Classification of Affective States in the Electroencephalogram

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M.Sc. Dirk Tassilo Hettich
aus Sindelfingen

Tübingen
2016

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	14.10.2016
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Wolfgang Rosenstiel
2. Berichterstatter:	Prof. Dr. hc. mult. Niels Birbaumer

Acknowledgments

First of all, I would like to thank all individuals who have helped, inspired, and/or motivated me during my doctoral study. None of this would have been possible without the help, discussions, and input from and with colleagues.

I thank Prof. Dr. Wolfang Rosenstiel for accepting me as his doctoral student, for providing me the opportunity to conduct research at the Faculty of Science, and for his valuable feedback and milestone guidance during my doctoral project.

I thank Prof. Dr. hc. mult. Niels Birbaumer for giving me the opportunity to work at the Institute of Medical Psychology and Behavioral Neurobiology, Tübingen, in an excellent, engaging, and international research environment.

In retrospect, everything began as a student opportunity to earn a little extra money at a scientific research institution during my B.Sc. in Bioinformatics. Here, I would also like to thank my former supervisor, Dr. Sebastian Halder, from whom I learned a lot regarding scientific practice during my time as a student. For valuable discussion, I thank my colleagues from the Institute, notably Elaina Bolinger. I also thank the research assistants Felix and Stefan for their help during studies and data acquisition. For their support, I thank Dr. Tamara Matuz, Dr. Hong-Viet Ngo, and Prof. Dr. Jan Born.

I thank Dr. Martin Spüler from the research group Neural Interfaces and Brain Signal Decoding of the Wilhelm-Schickard-Institute for Computer Science, Tübingen, whose close supervision and insights on machine learning practices were indispensable for the development of the analysis apparatus developed in the present work. I also thank the group members for valuable research discussions during the Monday meetings.

Most importantly, I thank my wife Daniela for being at my side, as well as my family and friends for keeping me sane in times of need. Last but not least, I thank our dog Raiko for dragging me away from the computer in regular intervals.

Abstract

Brain state classification for communication and control has been well established in the area of brain-computer interfaces (BCIs) over the last decades. BCIs are communication systems in which muscles or neural pathways are not passed for sending messages or commands to the external world. The goal of the present work is to investigate the feasibility of automatic affect recognition in the electroencephalogram (EEG) in different populations with a focus on feature validation and machine learning in order to augment BCIs by the ability to identify and communicate the users' inner affective state.

Currently, affect recognition studies conducted on EEG data are hardly comparable due to variable parameters in study design, machine learning approaches, and performance measures. Class size is identified as a main constraining factor.

The present work introduces a machine learning framework based on common machine learning practices suitable for affect recognition in the EEG. Two in-depth studies on affect induction and classification are presented.

In the first study, an auditory emotion induction paradigm that easily translates to a clinical population is introduced.The paradigm is designed with a focus on maximizing trial size while avoiding habituation. Based on stimulus valence, three affective states are defined (unpleasant, neutral, and pleasant). The paradigm is applied in a healthy and a population of individuals with cerebral palsy. The late positive potential is identified in the healthy population. Significant above chance group classification is achieved using time domain features for unpleasant vs. pleasant conditions.

In the second study, data of an emotion induction paradigm for preverbal infants are investigated. In infant-parent interaction, different emotions are induced in 6-month-old infants. Employing the machine learning framework, cross-participant classification of pleasant vs. neutral conditions is significantly above chance with balanced training data.

Furthermore, the machine learning framework is applied to the publicly available physiological affect dataset DEAP for comparison of results. Based on spectral frequency features, the framework introduced outperforms results published by the authors of DEAP.

The results strengthen the vision of the feasibility of a BCI that is able to identify and communicate the users' affective state.

Zusammenfassung

Gehirn-Computer-Schnittstellen (GCS) sind Kommunikationssysteme, die es Nutzern erlauben Nachrichten oder Befehle an die Umwelt zu senden, ohne dabei neurale Pfade oder Muskeln zu nutzen. Die Klassifikation von Gehirnzuständen im Elektroenzephalogramm (EEG) durch maschinelles Lernen ist über die letzten Jahrzehnte zunehmend verbessert worden. Die vorliegende Arbeit hat zum Ziel zu untersuchen, ob GCS durch die Fähigkeit ergänzt werden können automatisch den affektiven Zustand der Nutzer zu identifizieren und zu kommunizieren. Derzeit sind Studien über Affekterkennung im EEG nur schwer vergleichbar, da verschiedene Parameter wie Studiendesign, maschinelle Lernansätze und Performanzmaße sich stark unterscheiden. Die Klassengröße ist als ein Schlüsselparameter identifiziert. Die vorliegende Arbeit stellt ein Rahmenwerk für Affekterkennung in EEG vor, welches auf gängigen Praktiken im maschinellen Lernen basiert. Es werden zwei Studien zu Affektinduktion und Klassifikation im EEG vorgestellt. Die erste Studie beschreibt ein auditorisches Paradigma zur Emotionsinduktion, welches sich leicht auf eine klinische Population übertragen lässt. Das Paradigma ist so gearbeitet, so dass die Anzahl an verfügbaren Trials maximiert ist und gleichzeitig Habituation zu vermeiden. Basierend auf der Valenz der Stimuli werden drei affektive Zustände definiert (unangenehm, neutral und angenehm). Das Paradigma wird in einer gesunden und einer Population mit Zerebralparese angewendet. Das späte positive Potential ist als Korrelat von Affekt in der gesunden Population identifiziert. Klassifikationsergebnisse zwischen unangenehmen und angenehmen Zuständen sind signifikant über Zufall mit Merkmalen aus der Zeitreihe, wenn man die Gruppe betrachtet. Die zweite Studie untersucht die EEG-daten von präverbalen Kleinkindern von sechs Monaten aufgenommen während sie mit einem Elternteil interagieren. Gemäß dem Rahmenwerk konnte erfolgreich zwischen angenehmen und neutralen Zuständen klassifiziert werden und das in einem cross-subject Design mit balancierten Trainingsdaten. Darüber hinaus wird das Rahmenwerk auf den veröffentlichten DEAP Datensatz mit physiologischen Daten affektiver Zustände angewendet. Basierend auf spektralen Frequenz-domänen Merkmalen zeigt die vorgestellte Methodik höhere Performanz als mit der bereits veröffentlichten. Die Ergebnisse stärken die Vision, dass ein GCS fähig ist die affektiven Zustände von Nutzern zu identifizieren und zu kommunizieren.

Contents

List of Abbreviations

ANOVA	Analysis of variance
ANS	Autonomous nervous system
AUC	Area under the curve
BCI	Brain-computer Interface
BPM	Beats per minute
BVP	Blood volume pressure
CCA	Canonical Correlation Analysis
CLIS	Complete locked-in state
CNS	Central nervous system
CV	Cross-validation
DEAP	Database for Emotion Analysis using Physiological Signals
ECG	Electrocardiography
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
ERD	Event-related desynchronization
ERP	Event-related potential
ERS	Event-related synchronization
FDR	False discovery rate

FFT	Fast Fourier transformation
fMRI	Functional magnetic resonance imaging
fNIRS	Functional near infrared spectroscopy
fNIRS	Near infrared spectroscopy
FP	False positive
GSR	Galvanic skin response
HCI	Human-computer interaction
ICA	Independent Component Analysis
ICA	Independent component analysis
LDA	Linear discriminant analysis
LIS	Locked-in state
LOOE	Leave-one-out estimation
LPP	Late positive potential
MEG	Magnetoencephalography
MLP	Multi-layer perceptron
Nc	Negative central component
OFC	Orbitofrontal cortex
PCA	Principal Component Analysis
PET	Positron emission topography
PFC	Prefrontal cortex
PNS	Peripheral nervous system
QDA	Quadratic discriminant analysis
RBF	Radial basis function
ROC	Receiver operating characteristic

RSP	Respiratory activity
SCP	Slow cortical potential
SNS	Somatic nervous system
SVM	Support vector machine
SWLDA	Stepwise linear discriminant analysis
TN	True negative
TP	True positive

“Human behavior flows from three main sources:
desire, emotion, and knowledge.”

Plato (428 – 328 BCE)

1 Introduction

Human behavior is driven by constant interaction. Desire, emotion, and knowledge emerge while we as individuals interact with the surrounding world, other individuals, or our inner selves. For the interaction between humans, communication is indubitably most vital. During the course of human existence, we developed a plethora of verbal and non-verbal means of communication. In the beginning nonetheless, the fountain of human behavior and interaction was built upon emotions. Evolution provided our ancestors with the abilities to swiftly express, recognise, and evaluate emotions in order to adjust their immediate behavior accordingly. These abilities may seem somewhat hidden in everyday life nowadays, yet they continue to greatly guide our behavior, interaction, and therefore communication.

1.1. Motivation

Affective states (i.e. emotions, feelings, and moods) are key in personal and interpersonal everyday life. Expressing and understanding emotions not only influences cognitive processes and therefore behavior, yet also secures and maintains individual well-being on a basal level. Classic human-computer interaction (HCI), as the interaction between humans and computing systems, lacks affect as a communication channel, to date. The relatively young field of affective computing seeks to also incorporate psychophysiological information about the inner state of an individual into classic HCI. This interdisciplinary endeavour requires knowledge from many domains mainly consisting of computer science, psychology, and neuroscience.

Besides the commonly known possibilities classic HCI offers for healthy individuals to date, the development in providing disabled or paralyzed individuals a communication system has progressed over the last decades. Systems that do not require muscles or neural pathways to send messages or commands to the external world can be described as brain-computer interface (BCI) systems. Clinically, disabled or paralyzed individuals profit from novel therapeutic or rehabilitative measures offered by these systems. In order to communicate with families or caretakers, paralyzed individuals without verbal communication can send messages to a computer screen solely by their brain activity and a BCI system. Furthermore, partially or entirely paralyzed individuals are able to control orthoses or robotic arms with such systems.

BCI systems for communication have first been established during the '80s of the last century [1, 2]. Until now, advancements in their efficiency and reliability have been achieved. Nonetheless, communication by BCI systems can still be categorised as classic HCI. Research shows that affective states and communication are vital in personal and interpersonal life especially during development and in a population where communication is impaired. Thus, the development of BCI systems that are able to recognise and communicate individuals' affective states is of high clinical interest.

Furthermore, various commercially exploitable applications regarding affect recognition in healthy individuals in modern computing and communication systems can be thought of.

1.2. Problem Statement

Current state-of-the-art computing or communication systems lack the ability to communicate their users' affect based on their brain activity. Especially a motor-impaired patient population could benefit from affect recognition systems, which has not yet been targeted in research. To date, there are ambiguous results regarding affect recognition even in healthy. Therefore, experimental paradigms, that easily translate to patient populations, have to be designed and executed to record electrophysiological data that contain affective information for analysis. A sufficiently large trial size is a main constraining factor in affect induction and classification studies. Psychophysiological correlates of affect have to be investigated and confirmed in the data recorded, before classification. Once validated, only discriminating features of affect must be selected and subjected to classification. Finally, valid machine learning approaches have to be applied to train and test models that are able to predict the users' affective state.

1.3. Thesis Structure

To approach the problems outlined, this doctoral thesis is structured as follows.
Chapter 2 constitutes the theoretical background regarding models and electrophysiological correlates of affect, the idea of affective computing, an overview of brain-computer interfaces, the state-of-the-art in affect recognition, as well as the employed classification apparatus including feature selection and extraction as well as performance analysis strategies.
Chapter 3 addresses the design of an auditory emotion induction paradigm with a focus on maximizing trial size that easily translates to a patient population. The paradigm is applied in a healthy as well as in a motor-impaired population with cerebral palsy. Correlates of affect stated in the literature are investigated in the time and frequency domain, then only validated features are subjected to classification.
To investigate affective states and affect recognition on the most fundamental level, Chapter 4 outlines the analysis and classification of affective data recorded from 6-months-old infants' brains while preverbal infants interacted with one of their parents in emotional scenarios.
In order to validate the method developed in the previous chapters and compare its performance to existing approaches, it is applied to a publicly available affect dataset based on emotion induction by music videos as depicted in Chapter 5.
Chapter 6 concludes the work presented and outlines strategies for affect classification as derived during the course of the present work. Also, remaining issues and future directions will be discussed.

"A computer will do what you tell it to do, but that may be much different from what you had in mind."

Joseph Weizenbaum (1923 – 2008 CE)

2 Theoretical Background

This chapter provides a theoretical background regarding concepts and methods employed. Firstly, the term affect is coined and main theories of emotion are introduced. Secondly, the concepts of affective and physiological computing are explained. Thirdly, electrophysiological correlates of affect in the peripheral and central nervous systems are characterized. Lastly, brain-computer interface systems and their different components are discussed, also with respect to active, reactive, as well as passive input.

2.1. Affect - Emotions, Feelings, and Moods

Emotion is an ambiguous term whose meaning has been intensely debated by scientists and philosophers for centuries. Antonio Damasio's definition of emotions as ”bioregulatory reactions aimed at the promotion, directly or indirectly, of the sort of physiological states that secure not just survival, but [...] [also] well-being” [3] has afforded researchers a modern, popular, and practical starting point from which to address emotion. From this viewpoint, emotions are considered short-lasting (seconds to a few minutes), universal, and elicited by the evaluation of a stimulus like a person, event, or object. This description contrasts emotions to longer lasting moods (hours to days), which are considered to be tendencies towards certain emotions. This view is also in-line with work conducted by Scherer [4]. Furthermore, feelings are regarded as mental representations of physiological changes which occur during emotions. Emotions, feelings, and moods therefore constitute the term affect. To give an example for each of these aspects of affect: fear is an emotion, restlessness a feeling, and anxiety a mood. This work employs the terms affect and emotion as synonyms in order to refer to the investigated phenomena.

As emotions are a complex multi-dimensional phenomenon (e.g. see [5]), a complementing definition of emotion was given by Kleinginna and Kleinginna [6]:

> "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural-hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal- directed, and adaptive."

Notably, both definitions are based on hypotheses postulated over a century ago. On the verge of the 20th century William James and Carl Lange independently hypothesized that ”bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion” [7]. Thus following the James-Lange view, an emotion is experienced after bodily changes caused by preceding physiological changes. It is important to note that the focus in this view lied on bare-eyed observable bodily changes (e.g. fleeing or crying) and not so much on the micro-scale physiological changes preceding those observable, as investigated nowadays. Based on a critical examination of the James-Lange view, Walter B. Cannon presented his own theory in 1927 [8]. He postulated that emotional experiences are grounded on subcortical activity (e.g. in the thalamus) and that peripheral activity is not necessary for emotional experiences. Approximately 40 years later, Schachter and Singer also added an important theory on emotion in 1967. They proposed that bodily changes only qualify as emotions in evaluation of objects or events that are emotionally relevant and may be attributed to these changes [9] (also see Section 2.1.3).

Another approach to define the term affect is by contrasting it to cognition. This idea follows the notion that phenomena related to affect are subjective and intuitive whilst aspects of cognition are objective and explicable. However, it is increasingly reported that affect and cognition are strongly entangled [10, 3].

The ongoing debate in emotion theory is of high relevance to the area of affective brain-computer interface systems as well as their development. As outlined in Section 2.5, such systems rely on distinct physiological changes in the peripheral or central nervous system

used as control signals. Distinct physiological patterns are a prerequisite for the success of machine learning and ultimately affect recognition.
In the following, three main theories useful for the study of emotion in an affective computing context will be outlined: the basic emotion, the dimensional, and the appraisal theory of emotion. However, these theories will not be addressed in more detail throughout the course of the present work for the focus lies on the feasibility of affect recognition by physiological signals and machine learning.

2.1.1. Discrete Emotion Theory

Discrete emotion theory [11, 12] suggests that humans express emotions based on on the combination of basic emotions and that these emotions are universal, partially inherited, and physiologically distinguishable from one another. Ekman and Friesen [13] investigated emotion expression in different ethnicities and showed that basic emotions are present in all of them. However, they concluded that deviations within these are possible due to social learning. Based on previous work, Ekman suggested six basic emotions (anger, happiness, disgust, surprise, sadness, and fear) [14] of which facial expression examples are shown in Figure 2.1. Ekman later on extended the set of basic emotions by embarrassment, shame, guilt, pride in achievement, relief, satisfaction, sensory pleasure, amusement, contempt, contentment, and excitement [15]. Obviously, definitions on inconsistent sets of basic emotions are cumbersome. This inconsistency is also the main criticism of discrete emotion theory.

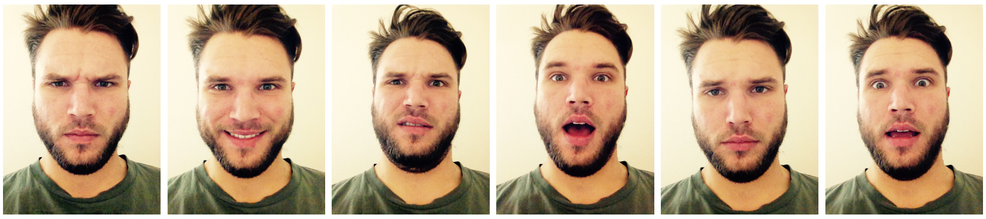


Figure 2.1.: Facial expressions examples of six basic emotions after Ekman [14] from left to right: anger, happiness, disgust, surprise, sadness, and fear)

2.1.2. Dimensional Emotion Theory

Dimensional emotion theory, originating from the model by Wundt [16], proposes that emotions are largely explained by the dimensions valence and arousal [17]. Valence is whether the emotion is subjectively felt as positive/pleasant or negative/unpleasant, and arousal is the subjective energetic activation from deactivated/calm to activated/excited associated with the emotion. The concept of valence has been found to be present in every

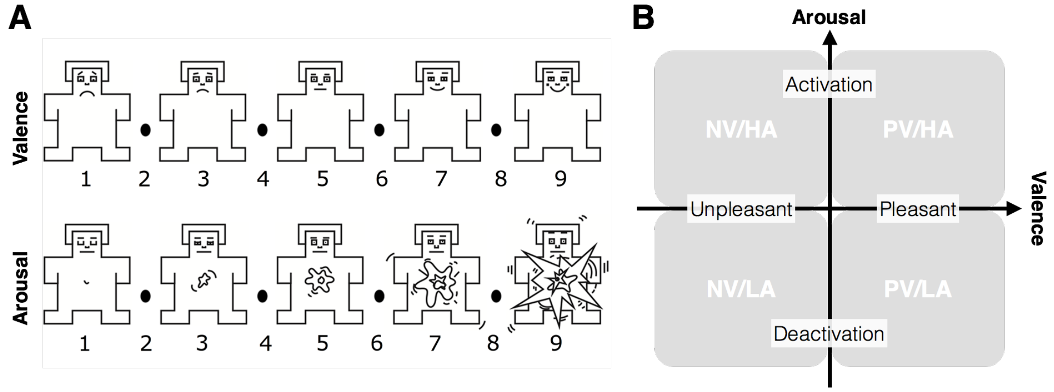


Figure 2.2.: **(A)** Self-assessment manikin in valence (top) and arousal dimension (bottom) on Likert-like scale from 1 to 9 after [18]. **(B)** Continuous valence from unpleasant to pleasant and arousal from deactivation to activation. Quadrants depict groups of negative valence / low arousal (bottom left), positive valence / low arousal (bottom right), positive valence / high arousal (top right), and negative valence / high arousal (top left).

culture [19, 20]. Even infants, a few days of age, feel pleasure or discomfort [21] and can distinguish between unpleasant or pleasant facial expressions in others [22]. As validated by factor analyses, linearly scaled valence and arousal spanning a two-dimensional space cover a wide range of discrete emotions and their combinations (Figure 2.2 B). Furthermore, additional dimensions such as dominance (being in control/being controlled) or tension, which appears similar to arousal, were found to explain more variance, however less consistently [5]. Consequently, discrete emotional responses can coherently be grouped into categories, e.g. an unpleasant, a neutral, or a pleasant category regarding valence. An important advantage of dimensional emotion theory is the possibility to easily obtain participants' self-reported valence and arousal in response to emotional stimuli. Participants rate the quality and intensity of an emotional response by the help of the self-assessment manikin (SAM) [23] (Figure 2.2 A) on a Likert-like scale.

