* 343 (2019). The proceedings of AmI, the 2018 European Conference on Ambient Intelligence., pp. 35–55. ISSN: 1571-0661. DOI: https://doi.org/10.1016/j.entcs.2019.04.009. URL: https://www.sciencedirect.com/science/article/pii/S157106611930009X.

[Fer+18] Luz Fernández-Aguilar et al. "Emotional Differences in Young and Older Adults: Films as Mood Induction Procedure". In: *Frontiers in Psychology* 9 (2018). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01110. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2018.01110.

[GG19] Atefeh Goshvarpour and Ateke Goshvarpour. "The potential of photoplethysmogram and galvanic skin response in emotion recognition using nonlinear features". en. In: *Australas. Phys. Eng. Sci. Med.* 43.1 (Nov. 2019), pp. 119–134.

[Hak+18] Lutfi Hakim et al. “Emotion Recognition in Elderly Based on SpO2 and Pulse Rate Signals Using Support Vector Machine”. In: *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. 2018, pp. 474–479. DOI: 10.1109/ICIS.2018.8466489.

[HS18] Terence K.L. Hui and R. Simon Sherratt. “Coverage of Emotion Recognition for Common Wearable Biosensors”. In: *Biosensors* 8.2 (2018). ISSN: 2079-6374. DOI: 10.3390/bios8020030. URL: https://www.mdpi.com/2079-6374/8/2/30.

[Kre10] Sylvia Kreibig. “Autonomic Nervous System Activity in Emotion: A Review”. In: *Biological psychology* 84 (Apr. 2010), pp. 394–421. DOI: 10.1016/j.biopsycho.2010.03.010.

[MGK18] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. “Detection of Negative Emotions and High-Arousal Negative-Valence States on the Move”. In: *2018 Advances in Wireless and Optical Communications (RTUWO)*. 2018, pp. 61–65. DOI: 10.1109/RTUWO.2018.8587888.

[Nas+21] Sara A. Nasrat et al. “Emotion Recognition in the Wild from Long-term Heart Rate Recording using Wearable Sensor and Deep Learning Ensemble Classification”. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021, pp. 1676–1678. DOI: 10.1109/BIBM52615.2021.9669553.

[PRP05] JONATHAN POSNER, JAMES A. RUSSELL, and BRADLEY S. PETERSON. “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology”. In: *Development and Psychopathology* 17.3 (2005), pp. 715–734. DOI: 10.1017/S0954579405050340.

[Rag+17] Martin Ragot et al. “Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors”. In: June 2017. ISBN: 978-3-319-60638-5. DOI: 10.1007/978-3-319-60639-2_2.

[ROM10] Kanlaya Rattanyu, Michiko Ohkura, and Makoto Mizukawa. “Emotion monitoring from physiological signals for service robots in the living space”. In: *ICCAS 2010*. 2010, pp. 580–583. DOI: 10.1109/ICCAS.2010.5669914.

[Sag+20] Stanisław Saganowski et al. “Emotion Recognition Using Wearables: A Systematic Literature Review - Work-in-progress”. In: *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2020, pp. 1–6. DOI: 10.1109/PerComWorkshops48775.2020.9156096.

[Sax+20] Piyush Saxena et al. “Reconstructing Compound Affective States using Physiological Sensor Data”. In: *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. 2020, pp. 1241–1249. DOI: 10.1109/COMPSAC48688.2020.00-86.

[Sch+19] Philip Schmidt et al. “Wearable-Based Affect Recognition—A Review”. In: *Sensors* 19.19 (2019). ISSN: 1424-8220. DOI: 10.3390/s19194079. URL: https://www.mdpi.com/1424-8220/19/19/4079.

[Shu+20] Lin Shu et al. “Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet”. In: *Sensors* 20.3 (2020). ISSN: 1424-8220. DOI: 10.3390/s20030718. URL: https://www.mdpi.com/1424-8220/20/3/718.

[Tak+21] Reika Takeshita et al. “Emotion Recognition from Heart Rate Variability Data of Smartwatch While Watching a Video”. In: *2021 Thirteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*. 2021, pp. 1–6. DOI: 10.23919/ICMU50196.2021.9638844.

[Wan+20] Zhu Wang et al. “EmotionSense: An Adaptive Emotion Recognition System Based on Wearable Smart Devices”. In: *ACM Trans. Comput. Healthcare* 1.4 (Sept. 2020). ISSN: 2691-1957. DOI: 10.1145/3384394. URL: https://doi.org/10.1145/3384394.