2.1.3. Appraisal Theory of Emotion

Appraisal theory of emotion states that emotions are the consequence of evaluations (appraisals) of stimuli such as persons, events, or objects that cause specific (emotional) reactions in specific contexts [9]. These evaluations may consist of a complex interrelated processing cascade involving cognitive, sensory, and neurophysiological components [4]. Different stimulus dependent analysis levels such as relevance, importance for current goals, coping-potentials, and normative significance are checked in this cascade. As a consequence, physiological and motor reactions are elicited or prepared, respectively.

Whilst discrete and dimensional emotion theory mainly attribute the emotional response to a stimulus, appraisal theory of emotion tries to shed more light on the process that led to an emotional response. Nonetheless, the proposed theories are not mutually exclusive yet rather cover different areas in the complex domain of affect and can also be combined (e.g. pleasant/unpleasant appraisal following the notion of valence) [24].

Related to the appraisal theory of emotion, Barrett [25] argues against discrete hard-coded emotions inside the brain yet for the evaluation of emotional stimuli in an introspective memory-like fashion.

Now that there is an understanding of affect and different theories of emotion, electrophysiological correlates of affect in the peripheral and central nervous system will be discussed. For further reading on the topic of emotion theories, the interested reader is directed to [21, 24].

2.2. Electrophysiological Correlates of Affect

Following the definition of emotion as bioregulatory reactions by Damasio [3], affect can be studied through psychophysiological signals from the peripheral and central nervous system, through audio recordings of speech signals, and through video-recordings of facial expressions. Research in computational linguistics as well as computer graphics provides a wealth of articles on the assessment of affect from audio- and video-signals (see [26] for review), respectively. However as for signal acquisition, both modalities require the active participation of users (e.g. verbal speech or facial muscle activity) which may not be available in disabled or paralyzed individuals. Thus, the focus lies on correlates of affect in the peripheral, and most importantly for the present work, the central nervous system.

2.2.1. The Peripheral Nervous System

The peripheral nervous system (PNS) comprises of nerve fibers and nerve cells outside the brain or spinal chord. It is divided into the somatic nervous system (SNS) and the autonomous nervous system (ANS). The former is responsible for voluntary muscle control as well as reflex behavior via efferent nerve fibers projecting from the central nervous system into the body. The latter consists of the sympathetic and the parasympathetic sub-system which act as control systems that maintain bodily functions. These sub-systems in turn form a nerve fiber network between the central nervous system, numerous internal organs, and various glands throughout the body. Short linguistic idioms for the function of the sympathetic, i.e. "fight or flight", as well as for the parasympathetic, i.e. "rest and digest", systems are well-known. Thus, the sympathetic system allocates bodily resources

for mental or motor activity, whereas the parasympathetic system brings the body into a relaxed state in order to maintain homeostasis by also managing intestinal activity.
Employing different recording techniques, analyses of peripheral signals have produced a multitude of findings regarding correlates of affect. Popular psychophysiological measures are electromyography (EMG), electrocardiography (ECG), skin conductivity by galvanic skin response (GSR), blood volume pressure (BVP), and respiratory activity (RSP).
Commonly, recordings of electrophysiological biosignals in the PNS require a technical setup consisting of electrodes attached to the body, impedance reduction by conductive measure (e.g. gel), amplification, filtering, digitization, and storage in a computer system.

Electromyography

Electromyography is the recording of electrical activity produced by the enervation of skeletal muscles via nerve cells employing electrodes. Already Darwin wrote exhaustively about the importance of posture and facial expressions related to emotion [27] which was continued by Ekman [15]. Thus, the EMG of facial muscles is a prominent location to derive psychophysiological information about emotion expression. With regard to the ANS and SNS, the notion of elevated emotional arousal and overall increased muscle tension is obvious. Furthermore, facial EMG discriminates emotional valence and arousal [28, 29]. The activity of the muscles *zygomaticus major* (smiling) as well as *corrugator supercilli* (frowning) are often recorded for affect recognition (see [30] for review). Interestingly, differences in facial emotion expression in non-depressed and depressed individuals during emotion imagery have been found [31]. Nonetheless, these types of measures are not of interest in search for physiological control signals of affect in a disabled or paralyzed population.

Electrocardiography

Electrocardiography is the recording of cardiovascular activity produced by the enervation of the heart employing electrodes. The heart is a unique muscle within the body that is responsible to maintain blood circulation by continuous contractions. To supply the body, the heart is connected to three vascular networks consisting of the pulmonary circulation, the coronary circulation, and the systemic circulation. Following this order, one network cycle of contraction starts at the lungs for O_2/CO_2 exchange in the blood, then goes to the heart, and finally reaches the rest of the body.
A common measure for cardiovascular activity is the heart beat rate. The average heart beat rate changes dramatically during lifetime. Newborns in the first month exhibit a high heart beat activity of 70 - 190 beats per minute (BPM) during rest. The upper bound decreases

greatly with age, as for infants (1 - 11 months of age) show on average a resting heart beat rate of 80 - 160 BPM. For healthy adults and children over 10 years of age, the heart beat rate ranges on average from 60 - 100 BPM. (Well trained athletes have an even lower resting heart beat rate of down to 40 BPM.)

With regard to the ANS, the heart beat rate is directly coupled with the activity of the sympathetic nervous system. If sympathetic activity increases, the heart beat rate increases and vice versa. Thus the heart rate may react to a stressing stimulus with increased activity to allocate resources for a potential fight or flight scenario. On the other hand, the heart beat rate is inversely coupled with activity of the parasympathetic nervous system. Thus, if parasympathetic activity increases, the heart beat rate decreases and vice versa (e.g. in the absence of stressors during relaxation, the heart rate decreases). Both relationships between heart rate and ANS activity are not mutually exclusive but can coexist [32]. The effects of arousing stimuli on the heart beat rate have been investigated in the literature by the presentation of pictures (see Section 2.5 in [33]), sounds (see Figure 5 in [34]), and videos (see Figure 1 A in [35]). Regarding dimensional emotion theory, stimulus valence has been found to influence heart rate. Accordingly, heart rate is decreased for unpleasant and increased for pleasant pictures [33] or scenic sounds [34]. Interestingly, Goldstein presented in his article "Thrills in response to music and other stimuli" [36] the impact of music on affective states of the listeners. Heart rate has also been found to be responsive to the emotional valence and arousal of music pieces [37]. Besides heart beat rate, blood volume pressure, as well as respiratory activity are further measures related to cardiovascular activity.

Skin Conductivity

Skin conductivity refers to altered conductivity of the skin due to sweat gland activity which is guided by the sympathetic part of the ANS. Most commonly skin conductivity is measured by the galvanic skin response. Hereby, two electrodes are attached to skin of the palm or food a couple of centimeters apart from each other, a small harmless current is then applied to one electrode and measured at the other to obtain the level of skin conductivity. Even non-perceivable deviations in gland activity in terms of sweat are measurable by GSR. Emotional arousal influences sympathetic activity and is thus found to alter sweat gland activity which in turn alters the GSR. Skin conductivity was higher for pleasant or unpleasant stimuli as compared to neutral ones. Arousal is robustly expressed in GSR in studies with arousing pictures [38, 39] and scenic sounds [34, 40]. Skin conductivity has been found to be influenced by music clips of different arousal levels [41].

2.2.2. The Central Nervous System

The central nervous system comprises of the brain and spinal chord. The brain is the central element in integrating information received by afferent nerve fibers as well as distributing information via efferent nerve fibers throughout the spinal chord to body parts. Anatomically, the brain is divided into various structures. The larger part of the brain, the cerebrum, can be divided into six lobes: frontal or (neo)cortex, temporal, parietal, occipital, limbic, and insular cortex. The other, smaller, part is the cerebellum.

Evidence for structures associated with affective responses have been found by a number of investigations employing different technical measures and approaches. Historically, evidence about brain structures was derived from lesion studies. Although lesion studies continue to benefit neuroscience, the possibility to record physiological activity of the brain has provided researchers with new insights. The first such tool was electroencephalography (EEG), introduced in the first third of the 20th century. This was followed by brain imaging techniques such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), functional near-infrared spectrometry (fNIRS), or magnetoencephalography (MEG) that have been introduced throughout the remainder of that century (all of which have their strengths and weaknesses). Based on brain imaging reviews [42, 43, 44, 45, 46], brain structures associated with affective processing and affective responses are briefly outlined.

The main network of affective processes is located towards the ventral structures of the brain. The limbic system which is a complex connection of parts located on both temporal sides of the thalamus ventral to the cerebrum has been identified as a vital part in processing sensory information, contextualizing, and estimating the effect of an internal or external emotionally relevant event. It includes but is not limited to structures such as the amygdala, the insular cortex, the hippocampus, as well as the striatum. Lesions in the amygdala have been found to interfere with both positive or negative emotional reactions. The amygdala is viewed as a connectivity hub of major sensory input from the thalamus and higher-order association areas of the cortex. Simultaneously, the amygdala projects to the brainstem controlling emotional responses such as behavioural responses (e.g. facial expressions, or freezing) or autonomic nervous system responses (e.g. endocrine responses that lead to sympathetic or parasympathetic de/activation). The insular cortex is associated with introspective features of the body (e.g. skin condition, posture, or information about inner organs) that are also integrated during the evaluation of the stimulus event. Further circuits such as the orbitofrontal cortex (OFC) and parts of prefrontal cortex (PFC) have been found to be active during affective processing in brain imaging studies (see [45] Section 5.3 ff.). The OFC is supposed to allow for flexible reactions during the integration of the initial coding of the stimulus event by the amygdala. The PFC is thought to link already integrated information about the stimulus event to visceromotor actions. Information integration and execution resulting from the interplay of these circuits has been described as an "affective neural reference space" [47] which serves as a "valence-general affective workspace". This view has been supported by the latest meta study on brain structures sensitive to emotional valence of stimuli [46].

The following will focus on correlates of affect in the EEG time- and frequency domain as EEG is the most common and well-established technique for non-invasive recordings of brain activity especially in the field of brain-computer interfacing. The next chapter is devoted to EEG since it is the main topic of the present thesis.

2.3. Electroencephalography

Electroencephalography is the recording of summed electrical activity along the scalp produced by the firing of neurons within the brain employing electrodes. Hans Berger introduced this technique for monitoring electrophysiological activity within the brain [48]. Berger discovered an oscillating pattern in the electrical signal within the frequency range 8 - 12 Hz which he named alpha waves, since it was the very first brain signal ever discovered. Subsequently, more repetitive patterns have been discovered and named. Oscillatory brain activity originates from synchronized events in billions of neurons. Table 2.1 gives an overview of typical brain oscillations and their frequency ranges.

Table 2.1.: Names and greek symbols of typical brain waves as well as their frequency ranges.

Name	Symbol	Frequency Range [Hz]
delta	δ	0 - 4
theta	θ	5 - 7
alpha	α	8 - 12
mu	μ	8 - 13
beta	β	13 - 30
gamma	γ	40 - 100

Electrophysiological recordings usually follow a common approach where electrodes are attached to the scalp following the standardized 10/20 electrode location system [49], amplified and digitized by an analog-to-digital converter at a certain sampling frequency. Formally, sampling is the process of converting a signal from a function of continuous time or space into a numeric sequence, i.e. a function of discrete time or space.

When recording EEG, typical filters are high-, and low-pass filters of 0.1 Hz to 1 Hz and 30 Hz respectively. Also a notch filter of 50 Hz (Europe) or 60 Hz (USA) is applied to filter out voltage phase artifacts of power lines. Other artifacts are electrogalvanic, movement, or electromyographic artifacts. The first two are filtered by the high-pass filter and the latter, being a high-frequency artifact, is filtered out by the low-pass filter.

In a referential montage, the EEG of each channel depicts the voltage difference of the electrode at that channel referenced to a designated electrode. The signal is also grounded to another electrode. Established positions for those are left or right mastoids as well as a frontal midline location for the study of hemispheric differences. The EEG is sampled in an analog-to-digital converter. Formally, sampling is the process of converting a signal from a function of continuous time or space into a numeric sequence, i.e. a function of discrete time or space. Commonly, absolute EEG amplitudes measured on the scalp are between 5 - 50 μV and decline with age [50]. Furthermore, frequency bands in infant EEG have been found to be shifted to the left in terms of frequency compared to adult EEG. Infant alpha waves have been found to be within 6 - 9 Hz during the first year into early childhood [51].

In the following, time- and frequency domain correlates of affect in the EEG will be outlined. Furthermore for each band in the frequency domain, their respective general functions are outlined followed by correlates of affect.

2.3.1. Time Domain

Time domain correlates of affect refer to altered deflections of event-related potential amplitudes in reaction to emotional stimuli.

Typically, event-related potentials are computed by averaging amplitudes in the EEG over multiple trials of the same stimulus condition relative to stimulus-onset. Averaging increases the signal-to-noise ratio of a stimulus-driven functional brain response with respect to background EEG activity.

Researchers have reported conflicting evidence that early components of visually-induced ERPs, e.g. P1, N1, or N2, are modulated by stimulus valence. These modulations are thought to reflect the increased attention towards emotional stimuli [52]. Emotional valence as well as arousal have been reported to modulate late components (greater 300 ms after stimulus-onset) known as the late positive potential (LPP) to varying degrees (see [53] for review). Late positive potential amplitudes have been found to be more positive to pleasant and unpleasant stimuli compared to neutral with respect to emotional valence, as well as for emotionally arousing stimuli [54, 55]. Although, emotional valence and arousal are entangled, findings about LPP amplitude modulations have been more consistent with regard to emotional arousal (at least for pictures [53]). Another ERP sensitive to affective stimulation is the P300 which is involved in attention towards the saliency of a stimulus. P300 amplitudes show a larger positive deflection for very rare and highly emotional stimuli [56]. Furthermore, the P300 is a common control signal for brain-computer interfacing (see Section 2.5.4).

In an affect manipulation study of 7-month-old infants, an altered ERP due to the presentations of fearful and happy faces has been described. The midlatency (700 ms) negative central component (Nc) of ERPs has been found to be more negative in response to fearful faces than to happy ones which is attributed to increased recruitment of attentional resources [57].

Although there is a large body of research regarding affective manipulation of ERP amplitudes, the LPP has not yet been classified in a machine learning approach which is addressed in the present work in Chapter 3. At the same time, the practical realization of employing ERPs as a control signal in affective brain-computer interfacing is rather difficult due to their requirement of repetitive stimulation and response averaging. In that regard, frequency domain features of affect could be promising as a control signal for an affect recognition system.

2.3.2. Frequency Domain

Frequency domain correlates of affect refer to deviations in the spectral dimension of transformed time domain EEG data attributed to affective manipulation.

To obtain the spectral dimension of an EEG recording, various techniques are available. Fast Fourier transform (FFT), spectral density estimation by autoregressive models (AR) after Welch [58] or the maximum entropy method (MEM) after Burg [59] are fundamental approaches for the transformation of discrete time signals into power spectra. (For further reading on the topic, the interested reader is referred to the excellent book by Oppenheim et al. [60].) Various typical oscillatory patterns in the EEG have been investigated throughout the last century with the help of spectral methods (see Table 2.1). The following will briefly review the (currently known) functions of these typical bands also with respect to affective processing in the second halves for each section and frequency band.

Delta Frequency Band

The delta frequency band is within the range of 0 - 4 Hz. Oscillations in this band are associated with sleep and an aroused brain as has been found in investigations of the thalamocortical network [61]. During wakefulness, delta oscillations have been attributed to homeostatic and motivational as well as during the transition into slow wave sleep (SWS) (see reviews [62, 63] and review [64] with emphasis on neurotransmitters). Based on a review on ERPs and oscillations [65], delta and theta activity has been theorized to possibly generate or influence the P300 ERP. With regard to affective responses to emotional facial expression, an increase in delta band power has been reported over posterior sites [66], however for all emotional conditions including neutral. A follow-up study attributed increased delta activity to stimulus updates [67]. Focusing on delta oscillations and affect, a study has found effects sensitive to emotional arousal and valence [68].

Theta Frequency Band

The theta frequency band is within the range of 4 - 8 Hz. There is a large body of research regarding the role of theta oscillations in cognition and affective responses. Fundamental research regarding working memory, as a part of cognition, has been conducted by Klimesch et al. [69, 70]. Specifically, theta band power is increased in response to higher workload demands and is thought to reflect information integration vital to executive function [71].

Regarding affect manipulation, increased theta activity has first been reported in 1950 [72] as "hedonic theta" occurring when pleasurable stimulation was aborted. Research in 6-month-old infants to 6-year-old children has revealed increased theta activity in response to pleasant stimuli [73]. In an infant population, literature findings suggest that the type (e.g. unpleasant or pleasant) of an emotional experience can be discerned by power differences across frontal hemispheres [74, 75, 76, 77]. Thereafter, elevated frontal left-hemispheric activity is associated with a pleasant emotional experience as compared to the contra-lateral hemisphere. An unpleasant or aversive emotional experience leads to relative higher frontal right-hemispheric activity as compared to the contra-lateral region. This concept has been introduced as appraisal theory of emotion (see Chapter 2.1.3). Recent investigations with different stimulus modalities have reported increased theta over frontal and/or parietal regions in response to arousing stimuli [78, 67]. Emotional valence has also been associated with increased theta activity in fronto-medial regions [37, 79]. During sleep, pre-frontal theta has been found to be relevant for emotional memory consolidation during rapid-eye movement (REM) sleep [80].

Alpha Frequency Band

The alpha frequency band is within the range of 8 - 12 Hz. This historically famous type of oscillatory activity is exhibited the most over parietal to occipital regions especially during wakeful relaxation when the eyes are closed. Increased alpha activity is thought to reflect inhibitory activity when certain brain regions are idle during wakeful relaxation [81, 73]. This has been validated by skin conductivity measures reflecting overall arousal in healthy adults [82] and 8-12-year-old children [83].

Furthermore, alpha activity decreases in the presence of sensory stimulation indicating the allocation of sensory input processing cortical regions. Alpha activity is associated with sensorimotor activity, whereas motor activity is emphasized. It is then known as the mu rhythm within a similar frequency range of 8 - 13 Hz. Anatomically, the mu rhythm is located over central regions where the frontal meets the parietal lobe (central sulcus). The mu rhythm is a well established control signal for brain-computer interfaces. It is coherently altered during motor execution but more importantly also during motor imagery which is detectable in single trial classification [84].

The notions of event-related desynchronization (ERD) as well as event-related synchronization (ERS) of alpha oscillations reflect a decrease and an increase in alpha band power, respectively. The method to compute ERS and ERD, which are not exclusive to alpha activity, has been introduced by Pfurtscheller and Da Silva [85]. Alpha ERS has been found during working memory tasks [81].

Regarding affect manipulation and originally stated by Davidson in 1982 [74], frontal alpha power asymmetry has been described [86, 87, 88]. Thereafter, alpha is increased over frontal left hemispheric regions in response to appetitive stimuli, as opposed to increased right hemispheric frontal alpha activity in response to aversive stimuli compared to corresponding alpha activity at the contra-lateral hemisphere, respectively. This is also known as the approach-withdrawal theory of affect for which Harmon-Jones et al. provide considerable work [89, 90, 91, 92]. Underlying lateralized neural structures responsive to specific stimulus properties have been thought to account for asymmetrical activity in the alpha band. However, latest evidence based on a review [46] has shown that the existence of specific neurons only responsive to certain types of emotional stimuli is unlikely (see explanation on neural structures of affect in the beginning of Section 2.2.2). Fox et al. have found asymmetrical brain activity in newborns in response to appetitive and aversive taste [93] and by facial-signs of emotion in 10-month-old children [76]. Although there are several studies reporting asymmetrical brain activity in the alpha band during affect manipulation, numerous studies failed to validate this effect [94, 95, 96]. Recently, frontal alpha asymmetry has been described to play a role in cognitive processes related to workload [97].

Beta Frequency Band

The beta frequency band is within the range of 13 - 30 Hz. Over the central sensorimotor cortex and with the lower frequency boundary overlapping with the mu rhythm, beta waves are associated with motor planning, motor execution, and processing of sensory input [98] (e.g. visual [99]). Thereafter, beta oscillations decrease during motor execution but increase when voluntarily withstanding movement impulses [100]. In the review "beta-band oscillations – signalling the status quo?" [101], the authors propose a general theory where beta oscillations govern the upkeep of sensorimotor areas in a buffer-like top-down fashion (see also [102]). Recently, beta band activity in auditory pathways has been linked to speech recognition [103].

Regarding affective manipulation with unpleasant and pleasant pictures or affect imagery, increased lateralized beta activity has been found [104, 105]. Decreased beta activity has been found for relevant emotional stimulus events as compared to neutral ones [106]. A recent study has also found decreased beta activity whilst pictures of emotional faces were viewed during simultaneous pain induction which initially led to an increase in beta band activity [107].

Gamma Frequency Band

The gamma frequency band is within the range of 40 - 100 Hz. Oscillations in the gamma band are thought to be important during cognitive processes, mainly information integration in cortical circuits [102, 108]. Furthermore, gamma band oscillations have been found during multi-sensory integration [109, 110] as well as during attention and memory relevant tasks [111]. Interestingly, dysfunctional gamma band oscillations have been associated with working memory and other cognitive deficits in schizophrenia (see [112] for review). Meditation experts have shown increased baseline gamma activity which is explained by highly trained selective attention (see [113] for review). In the infant brain, increased gamma oscillations have been associated during object recognition tasks which are also related to selective attention [77].

Regarding affective manipulation and emotional valence, a proportional relation in temporal gamma has been reported by [105]. Aversive pictures enhance mid gamma activity (40 - 45 Hz) shortly after stimulus-onset, whereas arousing pictures elicit higher gamma activity (46 - 65 Hz) 500 ms after stimulus onset compared to neutral, as has been reported by [114] (see also [115]). During pain induction, increased gamma band activity has been found over the somatosensory cortex whilst watching fearful faces as compared to angry faces which might reflect avoidance behavior [107]. However, if the appraisal of emotional stimuli does not yield a subjective emotional experience, decreased gamma band power has been reported [106]. Gamma oscillations haven been linked to emotional memory consolidation [116].

To conclude this section about central nervous system activity recorded by EEG, it is obviously non-trivial to attribute a plethora of reported cognitive or affective processes to specific neural substrates, event-related potentials, or specific frequency bands. Nonetheless, a consensus of the research body has been presented regarding affect manipulation of the late positive potential as well frontal alpha band power asymmetries in electroencephalography data. Furthermore, lateralized beta and higher gamma are worth investigating but estimated not as promising. Especially frequency domain correlates of affect in the slower bands are of interest as control signals for affective brain-computer interfacing.

The following section will further elaborate the ideas of affective and physiological computing.

2.4. Affective and Physiological Computing

Originating from the article by Rosalind Picard in 1995 [10], affective computing can be defined as the study and development of systems and devices that are able to recognize, interpret, process, and simulate human affect. The modern field of affective computing, combining the study of affect and computing, is an interdisciplinary endeavour mainly consisting of computer science, neuroscience, and psychology. As outlined before however, the origins of this field date back to the verge of the 20th century [7].

The key principles of user input in human-computer interaction, namely a typewriter-style keyboard and mouse, have not changed in their core since personal computers were first sold in 1965. To alleviate and improve user experience, human-computer interaction continues to study and to explore design principles in hard- and software. Numerous improvements including speed, size, and portability of computing systems and communication devices have been made, yet the core principles of user input remain unchanged. This classic mode of human-computer interaction is asymmetrical in terms of information exchange [117]. To elaborate on this thought, the machine is able to provide a plethora of information based on its inner state (e.g. CPU speed, RAM usage, information stored on hard drive(s), network connection to other machines, etc.), yet the inner state of the user (e.g. intent, cognitive, or affective state) remains hidden for the machine except for overt commands the user sends via keyboard and/or mouse. Thus, Allanson and Fairclough suggest a new mode of human-computer interaction, where system interaction is achieved by monitoring, analyzing, and responding to covert psychophysiological activity from the user in real-time [118]. Similar to brain-computer interfaces (see Section 2.5), such systems transform psychophysiological data into a control signal without any conscious actions from the user. To get back to the thought on asymmetry, the mode of interaction would be rendered symmetrical if such systems work seamlessly. Therefore, Fairclough proposes in his article on physiological computing an extension to the idea of affective computing. Thereafter, additional psychophysiological user input is not limited to affective information but also includes information about the user's cognitive state (e.g. workload, attention, or vigilance).

Brain-computer interface systems are in a sense physiological computing systems yet with the limitation of only conveying messages or commands to the external world by active or reactive physiological changes (in a sense emulating the keyboard or mouse). Currently, such systems still require experts for setup and operation. As of now, seamlessly working affective or physiological computing systems as described by Picard and Fairclough are dreams of the future, yet there is a growing body of research regarding affect classification (see Section 2.8). Nonetheless, the young fields of affective and physiological computing are expanding and the work presented here seeks to add information for the realization of the common goals.

The design and structure of brain-computer interface systems with a focus on affect is outlined in the following section.

2.5. Brain-computer Interfaces

In the original definition, BCI systems allow users to actively convey intent (e.g. messages or commands) to the external world without passing the brain's motor output pathways [119]. Recently, this definition was extended by additionally conveying information about the users' inner state (e.g. emotional, cognitive, or physical) [120]. At the same time, BCI input is not anymore exclusive to brain activity but further biosignals from the PNS are employed. In this context, multi-modal input BCIs are also referred to as hybrid-BCI (hBCI) systems [121]. The present work will employ the abbreviation BCI referring to brain-state-based control signals and hBCI for central- and peripheral-based control signals combined.

2.5.1. Overview

The BCI, being a communication or control system, is composed of input and output channels, components that translate input into output, as well as a protocol that controls interaction and timing of all components. Figure 2.3 shows these components and their basic interactions. After signal acquisition, the key part in any (h)BCI system is signal processing consisting of feature extraction and a translation algorithm that interprets extracted features into device commands.

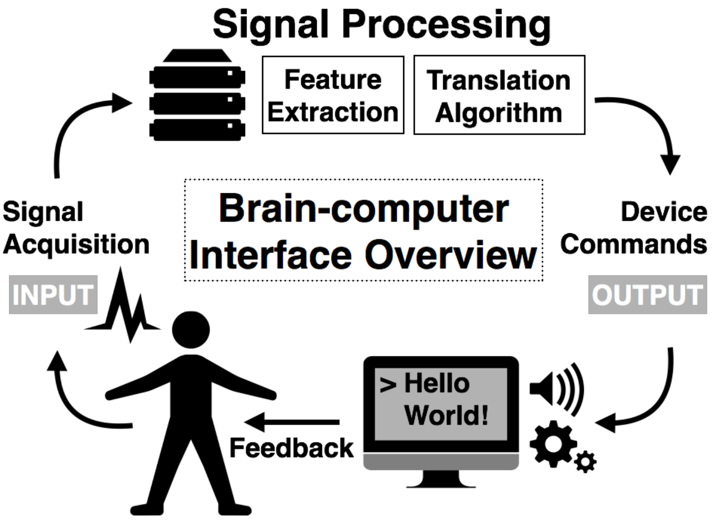


Figure 2.3.: Basic design and operation of any BCI system defined by [119]. Physiological input is recorded from the user and digitized. Meaningful features related to the intent or inner state of the user are extracted, translated into messages or device commands, and fed back to the user.

2.5.2. Active, Reactive, and Passive Input

In the original paper by Vidal [1], physiological activity recorded from the central nervous system (CNS) was proposed to serve as input to control hBCI systems. In active and reactive hBCI systems (Figure 2.4), users convey information by either voluntarily altering physiological signals (e.g. motor imagery) or by attention (e.g. oddball paradigm). With the formulation of passive hBCI systems [120], hBCI input was to be complemented by passively gaining information about the user's inner state (e.g. emotional, cognitive, or physical). Consequently, input signals have not anymore been limited to the CNS yet extended to the peripheral nervous system which contains vital information about the user's inner state (see Section 2.2). Thus, passive hBCI systems are formulated as an augmentation to established active and reactive BCI communication.

Figure 2.4 shows an overview of the described systematics of active, reactive, and passive input for which the original modules of a hBCI system still exist. Central to the idea of passive hBCI systems is the possibility that family members or caregivers act upon a detected inner state of the user. A number of ethical challenges arise with the setup of passive BCI systems, which shall not be discussed herin. However, the interested reader is directed to the ethical reviews [122].

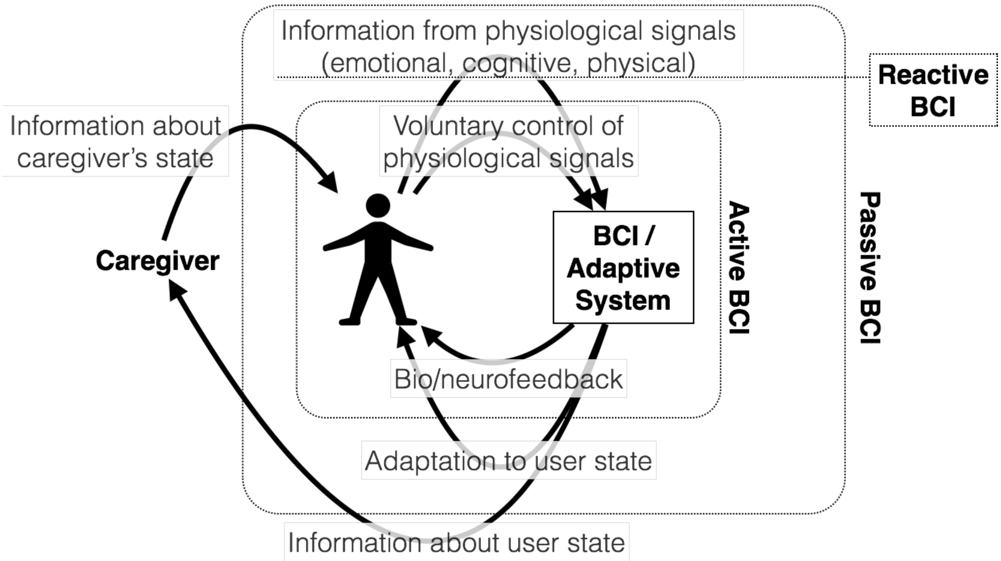

Information from physiological signals (emotional, cognitive, physical)
Reactive BCI
Information about caregiver's state
Voluntary control of physiological signals
Caregiver
BCI / Adaptive System
Active BCI
Passive BCI
Bio/neurofeedback
Adaptation to user state
Information about user state

Figure 2.4.: User-centered schematic of active, reactive, and passive BCI systems along with caregiver and interactions after [120].

2.5.3. Signal Acquisition

Various possibilities to record physiological signals from the CNS are available, but not necessarily practical for hBCI communication. Electrophysiology measured on the skin or scalp is a minimally, well-established, non-invasive method to acquire biosignals (see Chapter 2.3).

In terms of BCI, overall signal quality of non-invasive EEG is good enough to ensure reliable communication, i.e. classification accuracies of $\geq$ 70 % [123].

2.5.4. Control Signals

Slow cortical potentials (SCPs), mu- and beta rhythms, and the event-related potential P300 are possible signals for BCI control [124, 84]. SCPs are negative or positive polarizations in the EEG that last from 300 ms to several seconds [125]. ERPs are, as suggested by the name, neuronal reactions to visual, auditory, or other stimuli that result in amplitude deflections in the EEG [126]. Such EEG signals are time-, and phase-coupled. The neuronal source of these activations is the somatosensory cortex (lateral post-central gyrus) since activations are caused by somatosensory stimuli. Besides an evoked response, an induced one exists which is elicited in the cortex subsequent to ongoing higher mental processes. Such induced responses are rather time-, but not phase-coupled since they result from synchronization and desyncronization processes. As mentioned earlier, event-related synchronization and event-related desynchronization effects are observable in the synchronous alpha rhythm (8 - 12 Hz) after motions are imagined. Then the alpha rhythm desynchronizes over the sensorimotor cortex (central sulcus; pre-central gyrus) which is at this locus also known as mu rhythm. The P300 as the most common control signal for a selective attention and therefore a reactive BCI is described in the next paragraph.

P300

The P300 shows a reproducible positive amplitude at roughly 300 ms after stimulus onset and is typically measured most strongly by the electrodes covering the central-parietal lobe in the EEG [127, 128]. A P300 comprises of subcomponents such as the P3 and P3a, and a subsequent slow wave [126].

The P300 is mainly involved in the process of decision making and is therefore elicited if a target criterion is met. For example, two different tones, the first being the target and the second being a non-target, are randomly presented to a subject. The subject is told to count each occurence of the target tone, while the non-target tone is presented in greater abundance. A P300 is elicited each time the subject distinguishes a target. In short, the P300 response is evoked by attention to rare stimuli in a random order series of stimulus events (i.e. oddball paradigm) [127]. The robust reproducibility makes the P300 a common choice as a BCI control signal in the EEG [2].

2.5.5. Signal Processing

Signal processing in BCI systems consists of filtering, feature extraction, and translation into device commands. The filtering commonly includes artifact rejection of (eye) movement artifacts. For eliminating movement artifacts, statistical methods such as independent component analysis (ICA) can be employed [129], however yielding altered EEG due to the various shapes of movement artifacts. To discard well-defined eye movements, the electroocculogram (EOG) is recorded and then regressed out of the EEG [130]. The latter has proven useful in the analysis of EEG for BCI control.

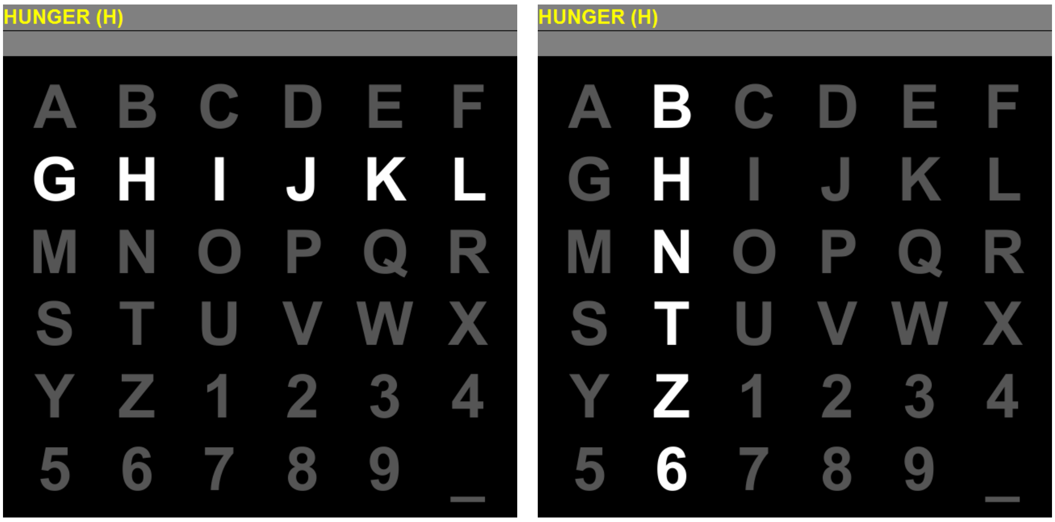
There is a multitude of feature extraction methods of which each has their pros and cons. The technicalities of the pleathora of feature extraction methods are not within the scope of this thesis. However, the fast feature selection method based on Pearson correlation by Spüler et al. [131] has proven to be very versatile and will be employed throughout the work presented. For further information on signal processing in BCI, the interested reader is directed to the reviews [132, 133].

The translation algorithm consists of a classification or regression algorithm that computes a model of the relation of physiological states in recorded data and target labels of classes or numeric values, respectively. Besides the well-established stepwise linear discriminant analysis (SWLDA) for the classification of brain states in the EEG, support vector machine (SVM) classifiers have been increasingly employed. As a note, linear or quadratic discriminant analysis (L/QDA) methods have also been employed. Lotte et al. provide an excellent review on this topic [123]. The authors state that SVMs are particularly efficient for BCI due to their regularization property as well as their immunity against the curse-of-dimensionality.

2.5.6. Application: P300 Speller

As a practical example of a BCI application, the P300 speller is described in the following. Also, trial size and bitrate are given for comparison to affect recognition studies outlined in Chapter 2.8.