[Wei99] Mark Weiser. “The Computer for the 21st Century”. In: *SIGMOBILE Mob. Comput. Commun. Rev.* 3.3 (July 1999), pp. 3–11. ISSN: 1559-1662. DOI: 10.1145/329124.329126. URL: https://doi.org/10.1145/329124.329126.

[Wen+14] Wanhui Wen et al. “Emotion Recognition Based on Multi-Variant Correlation of Physiological Signals”. In: *IEEE Transactions on Affective Computing* 5.2 (2014), pp. 126–140. DOI: 10.1109/TAFFC.2014.2327617.

[Zha+18a] Bobo Zhao et al. “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 (Smart-*

World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). 2018, pp. 346–355. DOI: 10.1109/SmartWorld.2018.00091.

[Zha+18b] Bobo Zhao et al. "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, pp. 346–355. DOI: 10.1109/SmartWorld.2018.00091.

Online Sources:

https://my.clevelandclinic.org/health/articles/17064-heart-beat, last accessed 10.06.2022

https://psu.pb.unizin.org/psych425/chapter/744/, Michelle Yarwood, last accessed 10.06.2022

https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal, last accessed 1.07.2022

https://wtcs.pressbooks.pub/pharmacology/chapter/4-2-ans-basics/, last accessed 10.06.2022

https://www.mdpi.com/2079-9292/11/3/496/htm, Stanislaw Saganowski, last accessed 27.07.2022

https://www.researchgate.net/figure/Empatica-E4-wristband-physiological-signal-monitoring-14_fig9_322206805, Dragos Datcu and Leon Rothkrantz, last accessed on 27.07.2022

https://www.researchgate.net/figure/The-three-dimensional-space-spanned-by-the-VAD-dimensions-For-a-more-intuitive_fig1_313056245, Sven Buechel, Johannes Hellrich and Udo Hahn, last accessed 04.07.2022

A. Appendix

PPG Features	Calculations Based on Python (Import Numpy, Pandas and Scipy)
IBI	Peak detection of raw PPG signal and get an array (ppgnn) Interbeat interval (IBI) = ppgnn.interpolate(method = "cubic")
HR	Heart Rate = (60 s × sampling frequency)/peak-to-peak duration HR = IBI.rolling (window, min_periods = 1, centre = True).mean()
SDNN	Standard deviation of IBI SDNN = IBI.rolling (window, min_periods = 1, centre = True).std()
SDSD	Standard deviation of the difference between adjacent ppgnn ppgdiff = pd.DataFrame(np.abs(np.ediff1d(ppgnn))) ppgdiff = ppgdiff.interpolate (method = "cubic") SDSD = ppgdiff.rolling (window, min_periods = 1, centre = True).std()
RMSSD	Root Mean Square of the difference between adjacent ppgnn ppgsqdiff = pd.DataFrame (np.power(np.ediff1d (ppgnn), 2)) ppgsqdiff = ppgsqdiff.interpolate (method = "cubic") RMSSD = np.sqrt (ppgsqdiff.rolling (window, min_periods = 1, centre = True).mean())

Figure 12: List of PPG-Signal's Extracted Features I. [HS18]

PPG Features	Calculations Based on Python (Import Numpy, Pandas and Scipy)
SDNN/RMSSD	Ratio between SDNN and RMSSD SDNN_RMSSD = SDNN / RMSSD
LF	Power Spectral Density (PSD) for low frequency range (0.04 Hz to 0.15 Hz) Y = np.fft.fft (IBI)/window, Y = Y [range(window//2)] LF = np.trapz (np.abs(Y[(freq $\geq$ 0.04) & (freq $\leq$ 0.15)]))
HF	PSD for high frequency range (0.16 Hz to 0.4 Hz) HF = np.trapz(np.abs (Y[(freq $\geq$ 0.15) & (freq $\leq$ 0.4)]))
LF/HF	PSD ratio between LF and HF LHF = LF / HF

Figure 13: List of PPG-Signal's Extracted Features II. [HS18]

EDA Features	Calculations Based on Python (Import Numpy, Pandas and Scipy)
EDA (filtered)	eda = raw EDA signal sampling at 100 ms B, A = signal.butter (2, 0.005, output = "ba") EDAf = signal.filtfilt (B, A, eda)
EDA (mean)	Getting rolling mean of filtered EDA raw signal (EDAf) EDAmean = EDAf.rolling (window, min_periods = 1, centre = True).mean()
EDA (std)	Getting rolling standard deviation of filtered EDA raw signal (EDAf) EDAstd = EDAf.rolling (window, min_periods = 1, centre = True).std()

Figure 14: List of GSR-Signal's Extracted Features I. [HS18]