A common software system for the realization of BCI paradigms is BCI2000 [134]. The P300 speller has often been realized with a regular alphabet, either visual or auditory [135, 136, 137]. Figure 2.5 depicts the matrix view of the P300 speller with flashing row and column, respectively.



HUNGER (H)

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	_

HUNGER (H)

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	_

Figure 2.5.: Example P300 speller matrix with letters, numerals, and underscore realized in BCI2000 with flashing row and column. User feedback is given on top of the screen in plain text.

Users are instructed to focus attention on the symbol-to-select in the matrix. Rows and columns intensify in a random order. Users select a symbol by focusing and/or counting the intensifications of a target symbol. Thus, a P300 is elicited each time the user identifies/counts the intensification of a target. Event-related potentials are classified by stepwise linear discriminant analysis in an online and offline setting.

User feedback is realized via an output-line on the computer screen above the matrix. On average, motor impaired individuals yield approximately 1.2 selections per minute with this setup [137].

In the context of machine learning, there are on average 180 trials available for the classification of one symbol in the P300 speller paradigm.

Symbol content of the P300 speller is exchangeable. For the illiterate, Blissymbols offer an augmentative and pictographic symbol language especially designed for individuals with speech impairments [138], e.g. motor-impaired individuals with cerebral palsy.

2.6. User Groups and Motivation

Brain-computer interfaces for communication and control have been well-established in a paralyzed population. Many different disorders can disrupt neuromuscular communication channels or render people paralyzed. Amyotrophic lateral sclerosis (ALS), brainstem stroke, brain or spinal cord injury, cerebral palsy (CP), muscular dystrophies, multiple sclerosis, and numerous other diseases impair the neural pathways that control muscles or impair the muscles themselves. These diseases disable patients in communication over time. Patients who are almost completely paralyzed, but have residual voluntary control over a few muscles, such as eye movement, eye blinks, or twitches with the lip, are referred to as being in the locked-in state (LIS). Patients may also be in the complete locked-in state (CLIS), e.g. in the end-stage of ALS, in which all motor control is lost [139]. The main targets for brain-computer interfacing have been individuals in the LIS and individuals in stroke rehabilitation. Naturally, the holy grail of brain-computer interfacing is to restore communication in the CLIS which has not been achieved to date.

Another population with severe motor impairments comprise individuals with cerebral palsy. Cerebral palsy is an umbrella term for non strictly defined motor impairments caused by damage to the newborn or infant brain up to three years of age (see [140] for review). The prevalence of CP ranges from 1 to 4 per 1000 births of a defined age range. Motor impairments often affect the movement apparatus to various degrees including spasticity and dyskinetic movements. Dyskinetic CP is usually accompanied by impairments to oral communication from early childhood. The absence of communication may cause the intricate emotional needs to be forgotten leading to psychological conditions. At 12 years of age, 40 % of children with CP require professional psychological help [141]. Furthermore, prolonged physical impairments may cause intense chronic pain which is often co-morbid with depression [142] and social isolation [143].

Individuals with CP have not been in the focus of BCI research yet. However, the benefits of brain-computer interfacing are of high clinical interest in this population. Since individuals with CP often lack the ability to exhibit emotions by quantifiable physical behaviours,

psychophysiological information from the PNS or CNS offers a promising alternative to access the their inner affective or cognitive state. Thus, passive affective BCIs pose a multitude of advantages for users, families, and caregivers. Brain-computer interfacing or psychophysiological affect have not yet been investigated in this population.
Furthermore, preverbal infants up to 6 months of age account for an interesting "model" for the study of affective processing due to their "purity" (i.e. less cultural learning). The vision behind affect classification in preverbal infants is to provide an emotional communication channel between the child on a sensory deprived or severely impaired caretaker. To date, preverbal infants have not been addressed in brain-computer interface or affect recognition research.

2.7. Classification

In machine learning or pattern recognition, classification is the process of identifying the class membership of an unknown observation based on training data. Two types of learning are distinguished from each other. In supervised learning, training data consist of class-determining data points and known class labels. Besides supervised classification, there exist regression methods which identify the outcome of an unknown observation on a continues scale based on training data. If class labels are unknown or unavailable, the process is known as clustering and referred to an unsupervised learning approach.
The present work deals with supervised learning problems. A variety of classification algorithms exist. These algorithms are often formulated as mathematical optimization problems and simply referred to as classifiers. The very first pattern recognition algorithm was introduced by Fisher in 1936 [144]. Fisher considered two normally distributed populations of data and has shown an optimal (Bayesian) solution in form of a quadratic function which, based on populations' characteristics, degenerates to a linear function. Subsequently, this linear or quadratic discriminant analysis algorithm has constituted the basis for many classification algorithms employed until today [123].

2.7.1. EEG Data and Classification Basics

In brain state classification, EEG data consists of a two dimensional matrix $X \in \mathbb{R}^{c,s}$, where $c \in \mathbb{N}$ is the number of channels and s the number of data samples. (Please note that the introduced notation $M \in \mathbb{R}^{i_1,\ldots,i_n}$ denotes the cardinality of a n dimensional matrix M consisting of values $m_{i_1,\ldots,i_n} \in \mathbb{R}$.) Matrix X is then epoched into intervals (usually according to stimuli) resulting in a three dimensional matrix $X_{epoched} \in \mathbb{R}^{e,c,s'}$, where e is the number of epochs and $s' = \lceil s/e \rceil$ is the number of samples per trial. Alternatively in an online scenario, a ringbuffer is employed that equals one trial when fully filled. For further processing, each two-dimensional epoch consisting of channels times samples is collapsed into a one-dimensional vector where channels are concatenated. Therefrom, features are extracted using an appropriate method, e.g. moving average filter [145] or feature selection based on R^2-values (Section 2.7.2), which results in a data matrix $X_{features} \in \mathbb{R}^{e,j}$, with $j \in \mathbb{N}$. The number of features is determined by the feature selection method or by hand. In classification, rows in $X_{features}$ consist of a feature vector of a single epoch that is associated with a discrete value, i.e. the class label, $y_i \in Y$. Well established normalized class values for y_i are -1 and 1. In regression, target values $y_i \in [a,b]$, where $a \in \mathbb{R}$ is the lower bound and $b \in \mathbb{R}$ the upper bound of the interval, are (usually) continous with data-specific resolution.

Each classification algorithm computes a model out of training data. Based on that model, predictions about the class affiliation of future incoming data are possible.

2.7.2. Feature Selection by R^2-values

To reduce the number of features, R^2-values between data and labels are computed for each feature and the features with the highest R^2-values are used for classification [131]. Correlation in statistics indicates the strength and direction of a linear relationship between two random variables. This coefficient is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. It is a measure describing the amount of variability in one variable that is explained by the other.

The basis of the coefficient of determination R^2 is the correlation coefficient R:

$$R = \frac{\mathrm{cov}_{xy}}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} \tag{2.1}$$

with N being the number of observations, $\bar{x}$ and $\bar{y}$ being the mean of the samples in X and Y, x_i and y_i being data points and s_x and s_y being the standard deviations of X and Y; $R \in [-1,1]$, where 1 means a positive correlation and -1 means a negative correlation. In other words, if $R = 1$ and x increases in a certain way, then y has to increase in a similar way. If $R = -1$ and x increases, then y has to decrease and vice versa.

This coefficient represents no causality between the variables x and y. A squared correlation coefficient represents this causality. It is then called the coefficient of determination [146].

2.7.3. Support Vector Machine

The support vector machine classifier was introduced by Vapnik in 1995 [147]. In its standard definition, the SVM is the formulation of a geometric and data-driven minimization problem that finds a hyperplane best separating datapoints of two classes under certain conditions. The SVM is also known as a large margin classifier for it finds a hyperplane from which the distances to datapoints of either class are maximal. The closest datapoints to the hyperplane are called support vectors. Euclidean distances from support vectors to the hyperplane are defined as $||w||$. In case of a linear kernel and a key point in SVM, $||w||$ can be expressed by a linear combination of support vectors [147].
The core principle of training a SVM model is best explained in an example. Consider n datapoints X and class labels Y, such that $(X;Y) = (x_1,..., x_n\ ;\ y_1,..., y_n)$, here $x_i \in \mathbb{R}^2$ and $y_i \in \{-1,1\}$. Let $w_o\ \cdot\ x_i + b_0$ be the optimal hyperplane in feature space.

$$\min ||w||, \text{ such that } \forall\ y_i,\ y_i(w\ \cdot\ x_i - b)\ \geq 1 \tag{2.2}$$

Figure 2.6 visualizes an example with $n = 7$ datapoints for each class.

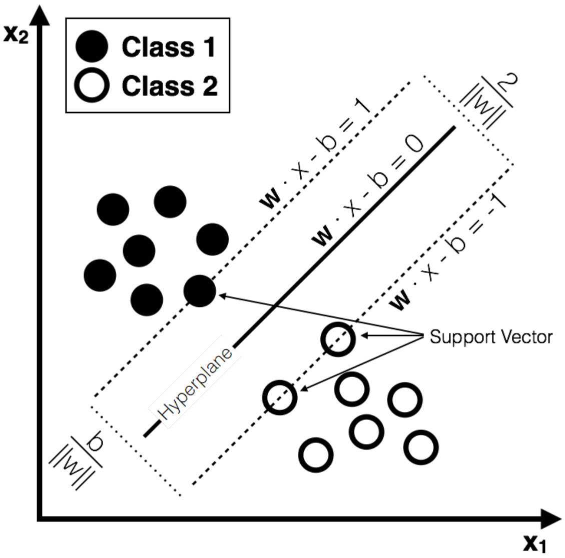


Figure 2.6.: Schematic of training a support vector machine model based on seven datapoints per class and two features.

Equation 2.2 describes a hard-margin SVM classifier. This approach is prone to overfitting the training data. In prediction, obtained models therefore suffer from a lack of generalizing future data. The introduction of soft-margin SVM classifiers overcomes the issue of overfitting. Therefore, a cost parameter C weighting errors and slack variabels ξ_i determining an error are introduced to the original equation.

$$\min_{w,\xi,b} \left\{ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \right\}, \text{ such that } \forall\, y_i,\; y_i(w \cdot x_i - b) \;\geq\; 1 - \xi_i,\; \xi_i \;\geq\; 0 \tag{2.3}$$

If C is small, the penalty for errors is minuscule leading to more errors and larger margin. If C is large on the other hand, the penalty for erros is considerable leading to a smaller margin. Lastly, if $C = \infty$, the hard-margin SVM is obtained, i.e. there are no mistakes in prediction.

Computationally, it is of interest how to solve the minimization problem. Vapnik has shown a re-formulation of the minimization problem (Equation 2.3) in two steps. Firstly, Lagrange multipliers have been introduced which lead to the following minimization problem.

$$\min_{w,\xi,b} \max_{\alpha,\beta} \left\{ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \;-\; \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1 + \xi_i] - \sum_{i=1}^{n} \beta_i \xi_i \right\}, \tag{2.4}$$
$$\text{such that } \alpha_i, \beta_i \;\geq\; 0$$

Secondly, the transformation of the equation into its dual form has been introduced. The dual form determines the lower bound for minimization problems. For convex problems, this lower bound equals the global optimum. This allows for efficient computation since weights $||w||$ and slack variable ξ_i have been eliminated. These concepts have been introduced by Platt and are known as sequential minimum optimization (SMO) [148, 149].

$$\text{Maximize } \forall\, \alpha_i :$$
$$\tilde{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\, \alpha_j\, y_i\, y_j\, \langle x_i, x_j \rangle, \tag{2.5}$$
$$\text{such that } 0 \;\leq\; \alpha_i \;\leq\; C \text{ and } \sum_{i=1}^{n} \alpha_i\, y_i \;=\; 0$$

For linearly separable data, this approach is perfectly fine. For non-linear data however, the so called kernel trick is necessary. The idea of a kernel function is to transfer the data into a higher-dimensional space without exactly knowing that space and furthermore without knowing the exact transfer-function. There are linear kernels $k(x,y) = \langle x,\; y \rangle$, polynomial kernels $k(x,y) = \langle x,\; y \rangle^d$, where $d \in \mathbb{N}$ and radial basis function (rbf) kernels $k(x,y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right)$. For kernel functions, certain formal conditions must hold (i.e. Mercer's Theorem). For an adequate disquisition on kernel functions, the interested reader is directed to [150]). The present work adheres to linear kernel functions.

Although SVM classifiers are in their core only applicable in binary classification problems, they can be extended to multi-class problems (e.g. conduct one-vs-one or one-vs-all

classification, then output highest performance value, majority voting, etc.).
Besides classification, it is possible to employ SVM models for regression.
Performance measures will be explained in the following section. However, the introduction of probabilistic output by Platt [151, 152] for SVM model predictions is of interest for the calculation of certain performance measures.
The open-source implementation of different SVM formulations by Chang and Lin [153] offers efficient calculations in many programming languages including MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States).

2.7.4. Performance Measures and Permutation Tests

To assess classification performance, the present work will investigate three measures: (i) classification accuracy, (ii) area under the curve (AUC) values, and (iii) F1-scores[1]. All three performance measures are different ratios of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
Once performance measures are introduced, permutation tests are discussed regarding the significance of machine learning performance.

Accuracy

Accuracy, as the most prominent performance measure in reporting classification results, is the ratio of TP plus TN divided by the number of test instances. To estimate the quality of classification, obtained accuracy is compared to the chance level of purely random classification. The chance level is dependent upon the numer of instances per class as well as the number of classes. This is best illustrated by a thought experiment. Assuming there is a dataset containing results of $n = 100$ coin tosses. The data consist of either heads or tails. Say coin toss results are highly skewed, heads occurred 90 times and tails 10 times. Now, if a classifier model was to always predict heads, it would achieve 90 TP and 0 TN. Thus, the classification accuracy is $\frac{90+0}{100} = 90$ % in this example.

Area Under Curve

As a second measure for assessing classification performance, area under the curve (AUC) values from receiver operating characteristic (ROC) curves can be computed (see [154] for review). AUC-values are based on true positive and true negative rates computed from thresholds of prediction probabilities of a classifier. The true positive rate is the ratio of TP

[1] The source code for feature reduction, classification by SVM, as well as the computations for accuracies, AUC-values, and F1-scores is freely available at `https://github.com/dthettich/BSClassify`

divided by TP plus FN, whereas the true negative rate is the ratio of TN divided by TN plus FP. To obtain a performance measure that is independent of thresholds, true positive rate and true negative rate are computed by varying thresholds ranging from 0 to 1 in 0.01 steps. The area under the resulting curve is the final AUC-value. As a note for interpretation, AUC-values range from 0 to 1 where 0.5 equals purely random classification, i.e. the classes are statistically identical, values exceeding 0.5 are better than random and vice versa.

F1-score

As a third measure of classification performance, F1-scores reflecting the harmonic mean of true positive rate and positive predictive value of a binary classifier can be computed. Positive predictive value is the ratio of TP divided by TP plus FP. Thus, F1-scores are computed by $\frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$. F1-scores also range from 0 to 1 with purely random classification at 0.5. Scores exceeding 0.5 are better than random and vice versa. Although F1-scores are claimed to account for class imbalance, these scores are unreliable under certain circumstances [155].

Permutation Tests

In a binary classification problem with balanced classes in which the number of instances per class is the same, chance level for accuracy is at 50 %. However, the individual significance level threshold of classification performance scales with the number of instances per class as well as the number of classes [156]. Individual significance level thresholds of classifier performance are obtained in permutation tests [157]. Therefore, for each dataset, classification performance is repeatedly evaluated in multiple iterations (typically 100 or 1000), where on each iteration the class label vector is randomly permuted. A common approach for the evaluation of classification performance is k-fold cross-validation (CV) or leave-one-out-estimation (LOOE). In the former, the dataset is divided into $k \in \mathbb{N}$ mutually exclusive sets. Then the classifier model is repetitively trained on $k-1$ sets and tested on the k-th set. Typical values for k are 5 or 10. In the former, classifier models are repetitively trained on $n-1$ samples of the dataset with size n and then tested on the n-th sample.

Individual significance level thresholds for classification are then obtained by sorting performance values in an increasing fashion and selecting values at the 5 % position for each dataset. If initially computed performances exceed obtained thresholds, classification ac-

curacies are significant at $p = 0.05$.
Since permutation tests are accurate but computationally exhaustive, [158] have shown that individual significance thresholds can be properly approximated for accuracy in the context of BCI research, assuming classification errors follow a binominal cumulative distribution. Accordingly for balanced classes, the individual significance level $c_i(\alpha)$ at a given significance threshold α is computed by the following MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States) code `binoinv(1-α,n,1/c)*100/n` , where n is number of samples per class and c the number of classes. This approximation is only applicable if all classes are balanced. In different circumstances to properly obtain classification accuracy, permutation tests are recommended.

2.8. State-of-the-art in EEG-based Affect Recognition

Since the emergence of affective computing, various attempts to classify affective states in the EEG offline have been conducted. Emotion elicitation paradigms follow either visual stimuli (e.g. pictures form the International Affective Digitized Sounds (IAPS) set [159]) auditory stimuli (e.g. scenic sounds from the International Affective Digitized Sounds (IADS) set [160] or music) , video clips, musical video clips, or emotional recall/imagery. Physiological recordings include signals from the CNS, PNS, or the combination of the two systems. For feature selection and subsequent classification, there is a manifold of techniques available.

2.8.1. Affect Recognition Studies

Selected descriptive affect recognition studies are summarized in the following.
Takahashi (2004) [161] conducted an affect recognition study by inducing pleasure and displeasure in an unknown number of participants with classical music (e.g. vivaldi) and music mixed with white-noise. EEG was recorded on 3 dry electrodes at frontal locations by a headband. Employing a SVM classifier and spectral features, he reported 62.3 % accuracy. In another study, Takahashi [162] induced 5 basic emotions (joy, anger, sadness, fear, and relaxed) with music videos in 12 participants. For each class, two trials are available resulting in 10 trials per participant. EEG was recorded on 3 channels at frontal locations. Employing signal statistics features and one-vs-all SVM, he reported 42 % accuracy in 5-class and 60 % in 2-class classification. In both studies, it is not clearly stated if classification was conducted on pooled or individual datasets. Emotional spectral correlates of affect as well as their potential differences were not statistically validated.
Chanel et al. (2006) [163] classified the arousal dimension by spectral power features of

specific frequency bands and locations relevant for emotion processing. Emotions were induced by a subset of the IAPS (100 pictures) in 4 participants while EEG was recorded on 64 electrodes. Stimulation time was 6 seconds. Trials were equally divided into 2-classes (calm and exciting), as well as 3-classes (calm, neutral, and exciting). The authors state that, labelings led to unbalanced classes (see [163] Figure 2) and adress this issue by imposing an a priori probability of 1/3. Classifiers were a naïve Bayes and Fisher's linear discriminant analysis (LDA). For every participant, LOOE was employed for performance evaluation in conjunction with accuracy. On group average in 2-class classification, 54 % accuracy were obtained in the Bayes approach and 55 % for LDA.

Chanel et al. (2009) [164] conducted another study in. In that study, emotions were induced by mental imagery of three states in a recall paradigm in 10 participants. States are defined in valance-arousal space as negatively excited, positively excited, and calm-neutral states. Mental imagery was cued with a descriptive image of the target state and lasted 8 seconds. EEG was recorded from 64 electrodes. Using LOOE and linear SVM, 3-class classification achieved 63 % accuracy and 2-class 70 %. Time-frequency features and the common information contained at each pair of electrodes served as features.

Horlings (2008) [165] conducted emotion induction in 10 participants by emotional pictures from IAPS while recording EEG on 19 electrodes. In the paradigm, 50 pictures were presented and the self-report of valence and arousal was obtained on a 5-point scale (the 5 classes). Employing various EEG features in a 3-fold CV in 5-class SVM classification, 32 % and 37 % accuracy were reported in the valence and arousal dimension. When the author only classified samples with self-report 1 and 5 in a 2-class approach (approximately 70 % of samples were removed), 71 % and 81 % accuracy in the valence and arousal dimension were reported. Class imbalances are not adressed and class size values are not clearly stated. Furthermore, it remains unclear whether classification was conducted on pooled or individual datasets.

Winkler et al. (2010) [94] investigated frontal EEG asymmetry [166] in response to emotional pictures similar to the IAPS in 9 healthy participants. They selected 48 negative, 48 positive, and 16 neutral pictures for presentation. Pictures were presented randomly for 6 seconds following self-report of valence and arousal by the help of the SAM. EEG was recorded from 32 electrodes. Statistically, significant differences in spectral power between hemispheres were not reported. To distinguish between negative vs. positive emotions, log alpha power features and a common-spatial patter (CSP) approach along with a LDA classifier in 5-fold CV repeated for 5 times were tested. Both approaches performed on group average with 56 % accuracy. On average, there were 78.6 trials available, yet numbers varied for each participant. The authors did not adress class imbalance.

Koelstra et al. (2012) [167] released a publicly available multi-modal physiological dataset of 32 participants for the study of human affective states. Emotion was induced by 40 music videos and self-report of valence, arousal, dominance, and liking was obtained. EEG was recorded from 32 electrodes. Statistically, spectral power and emotional dimension were validated. The authors employed spectral features across frequency bands as well as spectral differences between opposite electrodes for classification in a nïve Bayes classifier. In LOOE for 2-class classification between low and high valence as well as arousal, 57.6 % and 62.0 % accuracy were reported on average. The authors adress the issue of class imbalance by reporting F1-scores: 0.563 for valence and 0.583 for arousal. The authors state that F1-scores take class imbalance into account. Also, class ratios are reported: 57 % for valence and 59 % for arousal. In terms of validating classification performance, the authors perform right-tailed t-tests of F1-scores against 0.5, which they report was chance level for this measure. In that regard, the authors state that group average classification performance in valence and arousal are significant.

Gupta and Falk (2015) [168] employ the DEAP dataset and introduce graph theoretical features in order to account for highly interactive information transfer of active brain networks during emotional processing. They compare classification using spectral features against graph theoretical features. Performance increases of 11 % for valence and 7 % for arousal are reported by employing graph theoretical features. For each participant, LOOE was employed in rbfSVM. The authors specifically investigated spectral power and asymmetry features (as did Koelstra et al.), graph theoretical features, as well as the fusion of the two along with the number of features in relation to classification performance. For the first feature set, 52 % max. accuracy for valence and 54 % for arousal with 60 and 70 features are reported. For the second feature set, 63 % max. accuracy for valence and 61 % for arousal with 135 and 130 features are reported. The third feature set leads to 63 % accuracy for valence and 66 % accuracy for arousal with 350 and 167 features, respectively. The authors do not address class imbalance.

2.8.2. Literature Survey

To limit the search-space of free variables, Mühl et al. have reported a literature survey on the topic of affective computing [169]. The authors have conducted a literature review on the amount of publications including the terms "brain-computer interfaces; emotion affect; affective computing; emotion recognition; EEG fNIRS" has shown a substantial increase since the year 2000 (less then 5 articles) up to the year 2013 (almost 900 articles). From that time period, Mühl et al. provide an excellent survey of 18 curated studies regarding affect recognition from the CNS or PNS. They give information about the number of participants, emotion elicitation method, timing aspects, emotions assessed, signals/senors used, number of channels, signal processing, features, classification/regression, and performance for brain activity only (see Table 1 in [169]). All of these studies vary greatly in their experimental paradigms, methods used for analyses, and presentation of results rendering a clear state-of-the-art statement rather difficult. However, key aspects of the overview in [169] are outlined in the following.

On average, studies were conducted by $16.\overline{3}$ participants with 43.92 s emotional stimulation or recall time of 3.67 emotional classes and recorded from the CNS on 44 channels (28.71 channels, if 306 channel MEG study is excluded).

The studies have employed different materials for emotion elicitation (with counts): 8 IAPS, 3 IADS, 3 video clips, 2 musical 2 video clips, 2 music, 2 recall/imagery, 1 game with different difficulties, and 1 images of facial expressions.

The studies have employed different methods for recording physiological data from the CNS (with counts): 15 EEG, 1 MEG, and 2 fNIRS.

Classification or regression methods have been employed (with counts): 9 SVM, 2 LDA, 1 multi-layer perceptron (MLP), 1 naïve Bayes, 1 fuzzy clustering, 1 QDA, 1 logistic regression, and 1 ridge regression.

Besides different classification performance results, mainly regarding the number of classes, also regression results have been reported (counts with average accuracies): 6 two-class problems (68.33 %), 4 three-class problems (58.66 %), 2 four-class problems (81.50 %), 2 five-class problems (59.50 %), and 2 six-class problems (85.00 %).

To summarize this sample of affect recognition articles, there is a preference for the IAPS (followed by IADS), EEG recordings, and binary classification problems using a SVM classifier. The findings outlined are hardly comparable due to their substantial variances in experimental design and analysis methods. Classification performance measures are not comparable due to these variances, as is outlined in Section 2.7.4.

2.8.3. Key Parameter: Sample Size

A key parameter especially for brain state classification is the amount of trials also known as sample size. The outcome of brain state classification is directly related to the number of classes, class sizes, as well as class distributions [170, 156, 158]. These figures are defined by the experimental paradigm and therefore the number of trials. The amount of trials has not been provided in the overview cited above. However, Mühl provided in his PhD thesis a similar overview table containing information about 19 affect recognition studies based on EEG (see Table 1.1 ff. in [171]; study samples partially intersect). The average amount of trials of those studies is 111.2, however with a standard deviation of 226.1. If one study of this sample based on emotional recall with a trial number of 1000 in one subject is excluded, the mean of total trials is 61.3 with a standard deviation of 71.2. In comparison for example, the amount of trials available for the classification of one target symbol in the P300 speller is 180 in a two-class problem (30 samples target; 150 samples non-target). Recently, Brouwer et al. published recommendations to avoid common pitfalls in the analyses of brain signals that reflect cognitive or affective states [172]. The work presented here seeks to adhere to these with a focus on best practices for conducting and reporting classification results related to brain state classification of affect. Therefore, the present work will extend on these best practices with a focus on machine learning and classification of brain states of affect. Methodological pitfalls in brain state classification and classification performance reporting is outlined.

“ "They will never make a machine to replace the human mind—it does many things which no machine could ever do."
You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that! ´´

J. von Neumann, 1948. (Quoted by E. T. Jaynes in Probability Theory: The Logic of Science, p. 4.)

3 Auditory Affect Induction: Stimuli, Physiology, and Classification

This chapter provides an in-depth view on an auditory affect induction and classification study conducted in a healthy and motor impaired population.

The focus of this study is the investigation of affective processing in the EEG during auditory stimulation by emotional sounds with subsequent classification. Since emotional processing has not yet been researched in individuals with CP and their is no clear consensus on the strategies how to classify electrophysiological correlates of affect in a healthy population, the first milestone of this study is to investigate mentioned goals in a healthy population. These are first steps and groundwork towards an affective BCI for individuals with CP.

The results presented in the following have been obtained during the course of the European Union (EU) project: Augmented BNCI Communication[1] (ABC); supported by the Seventh Framework Programme (FP7) – EU Contract: FP7-ICT-2011-7-287774. The abbreviation BNCI stands for brain-neural-computer interface meaning a BCI system where input is not only limited to brain activity but extended to other signals (e.g. the periphery) as well.

Results presented in the following have been partially published [40, 18].

[1] `http://www.abc-project.eu/`

3.1. Participants

Healthy individuals as well as motor impaired individuals with cerebral palsy participated in the study which was approved by the Ethical Review Board of the Medical Faculty, University of Tübingen.

Healthy

Twenty-five right-handed healthy participants (12 female; age: 24.46 ± 3.17 years) with normal hearing participated in the study. Each participant was informed about the purpose of the study and signed informed consent prior to participation. All participants fully completed the experiment.

Cerebral Palsy

Four participants with cerebral palsy (2 female; age: 18 ± 2.16 years) with normal hearing participated in the study. Handedness was not present in 3 participants due to their motor impairments. Each participant was informed about the purpose of the study and their legal guardian signed informed consent prior to participation. With great effort, 'VPcb' signed informed consent themself. All participants fully completed the experiment.

3.2. Stimuli and Procedure

In an attempt to develop an emotion induction paradigm that yields a sufficiently large number of trials and which would easily translate to patient populations, the International Affective Digitized Sounds 2nd Edition (IADS-2) database [160] was utilized to induce emotion. Sounds in the database are 6 s long stereo audio recordings of scenic or everyday events. Using IADS-2 allows stimulation via the auditory sensory channel, which tends to be intact in many groups that cannot focus on or otherwise exploit visual information (e.g. patients with cerebral palsy). The auditory affect induction paradigm consisted of sixty audio files selected from the IADS-2. All sixty stimuli were categorized into 20 unpleasant events (e.g. vomit, growl, etc.), 20 neutral events (e.g. fan, rooster, etc.), and 20 pleasant events (e.g. baby, laughter, etc.). A list of all sounds with their respective categories is given in Supplementary Table A.1. All sounds were repeated in two separate blocks. Two pseudorandom sequences of consecutive, categorically disjoint sounds were generated for each participant, leading to 120 trials per participant.

Procedure

Healthy participants were seated in a comfortable chair approximately 1 m away from a laptop screen with a 15 inch diameter in a quiet room. Participants completed a German version of the Positive Affect Negative Affect Scale (PANAS) [173, 174] to evaluate current feelings prior to experimentation (see Appendix, Figure A.1). All participants were in a normal and relaxed state with no signs of substantial deviations. Standardized audiometry validated binaural hearing capabilities of each participant. The Presentation software kit (Neurobehavioral Systems, Inc.) was used for stimulus presentation. Auditory stimuli were presented via customary computer loudspeakers (Yamaha Co., Hamamatsu, Japan).

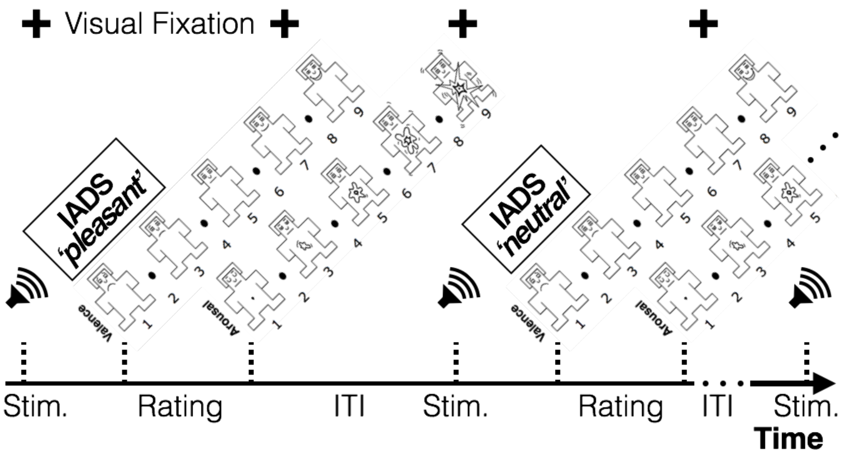


Figure 3.1.: Design of auditory emotion induction paradigm with annotations.

After attachment of electrodes, task instructions were given. Participants were asked to relax and to actively listen to the sounds presented whilst visually focusing a cross on the laptop screen. After presentation of a 12 s baseline sound, the first sequence of sounds was presented. To assess individual valence and arousal ratings, participants were asked to evaluate each sound after sound-offset with the help of the self assessment manikin (SAM) [23] by navigating a 9-point Likert-like scale using the cursor keys on the keyboard. The schematic SAM is shown in Figure 2.2 A. Pressing the up key first confirmed the selection for perceived valence followed by confirmation of the individual arousal rating also marking the end of the trial. The ITI varied randomly between 6 s and 14 s in order to maintain participants' task engagement. After presentation of 60 sounds, participants were allowed to relax their eyes and arms for 5 min. The second sequence of sounds was then presented in the same manner lacking the rating step. The design of the experimental paradigm for the rating run is depicted in Figure 3.1.

For participants with cerebral palsy, valence and arousal values were obtained before the sequential presentation of sounds. Therefore, an input form was realized within a webpage environment. Sounds were played consecutively, whereas after each sound the user input was obtained for valence and arousal. Following the acquisition of valence and arousal values, the regular experiment with sequential presentation of sounds in pseudorandom order was conducted.

3.3. Data Collection and Analysis

The electroencephalogram along with the vertical and horizontal electrooculogram as well as electrocardiography were recorded by active electrodes at 500 Hz sampling frequency and bandpass filtered from 0.1 Hz to 100 Hz (BrainProducts GmbH, Munich, Germany). Following the extended 10-20 system [49], EEG was recorded from Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, Pz, Tp9, Tp10, Fc1, Fc2, Cp1, Cp2, Fc5, Fc6, Cp5, and Cp6 all referenced to Fcz and grounded against Apz. Electrode locations on the scalp are depicted in Figure 3.2. Continuous EEG was corrected for vertical and horizontal eye movement artefacts [130]. EEG was segmented into 6 s long trials relative to stimulus onset. The data of two healthy participants had to be excluded from analysis due to excessive artefacts leading to $n_{healthy} = 23$ datasets for analysis. For cerebral palsy datasets, the amount of movement artifacts poses several problems. Nonetheless, all $n_{cp} = 4$ datasets were included for analysis.

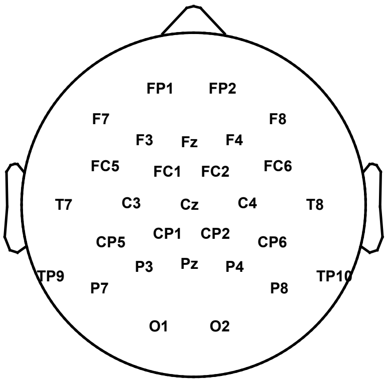


Figure 3.2.: Electrode locations.

3.3.1. Statistics

All data analyses were performed offline with a commercial software package (MATLAB 2014b, The MathWorks, Inc., Natick, Massachusetts, United States), FieldTrip [175], and custom code. For analysis of event-related potentials (ERPs), EEG was bandpass filtered from 0.1 Hz to 30 Hz with a two-pass Butterwerworth filter with order 6 and baseline corrected from -0.1 s to 0 s relative to stimulus onset. Grand average waveforms were computed for each valence category separately. Waveform differences in the time domain were tested for significance for conditions pleasant vs. neutral, unpleasant vs. neutral, and pleasant vs. unpleasant with a Wilcoxon test and corrected for multiple comparisons by false discovery rate (FDR) [176]. Power spectra were computed from time domain data (0 s to 1.4 s relative to stimulus onset) in 1 Hz frequency bins from 1 to 40 Hz by the method of Burg [59] with a model order of 32. Inter-hemispheric differences in power spectra of emotional conditions pleasant and unpleasant at electrode locations F3 and F4 were tested for significance across conditions with a FDR corrected Wilcoxon test. Additionally, scalp topography distributions of spectra for unpleasant minus neutral as well as for pleasant minus neutral conditions were investigated by ANOVA. To analyse if emotional stimuli had an overall effect on power spectra, conducted an ANOVA with factors participant, power per frequency band delta (1 - 4 Hz), theta (5 - 7 Hz), alpha (8 - 12 Hz), and beta (13 - 25 Hz), emotional condition (unpleasant, neutral, and pleasant), as well as channel.

3.3.2. Classification

Classification of valence categories was evaluated by postulating three binary classification problems: unpleasant vs. neutral, unpleasant vs. pleasant, and pleasant vs. neutral. In the following, classes are occasionally abbreviated with '-' for unpleasant, '0' for neutral, and '+' for pleasant.

Feature Extraction and Selection

Based on the neurophysiological analysis presented in results, features were extracted from channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6. To reduce the number of features, R^2-values between data and labels were computed for each feature and the features with the highest R^2-values were used for classification [131]. Initially, the number of features were varied. Only features that exceeded the mean of all computed R^2-values were taken into account for training the classifier model. On average, 1558 features were used for classification with this setting. As best practice however, only the 100 best scoring features in terms of R^2-values were retained for classification throughout the rest of analyses. As a rule of thumb, the number of features should approximately equal the number of samples (80 in the present study). Resulting feature selection masks were applied to test sets before assessing classification quality. As classifier, a support vector machine (SVM) with a linear kernel (C=1) using the libSVM implementation [153] was employed. SVMs have been proven to be well suitable for brain state classification especially in the field of BCI research [123]. Label predictions as well as prediction probabilities [152, 151] were obtained. All performance measures are obtained in a 10-fold cross-validation, i.e. for each participant, feature sets were divided into 10 mutually disjoint training and test sets resulting in 10 sets of 72 training and 8 test instances each. To assess classification performance, three measures were investigated: (i) classification accuracy, (ii) area under the curve (AUC) values, and (iii) F1-scores (see Section 2.7.4).

3.4. Results

Emotional categories unpleasant, neutral, and pleasant differed significantly from each other by IADS-2 normative valence as shown in Figure 3.3 B ($p < 0.01$, Wilcoxon test). Significant differences of literature IADS-2 and participants' self reported valence values were not observed in a Wilcoxon test. Participants' self report was correlated with literature IADS-2 valence values. Self reported valence values of all participants highly correlate with literature IADS-2 valence values ($r = 0.81$, $p < 0.001$) verifying the experimental paradigm.