ID	Designation	GSR-based Feature Description
1	Num_Peaks	Number of peaks for the whole record
2	Cond_mean_all	Mean conductance of the whole record
3	Cond_max	Maximum amplitude of the peaks
4	Cond_min	Minimum amplitude of the peaks
5	Cond_mean	Mean conductance of the peaks
6	Cond_rms	RMS of the peaks conductance
7	Cond_std	Standard Deviation of the peaks conductance
8	Cond_mean_abs	Mean absolute value of the peaks conductance
9	Resist_mean	Mean resistance for the whole record
10	Skewness	Skewness of the peaks distribution
11	Kurtosis	Kurtosis of the peaks distribution
12	d_1	First Degree Difference (FDD)
13	d_2	Second Degree Difference (SDD)
14	d_1_div_std	Ratio of FDD and Standard deviation
15	d_2_div_std	Ratio of SDD and Standard deviation
16	Specpow2_4	Power in the band 0-2.4Hz

Figure 15: List of GSR-Signal's Extracted Features II. [MGK18]

SKT Features	Calculations Based on Python (Import Numpy, Pandas and Scipy)
SKT (filtered)	skt = raw SKT signal sampling at 100 ms B, A = signal.butter (2, 0.005, output = "ba") SKTf = signal.filtfilt (B, A, skt)
SKT (mean)	Getting rolling mean of filtered SKT raw signal (SKTf) SKTmean = SKTf.rolling (window, min_periods = 1, centre = True).mean()
SKT (std)	Getting rolling standard deviation of filtered SKT raw signal (SKTf) SKTstd = SKTf.rolling (window, min_periods = 1, centre = True).std()

Figure 16: List of SKT-Signal's Extracted Features [HS18]

Table 1. Division of valence, arousal and dominance for the two-level fear evaluation modality.

Label	Valence	Arousal	Dominance
No fear (0)	(5; 9]	[1; 5)	[5; 9]
Fear (1)	[1; 5]	[5; 9]	[1; 5)

Table 2. Division of valence, arousal and dominance for the four-level fear evaluation modality.

Label	Valence	Arousal	Dominance
No fear (0)	[7; 9]	[1; 3)	[7; 9]
Low fear (1)	[5; 7)	[3; 5)	[5; 7)
Medium fear (2)	[3; 5)	[5; 7)	[3; 5)
High fear (3)	[1; 3)	[7; 9]	[1; 3)

Figure 17: Different Fear Levels [Băl+19]

Study	Setup	Data	Emotion-class	Method(s)	Features	Results	Model
Ayata et al., 2016	DEAP: open database with 32 subjects, 40 video clips	SC	4 class	EMD+RF	14	81-85% for arousal, 82-89% for valence	N/A
Ragot et al., 2017	19 subjects, 45 pictures (selected randomly)	SC, HR (E4)	9 class	SVM	9	70% for arousal, 66% for valance	independent
Ayata et al., 2018	DEAP: open database with 32 subjects, 40 video clips	SC, BVP	4 class?	RF	22	72.06% for arousal, 71.05% for valence	N/A
Zhao et al., 2018	15 subjects, self-chosen movie clips	SC, BVP, HRV, SKT (E4)	4 class	SFFS+SVM	28	76% in total	dependent
Goshvarpour et al., 2019	DEAP: open database with 32 subjects, 40 video clips	SC, BVP	4 class	PNN	3 feature extraction methods	88.57% for arousal, 86.8% for valence	independent
Al Machot et al., 2019	DEAP: open database with 32 subjects, 40 video clips	SC	4 class	CNN	raw data	85% for dependent, 82% for independent	both
Alskafi et al., 2021	16 debate session with pairs	SC, BVP, HRV, SKT (E4)	4 class	fine KNN	N/A	87.80% in total	N/A
Alskafi et al., 2021	16 debate session with pairs	SC, BVP, HRV, SKT (E4)	9 class	fine KNN	N/A	75.80% in total	N/A
Bulagang et al., 2021	20 subjects, 16 videos in VR environment	HR (E4)	4 class	SVM	3	46.7% in total	independent
Bulagang et al., 2021	10 subjects, 16 videos in VR environment	EDG/SC, HR (E4)	4 class	SVM	raw data	66% in total	independent
Bulagang et al., 2021	24 subjects, 16 videos in VR environment	HR, IBI (E4)	4 class	SVM	N/A	67.4% in total	dependent
Nasrat et al., 2021	80 subjects, self-report in daily life at random times	SC (MB2)	valence	CNN + SVM	5 image transformations	> 91% in total	N/A