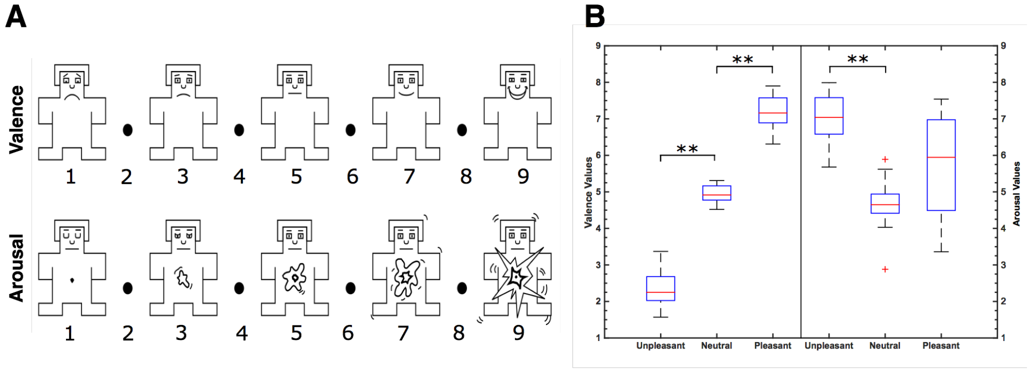


Figure 3.3.: **A**: Self-assessment manikin in the valence (top) and arousal dimension (bottom) [23]. **B**: Valence (left) and arousal (right) value distributions of IADS-2 sounds selected according to categories.

3.4.1. Event-related Potentials and Power Spectra

Healthy

The grand average event-related potential time locked to stimulus onset is shown in Figure 3.4 A for each valance category. Clear potentials are visible for responses to all categories. After a negative peak at approximately 200 ms, waveforms of low and high valence stimuli exhibit a stronger positive deflection than neutral valence stimuli that lasts approximately until 1400 ms. Figure 3.4 B depicts schematic scalp plots showing grand average responses on all channels for all categories on time points when amplitudes were minimal and maximal, respectively. Time points for minima and maxima were computed from channel Pz for each emotional condition. After stimulus-onset, amplitudes are more negative in frontal regions across categories. Topographies of responses to unpleasant and pleasant stimuli result in higher positive amplitudes over centro-parietal regions compared to neutral. Channels Cp1 and Cp2 exhibit the most prominent ERP waveforms with significant responses from 448 ms to 1400 ms for comparison of categories unpleasant and neutral, as well as pleasant and neutral (Figure 3.5 A). On Cp5 and Cp6, only pleasant and neutral responses are significantly different. Marginal inter-hemispheric waveform differences within the same category at electrode locations Cp1 and Cp2, as well as Cp5 and Cp6 were not significant ($p > 0.05$, FDR corrected Wilcoxon test).

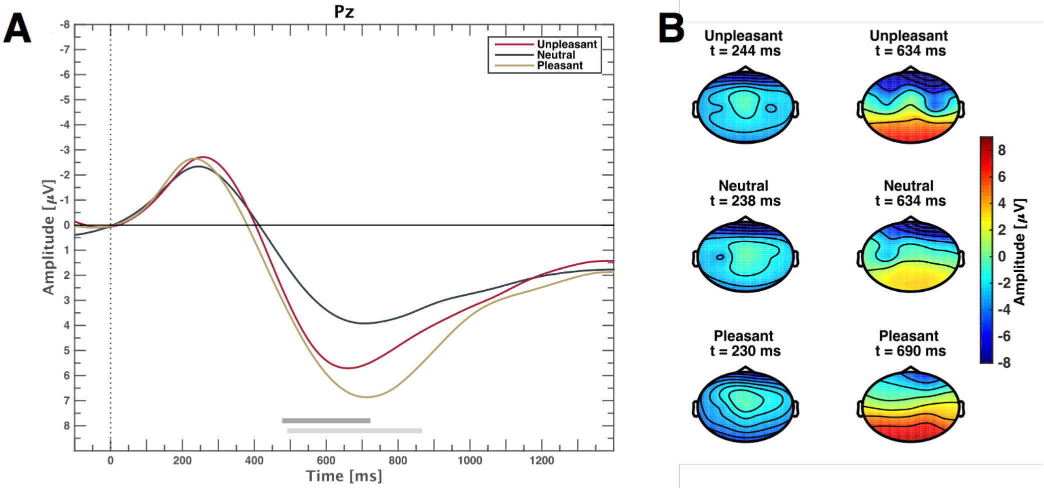


Figure 3.4.: (**A**) Event-related potentials averaged over all participants for unpleasant, neutral, and pleasant stimuli on midline electrode Pz. Grey horizontal bars depict significant differences between neutral and pleasant (light grey) or neutral and unpleasant responses (dark grey), ($p < 0.05$, FDR corrected Wilcoxon test). Differences between unpleasant and pleasant conditions are not significant ($p > 0.05$, FDR corrected Wilcoxon test). (**B**) Scalp plots showing the topographic distribution where grand average responses are minimal (left) and maximal (right) at electrode Pz for unpleasant, neutral, and pleasant stimuli.

In the frequency domain, it was expected that the processing of unpleasant sounds results in higher power in the alpha band (8 - 12 Hz) over right frontal hemispheric regions, whereas power would be elevated over left frontal brain regions for pleasant sounds [166, 177]. Figure 3.6 A depicts spectral differences of unpleasant and neutral and Figure 3.6 B between pleasant and neutral conditions in the frequency bands delta (1 - 4 Hz), theta (5 - 7 Hz), alpha (8 - 12 Hz), and beta (13 - 25 Hz). Unpleasant minus neutral condition shows

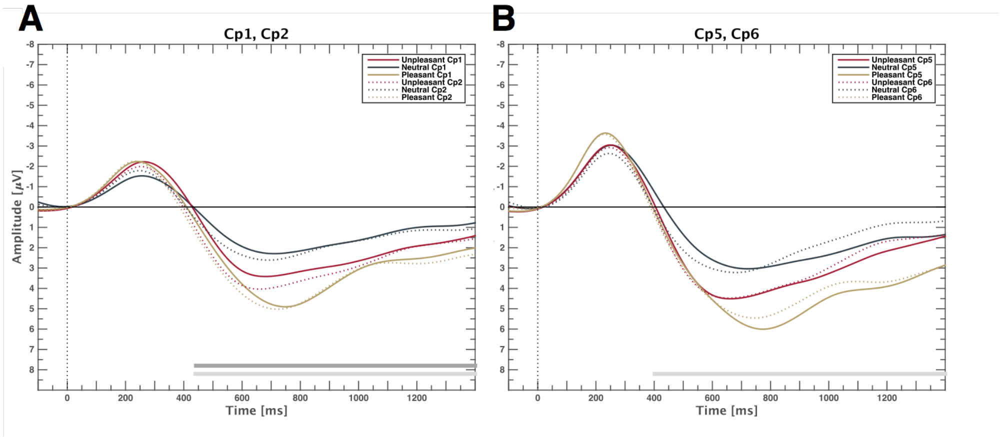


Figure 3.5.: Event-related potentials averaged over all participants for unpleasant, neutral, and pleasant stimuli on temporal electrodes Cp1 and Cp2 (**A**) as well as Cp5 and Cp6 (**B**). Grey horizontal bars depict significant differences between neutral and pleasant (light grey) or neutral and unpleasant responses (dark grey), ($p < 0.05$, FDR corrected Wilcoxon test). Differences between unpleasant and pleasant conditions are not significant ($p > 0.05$, FDR corrected Wilcoxon test).

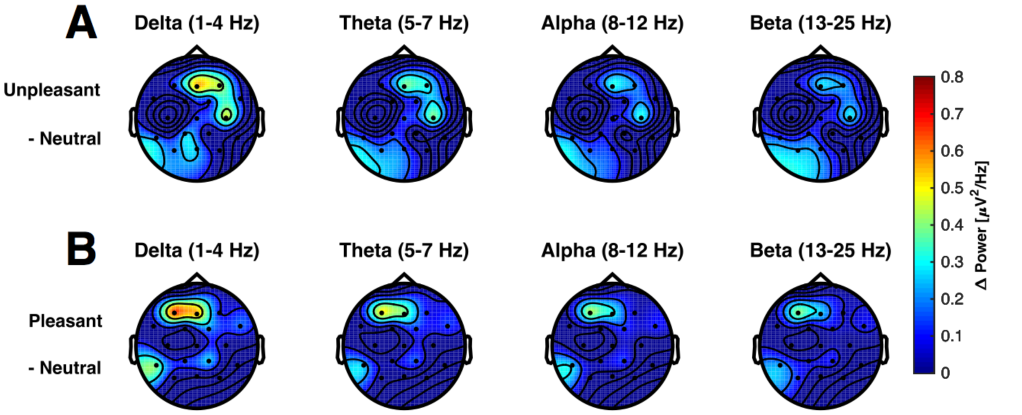


Figure 3.6.: (**A**) Scalp topography plots of grand average spectral differences for unpleasant minus neutral and (**B**) pleasant minus neutral valence categories for different frequency bands.

higher power in delta, theta, and alpha frequency bands over frontal, right hemispheric channels Fz, F4, and Fc2. Left temporal parietal power slightly increases in the beta band for this condition. Condition pleasant minus neutral shows an inverted effect where power is higher over frontal left hemispheric electrode locations Fz and F3, as well as a marginal power increase on P7. These power differences were not significant ($p > 0.05$, Bonferroni corrected ANOVA).

To investigate frontal alpha power asymmetry, power spectra of responses to pleasant and unpleasant stimuli on frontal electrode locations F3 and F4 were compared. As expected, pleasant stimuli exhibit on average higher power compared to unpleasant in the frequency range 1 Hz - 30 Hz on F3 (pleasant; 2.19 ± 1.39 μV^2/Hz; unpleasant: 1.76 ± 1.13 μV^2/Hz), and vice versa on F4 (pleasant: 1.97 ± 1.17 μV^2/Hz; unpleasant: 2.11 ± 1.24 μV^2/Hz), however not significant ($p > 0.05$, FDR corrected Wilcoxon test).

Cerebral Palsy

Analysis of cerebral palsy EEG data poses severe problems regarding artifacts. Motor disabilities including spastic and dyskinetic movements are intrinsic to cerebral palsy. When unphysiological trials with amplitudes greater 100 μV are rejected, only a very small number of trials can be retained for analysis (SC01: 12 trials, SC02: 0 trials, SC03: 45 trials, and SC04: 60 trials). Therefore, the following results have to be treated not only with caution but are an example of most likely false interpretations of machine learning analysis.

The grand-average responses in individuals with cerebral palsy to unpleasant, neutral, and pleasant sounds is shown in Figure 3.7 A recorded at midline electrode Pz. There is a positive deflection for all conditions starting at around 400 ms post stimulus-onset, peaks at 450 ms, declines until 1200 ms post stimulus. The deflection is strongest for unpleasant sounds. Against expectation, pleasant sounds evoked the smallest deflection of amplitudes.

As shown in Figure 3.7 B, schematic scalp topography plots of the different conditions show a high occipital activation during the maxima of amplitudes for unpleasant and neutral sounds.

Topological differences in the conditions unpleasant minus neutral and pleasant minus neutral are depicted in Figure 3.8 A and B, respectively. Activity in the delta band (1 - 4 Hz) show lateral (A) and central differences (B). Theta band activity differences are almost absent. Alpha and beta differences show lateral and frontal differences in A and B, respectively. Lateriazation effects could not be observed. All differences were subjected to a statistical test employing ANOVA which has not led to significant differences across conditions.

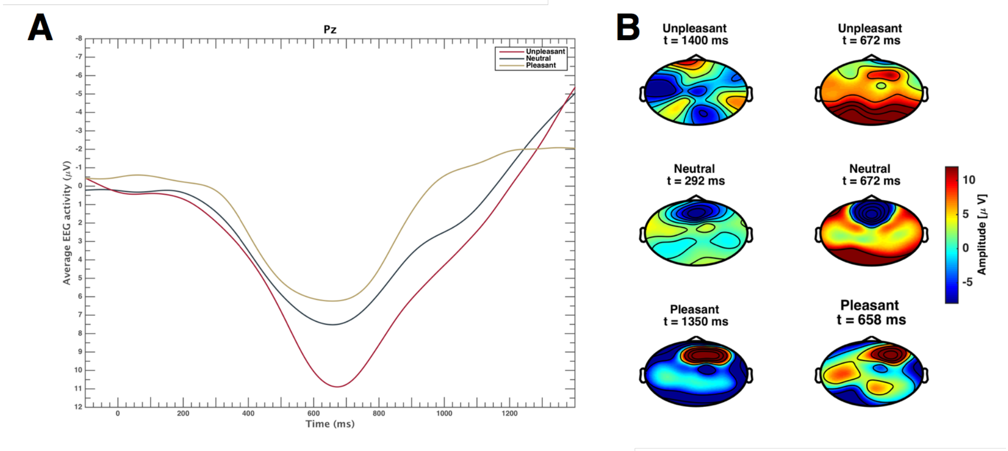


Figure 3.7.: (**A**) Event-related potentials averaged over all participants with CP for unpleasant, neutral, and pleasant stimuli on midline electrode Pz. Amplitude differences are not significant, ($p > 0.05$, FDR corrected Wilcoxon test). (**B**) Scalp plots showing the topographic distribution where grand average responses are minimal (left) and maximal (right) at electrode Pz for unpleasant, neutral, and pleasant stimuli.

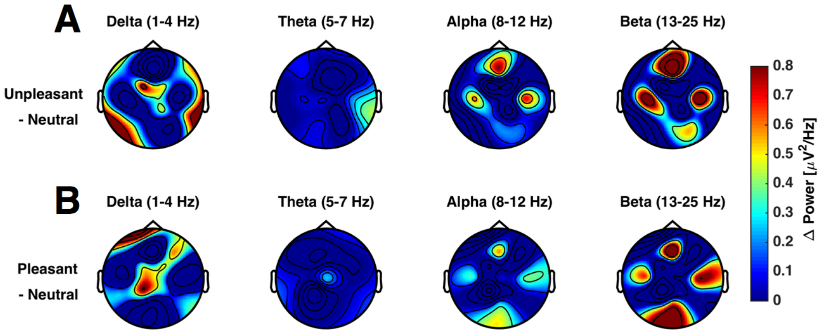


Figure 3.8.: (**A**) Scalp topography plots of grand average spectral differences in CP for unpleasant minus neutral and (**B**) pleasant minus neutral valence categories for different frequency bands.

3.4.2. Time Domain Classification

Healthy

Classification was conducted on time domain EEG data where significant differences were observed between conditions on channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6. Three binary classification problems were postulated according to valence categories: unpleasant vs. neutral, unpleasant vs. pleasant, and pleasant vs. neutral. Table 3.1 depicts average group classification accuracies, AUC-values, and F1-scores. Average group level accuracies and AUC-values for binary classification of unpleasant vs. pleasant and pleasant vs. neutral are significantly above chance.

Table 3.1.: Healthy mean classification accuracies, AUC-values, and F1-scores based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 obtained in 10-fold cross-validation. Columns indicate classes of respective binary classification problems ('-' unpleasant, '0' neutral, '+' pleasant). Classes are balanced with 40 instances each. Stars indicate significant group differences in a right-tailed t-test against 50 for accuracy and 0.5 for AUC-values and F1-scores with $p < 0.05$ and $p < 0.01$, respectively.

	'-' vs. '0'	**'-' vs. '+'**	**'+' vs. '0'**
Accuracy	49.99 %	53.39 % **	53.21 % *
AUC-value	0.49	0.54 **	0.54 *
F1-score	0.46	0.51	0.51

Individual chance levels are derived from permutation test results at $\alpha = 0.5$. Average chance levels for each performance measure are shown in Table 3.2.

Table 3.2.: Average healthy individual chance levels of classification at significance threshold $\alpha = 0.5$ obtained by permutation tests for the performance measures accuracy, AUC-value, and F1-score based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 in 100 iterations. Columns indicate classes of respective binary classification problems ('-' unpleasant, '0' neutral, '+' pleasant). Classes are balanced with 40 instances each.

	'-' vs. '0'	**'-' vs. '+'**	**'+' vs. '0'**
Accuracy	49.53 ± 0.99 %	49.29 ± 0.99 %	49.46 ± 1.18 %
AUC-value	0.50 ± 0.01	0.49 ± 0.01	0.49 ± 0.01
F1-score	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.01

Complete individual chance level results are shown in the Appendix Chapter A, Table A.2. Individual chance levels obtained by permutation tests are not significantly different to expected chance levels at 50 % or 0.5 for their respective performance measures (two-tailed t-test, $p < 0.001$).

Individual participant classification results are shown in Table 3.3.

Table 3.3.: Healthy individual classification accuracies, AUC-values, and F1-scores based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 obtained in 10-fold cross-validation. Columns indicate classes of respective binary classification problems (’-’ unpleasant, ’0’ neutral, ’+’ pleasant). Classes are balanced with 40 instances each. Stars indicate significant group differences in a right-tailed t-test against 50 for accuracy and 0.5 for AUC-values and F1-scores with $p < 0.05$ and $p < 0.01$, respectively.

	’-’ vs. ’0’			**’-’ vs. ’+’**			**’+’ vs. ’0’**		
Participant	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**
S01	**63.75 %**	**0.65**	**0.64**	**67.50 %**	**0.73**	**0.68**	46.25 %	0.49	0.46
S02	54.82 %	0.52	0.51	59.29 %	0.54	0.56	55.00 %	0.54	0.57
S03	52.50 %	0.55	0.53	51.25 %	0.49	0.49	52.50 %	0.52	0.49
S04	53.75 %	0.52	0.51	56.25 %	0.58	0.53	51.25 %	0.46	0.48
S05	56.25 %	0.58	0.56	53.75 %	0.57	0.53	**73.75 %**	**0.82**	**0.71**
S06	45.00 %	0.45	0.39	57.50 %	0.59	0.60	55.00 %	0.54	0.50
S07	46.25 %	0.47	0.38	46.25 %	0.48	0.47	57.50 %	0.56	0.57
S08	41.25 %	0.42	0.30	50.00 %	0.49	0.43	57.50 %	0.59	0.60
S09	46.25 %	0.54	0.48	40.00 %	0.40	0.33	60.00 %	0.59	0.57
S10	47.50 %	0.49	0.40	48.75 %	0.54	0.45	57.50 %	0.62	0.56
S11	47.50 %	0.45	0.49	60.00 %	0.58	0.57	48.75 %	0.49	0.44
S12	47.50 %	0.47	0.45	56.25 %	0.60	0.53	58.75 %	0.58	0.55
S13	55.00 %	0.51	0.54	50.00 %	0.50	0.44	62.50 %	0.65	0.58
S14	47.50 %	0.41	0.30	48.75 %	0.49	0.37	46.25 %	0.51	0.38
S15	50.00 %	0.49	0.39	51.25 %	0.49	0.51	48.75 %	0.43	0.48
S16	43.75 %	0.42	0.43	60.00 %	0.60	0.53	47.50 %	0.52	0.50
S17	51.25 %	0.49	0.52	53.75 %	0.57	0.49	43.75 %	0.41	0.33
S18	43.75 %	0.42	0.38	53.75 %	0.52	0.60	45.00 %	0.46	0.41
S19	57.50 %	0.53	0.55	50.00 %	0.47	0.51	48.75 %	0.49	0.42
S20	43.75 %	0.48	0.40	53.75 %	0.51	0.51	60.00 %	0.60	0.60
S21	48.75 %	0.46	0.44	53.75 %	0.47	0.53	41.25 %	0.45	0.41
S22	52.50 %	0.54	0.47	57.50 %	0.58	0.54	55.00 %	0.55	0.56
S23	53.75 %	0.46	0.51	48.75 %	0.52	0.49	51.25 %	0.53	0.55
Mean	49.99 %	0.49	0.46	53.39 % **	0.54 **	0.51	53.21 % *	0.54 *	0.51

Regarding individual classification results in terms of significance levels, only one participant exceeded 62.5 % accuracy for unpleasant vs. neutral and unpleasant vs. pleasant. In the classification of pleasant vs. neutral, one participant exceeded the individual significance level 70.0 % (depicted in bold font; Table 3.3). With a significance threshold of $\alpha = 0.05$, on average 1 in 20 participants was expected to exceed the individual significance level by chance.

To give a valid estimate of individual significance thresholds of classification for the respective performance measure, permutation tests were conducted. Table 3.4 shows individual significance levels at $p = 0.05$ for each participant (for comparison, individual classification performances are shown in Table 3.3). Two participants exceed individual significance levels in all performance measures for the classification of unpleasant vs. neutral, unpleasant vs. pleasant, and pleasant vs. neutral, respectively. One participant slightly exceeded the individual significance level for AUC-values, however not for accuracy nor F1-score. Average accuracy significance thresholds obtained by permutation tests prove the binomial estimate of 62.5 % only with deviations lesser than 0.5 %.

Table 3.4.: Healthy individual significance levels of classification at significance threshold $\alpha = 0.05$ obtained by permutation tests for the performance measures accuracy, AUC-value, and F1-score based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 in 100 iterations. Columns indicate classes of respective binary classification problems (’-’ unpleasant, ’0’ neutral, ’+’ pleasant). Classes are balanced with 40 instances each.

	’-’ vs. ’0’			**’-’ vs. ’+’**			**’+’ vs. ’0’**		
Participant	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**
S01	61.25 %	0.63	0.61	63.75 %	0.67	0.64	61.25 %	0.64	0.60
S02	61.07 %	0.65	0.61	63.21 %	0.65	0.63	63.75 %	0.65	0.62
S03	63.75 %	0.66	0.65	62.50 %	0.66	0.63	63.75 %	0.63	0.63
S04	63.75 %	0.63	0.65	63.75 %	0.63	0.62	60.00 %	0.65	0.60
S05	62.50 %	0.63	0.62	65.00 %	0.68	0.64	61.25 %	0.62	0.61
S06	63.75 %	0.62	0.64	63.75 %	0.64	0.65	65.00 %	0.66	0.63
S07	63.75 %	0.69	0.62	66.25 %	0.65	0.65	63.75 %	0.64	0.63
S08	62.50 %	0.63	0.62	61.25 %	0.63	0.61	61.25 %	0.66	0.63
S09	65.00 %	0.65	0.63	63.75 %	0.65	0.66	62.50 %	0.64	0.63
S10	60.00 %	0.62	0.59	63.75 %	0.66	0.63	61.25 %	0.63	0.64
S11	63.75 %	0.64	0.64	61.25 %	0.67	0.62	63.75 %	0.64	0.63
S12	61.25 %	0.65	0.62	62.50 %	0.66	0.63	62.50 %	0.64	0.64
S13	62.50 %	0.60	0.64	60.00 %	0.62	0.62	62.50 %	0.64	0.63
S14	65.00 %	0.67	0.66	62.50 %	0.67	0.64	63.75 %	0.66	0.62
S15	60.00 %	0.63	0.61	60.00 %	0.64	0.61	62.50 %	0.64	0.63
S16	63.75 %	0.65	0.64	65.00 %	0.66	0.64	61.25 %	0.63	0.61
S17	65.00 %	0.62	0.64	62.50 %	0.67	0.64	62.50 %	0.66	0.64
S18	58.75 %	0.62	0.60	62.50 %	0.62	0.62	62.50 %	0.63	0.62
S19	60.00 %	0.66	0.59	61.25 %	0.64	0.60	62.50 %	0.64	0.62
S20	63.75 %	0.65	0.63	63.75 %	0.65	0.63	62.50 %	0.65	0.62
S21	65.00 %	0.64	0.66	62.50 %	0.67	0.64	62.50 %	0.67	0.62
S22	62.50 %	0.60	0.63	61.25 %	0.64	0.61	62.50 %	0.64	0.65
S23	63.75 %	0.65	0.64	62.50 %	0.62	0.63	65.00 %	0.63	0.66
Mean	62.71 %	0.64	0.63	62.80 %	0.65	0.63	62.61 %	0.64	0.63

Classification Cerebral Palsy

Classification was conducted on time domain EEG data on channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6. However as opposed to healthy data results, differences in ERP amplitudes are not significant. Furthermore, it has been stated that cerebral palsy data are prone to artifacts which in turn lead to classification results most likely stemming from artifacts rather than real physiological effects. The results reported here are an example for the importance of correct data processing in the context of domain knowledge, i.e. numbers are patient. Table 3.5 depicts average group classification performance measures for three binary classification problems as described above.

Table 3.5.: Cerebral palsy mean classification accuracies, AUC-values, and F1-scores based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 obtained in 10-fold cross-validation. Columns indicate classes of respective binary classification problems (’-’ unpleasant, ’0’ neutral, ’+’ pleasant). Classes are balanced with 40 instances each. Stars indicate significant group differences in a right-tailed t-test against 50 for accuracy and 0.5 for AUC-values and F1-scores with $p < 0.05$ and $p < 0.01$, respectively.

	’-’ vs. ’0’	**’-’ vs. ’+’**	**’+’ vs. ’0’**
Accuracy	53.04 %	50.80 %	54.20 % **
AUC-value	0.55 **	0.50	0.52 *
F1-score	0.49	0.49	0.52 *

Individual chance levels are derived from permutation test results at $\alpha = 0.5$. Average chance levels for each performance measure are shown in Table 3.6.

Table 3.6.: Average cerebral palsy individual chance levels of classification at significance threshold $\alpha = 0.5$ obtained by permutation tests for the performance measures accuracy, AUC-value, and F1-score based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 in 100 iterations. Columns indicate classes of respective binary classification problems (’-’ unpleasant, ’0’ neutral, ’+’ pleasant). Classes are balanced with 40 instances each.

	’-’ vs. ’0’	**’-’ vs. ’+’**	**’+’ vs. ’0’**
Accuracy	53.48 ± 2.61 %	49.51 ± 1.42 %	55.40 ± 10.30 %
AUC-value	0.53 ± 0.03	0.50 ± 0.01	0.55 ± 0.10
F1-score	0.52 ± 0.05	0.45 ± 0.09	0.54 ± 0.13

Although accuracy, AUC-value, and F1-score on average exceed 50 % or 0.5 for conditions ’-’ vs. ’0’ and ’+’ vs. ’0’, deviations to estimated chance levels are not significant (two-tailed t-test, $p < 0.01$).

Complete individual chance level results are shown in the Appendix, Table A.2.
As mentioned, amplitude differences are not significant, yet the group average AUC-value is significantly above chance for '-' vs. '0'. For the condition '+' vs. '0', all performance measures are significantly above chance. Table 3.7 shows individual classification performances. The bold accuracy value for SC02 in '+' vs. '0' indicates above individual significance. Thus one participant exceeded individual significance in accuracy. In four participants, 0.2 participants are expected to exceed the individual significance level.

Table 3.7.: Cerebral palsy individual classification accuracies, AUC-values, and F1-scores based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 obtained in 10-fold cross-validation. Columns indicate classes of respective binary classification problems ('-' unpleasant, '0' neutral, '+' pleasant). Classes are balanced with 40 instances each. Stars indicate significant group differences in a right-tailed t-test against 50 for accuracy and 0.5 for AUC-values and F1-scores with $p < 0.05$ and $p < 0.01$, respectively.

	'-' vs. '0'			'-' vs. '+'			'+' vs. '0'		
Partic.	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**
SC01	50.00 %	0.57	0.35	45.00 %	0.39	0.29	55.00 %	0.49	0.49
SC02	60.00 %	0.58	0.53	50.00 %	0.51	0.49	**57.50 %**	0.53	0.51
SC03	55.89 %	0.51	0.62	49.46 %	0.51	0.62	53.04 %	0.56	0.56
SC04	46.25 %	0.52	0.47	58.75 %	0.57	0.56	51.25 %	0.51	0.52
Mean	53.04 %	0.55 **	0.49	50.80 %	0.50	0.49	54.20 % **	0.52 *	0.52 *

Individual significance levels of classification for cerebral palsy data are depicted in Table 3.8. Obtained individual significance levels for accuracy deviate approximately 4.5 % from the pre-computed value of 62.5 %.

Table 3.8.: Cerebral palsy individual significance levels of classification at significance threshold $\alpha = 0.05$ obtained by permutation tests for the performance measures accuracy, AUC-value, and F1-score based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 in 100 iterations. Columns indicate classes of respective binary classification problems ('-' unpleasant, '0' neutral, '+' pleasant). Classes are balanced with 40 instances each.

	'-' vs. '0'			'-' vs. '+'			'+' vs. '0'		
Participant	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**
SC01	57.50 %	0.60	0.57	57.50 %	0.58	0.49	58.75 %	0.60	0.56
SC02	65.00 %	0.67	0.61	55.00 %	0.55	0.53	56.25 %	0.57	0.54
SC03	61.96 %	0.64	0.66	56.07 %	0.56	0.63	75.54 %	0.80	0.78
SC04	58.75 %	0.59	0.59	58.75 %	0.59	0.58	55.00 %	0.53	0.54
Mean	60.80 %	0.63	0.61	56.83 %	0.57	0.56	61.38 %	0.62	0.61

3.5. Discussion

In this study, neural responses to emotion-laden sounds by recording EEG, in the context of affective computing, were investigated in a healthy and a population with cerebral palsy. An auditory emotion induction paradigm also suitable for the study of affect in disabled individuals where visual fixation is absent has been introduced. Following the dimensional model of emotion, sounds were divided by valence into three categories: unpleasant, neutral, and pleasant. Participants' self report of valence values strongly correlated with literature reported IADS-2 values ($r = 0.78$, $p < 0.001$). In healthy data, time domain EEG data analysis showed significant grand average waveform differences related to stimulus valence categories. Inter-hemispheric spectral power differences in the frequency domain related to stimulus valence were not significant. However there was a significant overall effect of stimulus valence to power spectra. Time domain EEG data were subjected to classification using SVM. In healthy data, group level significance for the classification of unpleasant vs. pleasant (53.39 % accuracy, 0.54 AUC-value) and pleasant vs. neutral (53.21 % accuracy, 0.54 AUC-value) conditions. Two participants reached significant individual classification performance in two (unpleasant vs. neutral and unpleasant vs. pleasant) and one condition (pleasant vs. neutral) was found in healthy data.

As for cerebral palsy data, time domain ERP differences were not significant. Classification performance exceeded chance significantly for the condition unpleasant vs. neutral in AUC-values and for condition pleasant vs. neutral in all performance measures. At the same time, classification results must be treated with some reservation due to CP data are potentially prone to artifacts.

For comparison, the methods introduced are applied to the DEAP dataset with some alterations. Complete results regarding the DEAP dataset and a discussion are depicted in Chapter 5. The 'DEAP: a Database for Emotion Analysis using Physiological Signals' dataset has been publicly released by Koelstra et al. [167]. It is a multimodal dataset aimed at the analysis of human affective states.

3.5.1. Event-related Potentials and Power Spectra

Healthy

Neurophysiological results in the time domain are consistent with results from earlier studies on affective picture perception [178, 55]. Emotional sounds (either unpleasant or pleasant) evoked a larger positive deflection than neutral event-related potentials. After an N2 component, positive deflections begin approximately 400 ms after stimulus-onset and last until approximately 1400 ms for unpleasant and pleasant stimuli. Positive deflections to pleasant stimuli are on average stronger compared to those of unpleasant stimuli, however not significantly. Amplitude differences between neutral and unpleasant or neutral and pleasant conditions are significant over midline and centro-parietal electrode sites. Waveforms at electrodes Cp1 and Cp2 exhibit prolonged positive deflections. Although not as prolonged, these results are in line with late positive potential data of [55] during the processing of emotion-laden pictures. The observable N2 preceding the LPP is attributed to auditory processing (see [179] for review). An inter-hemispheric effect of amplitude differences when comparing ERPs of the same condition at Cp1 and Cp2 or Cp5 and Cp6 could not be observed.

Frontal inter-hemispheric differences in frequency domain power related to stimulus valence reported by [180] could not be confirmed. Average power was increased at F3 for pleasant (and decreased for unpleasant) as well as increased power at F4 for unpleasant (and decreased for pleasant) stimuli, however not significant. Similarly, lateralised power differences in frequency bands between unpleasant minus neutral or pleasant minus neutral conditions were not significant. Nonetheless, a significant effect of stimulus valence to spectral power has been found confirming the altered brain activity during processing of stimuli. It is arguable that (not significant) effects in the frequency domain related to hemispheric differences in power and stimulus valence in the present study are attributed to substantial experimental design differences compared to the original study by [180]. The experimental paradigm in that study employed five 60 s video clips to induce two emotional states (happy and disgust), as well as baseline activity. The first video clip accommodated the participant with the experiment, the subsequent two were clips to induce a positive, and finally two clips to induce a negative emotional condition. Thus, the authors remained with a small number of trials whilst obtaining a relatively large amount of EEG data for analyses. In the present study, the total amount of "emotional" EEG recorded seems to be not sufficient to result in significantly measurable power differences in the frequency domain. However, our results in the time domain clearly show the LPP as a neurophysiological marker of valence whilst frequency domain results only in trend. It is to note that the present study is framed within the context of affective computing with focus on the amount of trials whilst retaining a design with controlled stimuli that easily translates to patient populations.

Cerebral Palsy

Event-related potentials in cerebral palsy data show a trend of late positivity in all conditions. However, only unpleasant sounds evoked the strongest positive deflection which is concurrent with the hypothesis. Responses to pleasant stimuli are less positive than neutral which are explained by right-hemispheric frontal artifacts as seen in the topographic ERP plots. The number of trials has been identified as a key parameter for successful machine learning. Therefore, rigorous artifact rejection could not be applied for the loss of a substantial amount of trials in the data recorded from this population. Behaviourally, spastic and dyskinetic movements have been observed to be coupled with participants arousal for inhibition abilities are diminished in CP [181]. This is particularly problematic during stimulus-onset and partially during stimulation itself.

3.5.2. Classification Performance Assessment

The assessment of classification performance is strikingly influenced by the number of classes, class sizes, as well as class distributions. Thus, it is of utmost importance to clearly report these figures, i.e. two classes with 40 instances each in the present study. Performance metrics such as accuracy, AUC-values, and F1-scores entail a couple of methodological problems. Classification accuracy, as the ratio between correctly classified instances and all instances, is probably the most prominent measure for classification quality assessment. In a generic two-, three-, or n-class classification problem, a straight-forward approach is to evaluate classification accuracy in a 10-fold cross-validation and investigate the deviation of obtained accuracy from random classification, i.e. the so called chance level at 50 %, $33.\overline{3}$ %, or $\frac{100}{n}$ %, respectively. The most severe problem is that this computation of chance level is only valid for balanced classes, i.e. the number of instances per class is the same for all classes. Complying with this prerequisite, accuracy computed by 10-fold cross-validation is a valid measure to estimate classification performance against the chance level. As will be outlined in the following, the performance assessment in brain state classification on a participant level requires further measures. From a theoretical point of view, individual significance thresholds in classification only hold for an unlimited number of training and testing instances [156]. Although this limitation is commonly accepted in the machine learning community, it seems not well-established in interdisciplinary fields such as affective computing where studies are especially prone to a small number of trials. To properly estimate individual significance thresholds of classification, it is strongly encouraged to conduct permutation tests. These tests are not only independent of the performance measure, but also independent of class distributions. Since permutation tests can be time consuming, it is suggested to compute individual chance levels according to [158]. Nonetheless, it is to emphasise that this approach is only valid for accuracy and if classes are balanced. In this regard, it is strongly encouraged to design studies such that trials are equal across experimental conditions. If class distributions are skewed however, (e.g. due to technical failures or processing steps), it is suggested to assess classifier performance by

AUC-values. Statistics for group level analyses are similar to accuracy. On the participant level however, permutation tests are again a must. The interested reader is directed to the introductory article by [182] for more information on AUC-values. The main disadvantage of F1-scores is that true negatives are neglected in their computation. Thus, F1-scores are known to be unreliable under certain circumstances [155]. In terms of statistical analyses, the same policy as for AUC-values applies.

3.5.3. Time Domain Classification

For single trial classification of time domain LPP data, it could be shown that classification of unpleasant vs. pleasant and pleasant vs. neutral was possible with accuracies and AUC-values above chance at group level. Data processing cascade was common to BCI practices. Fast feature reduction and selection based on R^2-values along with binary support vector machine classification yielded best results with 100 features and a linear kernel. However, classification only reached average accuracies of about 53 %, which are only significant at group level and not at participant-level. Thereby the application of machine learning methods merely serves as a confirmation that there are valence-related effects in the data, but that these effects are too small, so that the application for automatic affect recognition is not feasible with the presented approach.

In comparison with other studies, [167] conducted emotion induction by videos and also reported significant above chance level classification of EEG data regarding positive and negative valence. With an accuracy of 57.6 % they obtained results in a similar range as ours although a bit higher. However, these results are not directly comparable, as the classes were not evenly distributed, which stresses the importance of using measures like AUC to compare results with different class distributions across studies. In the present study, classification was also done solely in the time domain using the LPP while [167] used the power spectrum. As only validated features of neurophysiological emotional processing were classified, power spectra were not classified since our findings regarding inter-hemispheric frontal power difference related to emotional processing were not significant. Nevertheless, the classification performance in both studies is currently too low to be feasible for automatic affect recognition. This shows that besides better strategies for reporting and assessing classification performance, also better methods for EEG signal processing are needed to reduce the amount of noise in the data and improve affective classification.

3.6. Conclusion

Healthy

Neural responses to emotion-laden sounds were validated in the time- yet not in the frequency domain. The visually evoked LPP as a neurophysiological marker of emotional processing was investigated. Inter-hemispheric frontal differences in spectral power were not significant. Following a BCI processing cascade, classification results of LPP for valence were significantly above chance at group level.

Cerebral Palsy

Neural responses to emotion-laden sounds were not found to be significant, neither in the time nor in the frequency domain. Different positive amplitude deflections in response to stimuli categories could be observed, yet non-significant. On a group level, classification performance was significantly above chance for condition pleasant vs. neutral. Nonetheless, classified features could not be statistically validated. Furthermore, artifacts could not be fully excluded in the data recorded from this population due to the intrinsic difficulties of spastic and dyskinetic movements.