Figure 18: Comparison of 12 Studies: Complete table

Feature set	Attributes
(FS-10)	Minimum, maximum, average, standard deviation, variance, skewness, kurtosis, median, zero crossings, mean energy
(FS-14)	Feature 10 set, 3rd, 4th, 5th, 6th moments
(FS-18)	Feature 14 set, mean absolute value, maximum scatter difference root mean square, mean absolute deviation
(FS-22)	Feature 18 set, 1st degree difference, 2nd degree difference 1st degree difference divided by standard deviation, 2nd degree difference divided by standard deviation

(a)

Attribute	Formula	Attribute	Formula
Min	$\min\{X_n\}$	Skewness	$\sum_{n=1}^{N}(X_n - AM)\frac{3}{(N-1)SD^3}$
Max	$\max\{X_n\}$	Kurtosis	$\sum_{n=1}^{N}(X_n - AM)\frac{4}{(N-1)SD^4}$
Arithmetic mean (μ)	$\frac{1}{N}\sum_{n=1}^{N} X_n$	Median	$\frac{(\frac{N}{2})^{\text{th}}+(\frac{N}{2}+1)^{\text{th}}}{2}$ or $(\frac{N+1}{2})^{\text{th}}$
Mean Absolute	$\frac{1}{N}\sum_{n=1}^{N} \lvert X_n \rvert$	Moment (kth order)	$\frac{1}{N}\sum_{n=1}^{N} X_n^k$
Root Mean Square	$\sqrt{\frac{1}{N}\sum_{n=1}^{N} X_n^2}$	First Degree Difference	$\frac{1}{N-1}\sum_{n=1}^{N} \lvert X_{n+1} - X_n \rvert$
Standard Deviation (SD)	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(X_n - AM)^2}$	Second Degree Difference	$\frac{1}{N-2}\sum_{n=1}^{N} \lvert X_{n+2} - X_n \rvert$

(b)

Figure 19: Feature-Set of Ayata et al. [AYK18]

Feature	*Average*	*IG*
P_peak_LF	5.06±2.43	0.113
P_peak_MF	24.37±6.25	0.119
P_peak_HF	7.42±3.64	0.156
P_power_LF	42.44±10.86	0.145
P_power_MF	45.90±9.42	0.264
P_power_HF	35.63±8.65	0.041
H_meanValue	0.76±0.32	0.137
H_SDNN	0.13±0.08	0.235
H_maxValue	0.91±0.15	0.042
H_minValue	0.61±0.24	0.011
H_STDD	2.38±0.39	0.164
H_SD1	4.23±2.27	0.188
H_SD2	8.97±3.65	0.143
H_SD12	0.63±0.22	0.201

(a)

H_NN50	4.12±2.68	0.203
H_power_VLF	5.65±4.32	0.035
H_power_LF	8.36±5.48	0.154
H_power_HF	7.23±4.16	0.176
E_mean	0.54±0.32	0.138
E_mean_SCL	0.48±0.37	0.214
E_std_SCL	0.35±0.58	0.195
E_num_SCR	38.63±13.52	0.187
E_mean_SCR	0.02±0.04	0.114
E_std_SCR	0.01±0.02	0.161
T_maxValue	34.21±3.21	0.067
T_minValue	27.99+4.38	0.058
T_average	32.94±2.64	0.107
T_STD	1.05±0.89	0.203

(b)

Figure 20: Features of Zhao et al. [Zha+18b]

Classification	Best emotion-related features
Arousal	*P_peak_MF, P_peak_HF, P_power_MF, H_meanValue, H_SD1, H_SD2, H_power_LF, E_mean_SCL, T_average*
Valence	*P_peak_LF, P_peak_MF, P_power_MF, H_meanValue, H_SD1,E_std_SCL, E_mean_SCL, E_mean_SCR, E_std_SCR, T_STD*
Four emotions	*P_peak_LF, P_peak_MF, P_peak_HF, P_power_MF,H_meanValue, H_SD1, H_SD2,H_power_LF,E_num_SCR, E_mean_SCL,E_mean_SCR, T_STD, T_average*

Figure 21: Best Features of Zhao et al. [Zha+18b]

	emotional expressivity			emotional experience
	Valence	Arousal	Motivation	Heart rate
anger	-	women>men	-	decline: men>women
amusement	-	women>men	-	decline: men>women
pleasure	-	women>men	-	decline: men>women
horror	women<men	women>men	avoidance: women>men	-
disgust	women<men	women>men	avoidance: women>men	-
sadness	-	women>men	-	-
surprise	-	-	-	-

"-" means no gender difference.

Figure 22: Differences in Gender [Den+16]