"We shall not cease from exploration, and the end of all our exploring will be to arrive where we started and know the place for the first time."

T. S. Eliot (1888 – 1965 CE)

4 Affect Classification in Infants

Infants account for an interesting and promising model for the study of emotion due to their purity (i.e. less cultural learning) in reactions to external world stimuli.

This chapter investigates the feasibility of affect recognition in an infant population. Therefore, emotional responses of infants induced by infant-parent interaction were recoded and classified by means of EEG.

Electrophysiological correlates of three affective states (i.e. unpleasant, neutral, and pleasant) will be investigated by analyzing data obtained with this novel paradigm. Specifically, frontal inter-hemispheric spectral differences related to emotional experiences will be validated. Significant power spectral features of the three affective states will be employed in a machine learning approach to discern the possibility of their automatic classification by means of EEG.

4.1. Participants

Twenty-eight healthy infant-parent pairs participated in the study which was approved by the Ethical Review Board of the Medical Faculty, University of Tübingen. Infants were between 4 to 6 months (13 female). Parents were informed about the purpose of the study and gave consent. The data of 3 subjects had to be excluded from data analysis due to movement artifacts. Thus, 25 datasets remained for analysis.

4.2. Design and Procedure

The design of a successful emotion induction paradigm for infants is a demanding endeavor. Most studies were based on standardized stimuli [74, 75, 76, 77]. However, the attentional focus of infants is quite limited. The focus on potentially non-interesting stimuli is thus endangered. An infant-parent interaction-based emotion induction paradigm was thus designed by Elaina Bolinger and colleagues at the Institute of Medical Psychology and Behavioural Neurobiology, Tübingen.
In an attempt to design an emotion induction paradigm suitable for infants, emotions are elicited by the natural interaction between an parent and their infant in different scenarios. Therefore, six different scenarios were defined: love, peek-a-boo, sing, jack-in-the-box, rash, and electrical outlet).

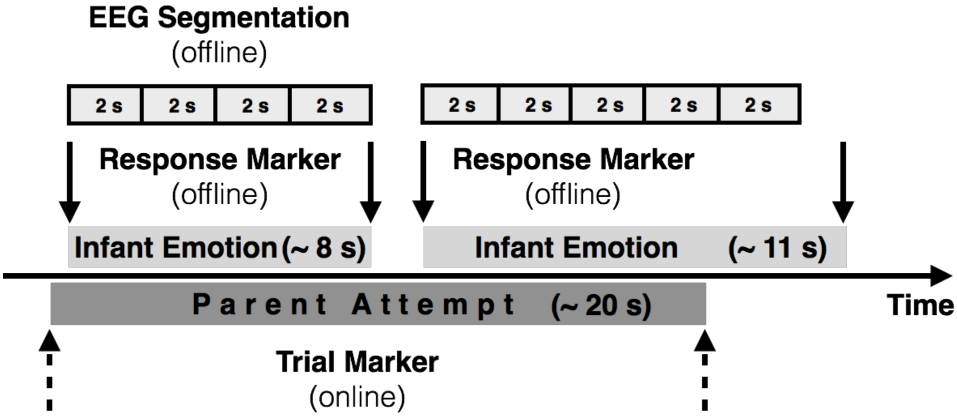


Figure 4.1.: Trial structure overview with online manually set trial markers along with offline set response markers (behavior scoring) and offline EEG segmentation. Parents enact selected scenario upon experimenter's cue called 'Parent Attempt'. Post-hoc offline scoring identifies emotionally relevant data within the larger trial called 'Response Marker'. Emotional responses are further segmented offline into 2 s long trial segments subjected to analysis.

For successful interactions, the parent first ensured the attentional focus of their infant. Subsequently, the different scenarios were carried out upon cues from the experimenter. In the love scenario, the parent expressed their strong positive affection towards their child. In peek-a-boo, the parent hid their face behind a writing pad and suddenly reappeared. In the sing scenario, the parent sang their or their infant's favorite song. In the jack-in-the-box scenario, the parent operated a musical box out of which a clown sprung after sufficient rotations of the mechanism. In the rash scenario, the parent expressed their strong concern about an imaginary rash on their infant's skin. In the electrical outlet scenario, the parent expressed fear caused by imagining that their infant would grasp an electrical outlet and thus warned their child.

Data without interaction served as baseline. Scenarios were grouped into three emotional categories of unpleasant, neutral, and pleasant experiences. Accordingly, the first four introduced scenarios (love, peek-a-boo, sing, and jack-in-the-box) are grouped as pleasant, while rash and electrical outlet scenarios are grouped as unpleasant. Baseline activity where no overt emotion was present is regarded as neutral.

Data were recorded in two sessions separated by one to four days. In each session, the parent enacted scenarios following the experimenter's cues. The duration of interaction varied naturally. The experimenter set relevant stimulus triggers manually and online. EEG and EOG were recorded simultaneously. EEG was recorded at electrode sites F3, Fz, F4, C3, C4, P3, Pz, P4 and Cz as the reference with special infant electrode cap (BrainProducts GmbH, Munich, Germany) at sampling frequency $F_s = 256$ Hz (SOMNOmedics GmbH, Randersacker, Germany). The topographical distribution of channels is shown in Figure 4.2 A. Furthermore, video recordings of infant-parent interactions were stored for later behavioral analysis. EEG data were subject to eye movement artifact correction [130]. Furthermore, trials that exceeded 100 μV were discarded.

An overview of the trial structure including online cues given by the experimenter for the parent to enact one of the scenarios (i.e. parent attempt) as well as the offline EEG segmentation based on behavioral scoring is shown in Figure 4.1.

Behavioral Analysis

Video recordings were post-hoc scored for behaviorally significant emotional expressions of the infants. Figure 4.1 shows the trial structure and post-hoc scoring in an example trial. The behavioral scoring was conducted by two independent experts at the Institute of Medical Psychology and Behavioural Neurobiology, Tübingen. Unambiguous criteria for in- or exclusion are depicted in italics. Other indicators carry some degree of ambiguity. Successfully identified emotional responses were segmented into 2 s long sub-trials with the same emotional label.

As for behavioral scoring, inclusion and exclusion criteria are depicted in Table 4.2. Unambiguous criteria for in- or exclusion are depicted in italics. Other indicators carry some degree of ambiguity.

Table 4.1.: Inclusion and exclusion criteria for successful emotional expression. Criteria written in italics constitute strong indicators of the respective condition. The rightmost column shows indicators for baseline activity.

Positive	Negative	Baseline
Eye contact	*Fussing*	No overt emotion
Smiling	*Crying*	Parent talking
Laughter/giggling	Avoidance (arching, gaze aversion)	
Cooing/babbling	Pushing away	
Approach (reaching)		
Happy dance		

4.3. Statistical Analysis

Since emotion expression was elicited by the naturalistic interaction with one of the infants' parents, stimulus triggers were variable. This suggests the analysis of the spectrum rather than event-related potentials (see Chapter 3).

Power spectra were computed by using autoregressive models, which were estimated using the maximum entropy method by [59]. A model order of 16 was used. The frequency range of 1 - 9 Hz was the main focus of analysis. For further analysis, the logarithm function was applied to the power spectra.

Inter-hemispheric frontal asymmetry was measured by subtracting the power at right-hemispheric electrode site F4 from left-hemispheric F3. Resulting numbers represent an index of the degree of lateralized activity. A more negative index indicates relative higher left-hemispheric power of EEG activity. A more positive index indicates relative higher right-hemispheric power. Differences between the three emotional conditions unpleasant, neutral, and pleasant were then tested for significance by a Wilcoxon test on each frequency bin of 1 Hz and corrected for multiple comparisons [176]. The same analysis is conducted for electrode-pairs C4 and C3, P4 and P3, as well as the average of spectra differences across conditions at F4, C4, P4 and F3, C3, P3.

The topographical distribution of power spectra in conditions unpleasant minus neutral as well as pleasant minus neutral is depicted in Figure 4.4. Obtained resolution is constrained by the 8 recorded EEG channels.

To estimate the separability of conditions pleasant and neutral in a machine learning approach, R^2-values are computed as shown in Figure 4.5.

4.4. Classification

Initially, to estimate the separability of valence conditions in a machine learning approach, coefficient of determination R^2-values were computed on a complete dataset of all subjects concatenated between the three binary classification problems: unpleasant vs. neutral, unpleasant vs. pleasant, and pleasant vs. neutral. Based on the correlation coefficient R, R^2 values are a measure for the proportion of variance in the dependent variable (i.e. EEG data) that is predictable from the independent variable (i.e. class). Classes in the complete dataset were balanced to the smallest amount of trials across conditions such that all classes remain with the same amount of trials. Balancing classes in machine learning ensures that estimated classifier performance is accurate. In order to estimate overall classification performance, classification was conducted on this complete dataset in a 10-fold cross-validation. Based on the results of this first analysis, cross-subject classification was conducted to estimate the generalization of features across subjects. Classification performance was assessed in a cross-subject approach with a SVM classifier using leave-one-subject-out-estimation (LOSOE). Therefore, a training model was computed based on $n-1$ datasets. Performance was then calculated using the model to predict labels of the n-th dataset which was excluded from training the model. This was conducted for all subjects. Datasets for training were balanced according to the smaller sized class. Therefore, trials from the larger sized class were randomly discarded until the number of trials was the same in both classes. The classification cascade is explained as follows.

Feature Selection & Extraction

Based on statistically significant differences (see Results section) between pleasant and neutral conditions of lateralized power spectra differences in the range of 1 to 9 Hz, binary classification was conducted using features of these frequency bands, i.e. power values between 1 to 9 Hz of all electrodes of one trial. The 100 best scoring features were then automatically selected with the fast feature selection method based on R^2-values [131].

Support Vector Machine

As classifier, a support vector machine (SVM) with a linear kernel (C = 1) using the libSVM implementation by [153] was employed. In its standard definition, the SVM is the solution of a geometric and data-driven minimization problem that finds a hyperplane best separating datapoints of two classes under certain conditions [147]. SVMs have been proven to be suitable for brain state classification especially in the field of BCI research due to their regularization property making them robust against the curse-of-dimensionality

[123]. In addition to the predicted labels, a probability estimate was obtained [152, 151].

Performance Measures

To assess classification performance, three measures were investigated: (i) classification accuracy, (ii) area under the curve (AUC) values, and (iii) F1-scores (see Figure 4.6). Please reger to Chapter 2.7.4 for a detailed description about performance measures.

Individual significance levels of classification performance at $p = 0.05$ were computed employing permutation tests [157]. For each dataset, classification performances were obtained in 100 iterations where in each iteration the label vector was randomly permuted. Performances were sorted in descending order and then the values at position 5, which equals the significance threshold at $p = 0.05$, were obtained as individual significance thresholds. Individual classification performances computed by original labels were significant, if they exceeded the obtained thresholds. Class ratios were computed as a ratio of the larger class relative to the total amount of samples in both classes combined.

4.5. Results

4.5.1. Behavioral Scoring & Number of Trials

Results of the behavioral scoring are shown in Table 4.5.1. For the different scenarios, the frequency of average attempts by the parent, the emotional scenario success rate, as well as the infant's average response lengths in seconds are depicted. The success rate indicates the amount of trials of identified valence during scenarios by two independent experts according to criteria shown in Table 4.2 in relation to the total amount of interaction attempts. The most successful scenario was the rash scenario with 82 % correct rate. The highest attempt frequency and the least successful scenario was jack-in-the-box with 19 % success rate. Response lengths varied between 8 seconds in the jack-in-a-box scenario to 26 seconds in the peek-a-boo scenario.

The distributions of unpleasant, neutral, and pleasant trials across all subjects for each condition is shown in Figure 4.2 B. After the rejection of artifacts, the following numbers of trials remained for each subject. For the unpleasant condition, there were on average 18.56 ± 14.81 trials. For the neutral condition, there were on average 96.44 ± 56.72 trials. Finally, the pleasant scenario led to 89.88 ± 67.61 trials.

Table 4.2.: Identification of valence by two independent experts including grouping ('+' pleasant; '-' unpleasant) with average emotional scenario attempt number of the acting parent. The success rate indicates the amount of trials of behaviorally identified valence during scenarios according to criteria shown in Table 4.2 in relation to the total amount of interaction attempts. Average response lengths of the infants' reactions is shown seconds.

Scenario	**Avg. Attempt No.**	**Success Rate**	**Avg. Response Length** $[s]$
Love (+)	8.67 ± 3.94	43 %	18.47 ± 14.37
Sing (+)	7.17 ± 3.36	27 %	15.95 ± 11.32
Jack-in-the-box (+)	15.00 ± 7.24	19 %	8.36 ± 3.38
Peek-a-boo (+)	7.43 ± 1.89	50 %	25.57 ± 22.49
Rash (-)	4.07 ± 1.46	82 %	15.85 ± 7.03
Electrical outlet (-)	3.23 ± 1.36	44 %	10.70 ± 5.57

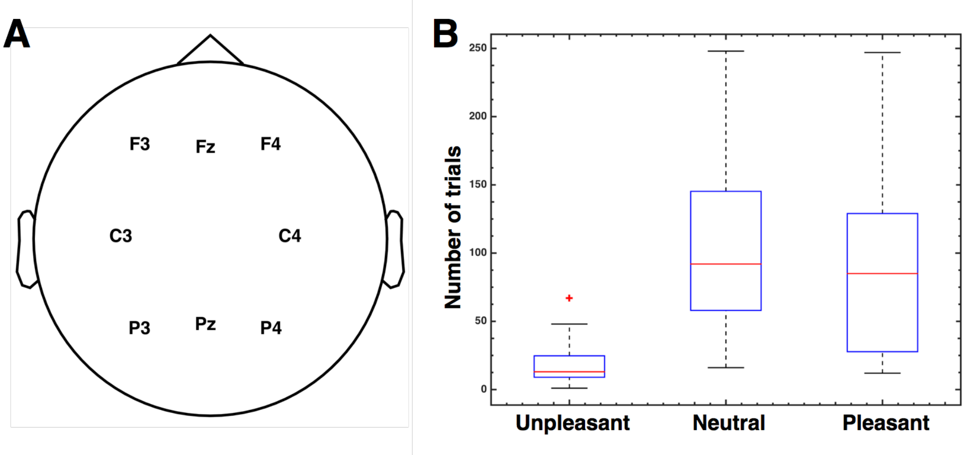


Figure 4.2.: (**A**) Topographical scheme of electrode locations. (**B**) Distributions of successful trials for emotional conditions unpleasant, neutral, and pleasant after discard of trials prone to artifacts.

4.5.2. Neurophysiological Measures

Spectral power analysis regarding a lateralization according to emotional experience is conducted in the range of 1 to 9 Hz. An index regarding the lateralization of power is computed by subtracting spectra of electrodes at opposite hemispheres. These are shown in Figure 4.3 for F4 and F3 (A), C4 and C3 (B), P4 and P3 (C), as well as average activity at F4, C4, P3 and F3, C3, P3 (D). Results regarding the lateralization index are shown in Figure 4.3. There was a significant difference in the lateralization index between pleasant and neutral conditions from 3 to 9 Hz in F4 minus F3 (Figure 4.3 A). According to the hypothesis, the lateralization index shows higher frontal left-hemispheric activity for pleasant than for unpleasant responses with increased activity on the right-hemisphere. At C4 minus C3 (Figure 4.3 B), differences between neutral and pleasant as well as unpleasant are significant between 3 to 6 Hz where the index is more negative for unpleasant than for pleasant responses indicating increased left-hemispheric activity during unpleasant responses more

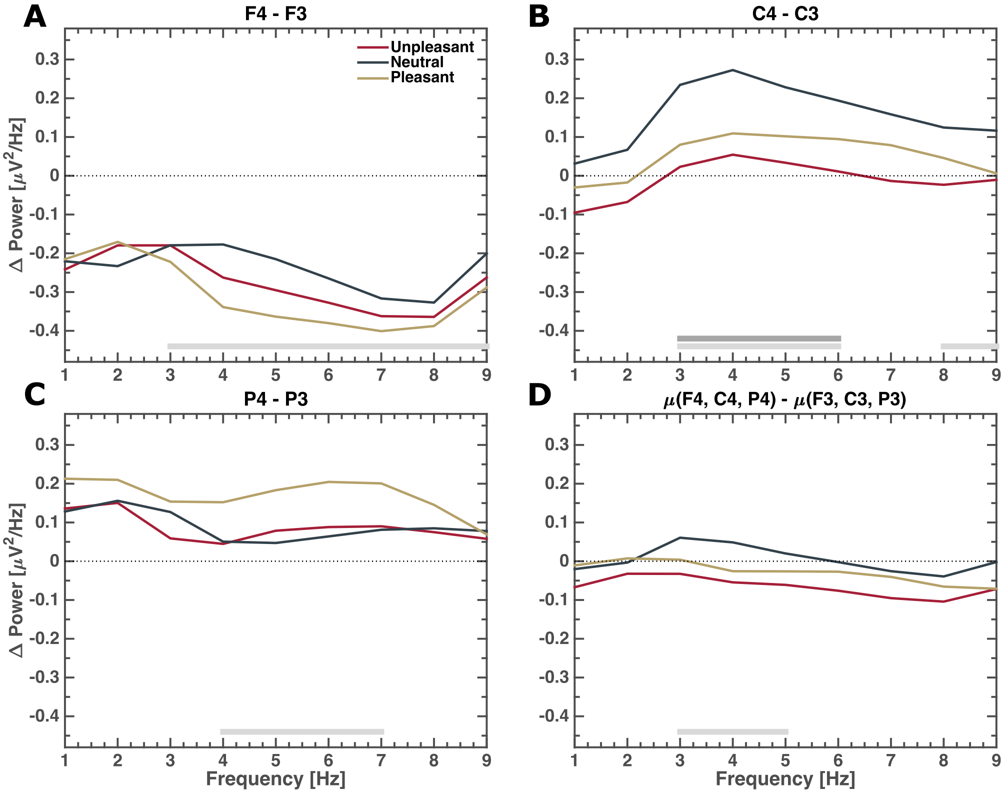


Figure 4.3.: Asymmetry comparison in power between different electrode locations and conditions across all subjects. Grey horizontal bars depict significant differences between neutral and pleasant (light grey) or neutral and unpleasant responses (dark grey) ($p < 0.01$, FDR corrected Wilcoxon test).

posterior. Furthermore, differences between neutral and pleasant conditions are significant between 8 to 9 Hz with similar lateralization. Parietal hemispheric differences in power for conditions pleasant and neutral are significant between 4 to 7 Hz where the index is also more negative for unpleasant than for pleasant responses (Figure 4.3 C) indicating again increased activity on the right-hemisphere. The averaged power of F4, C4, and P4 minus the averaged power of F3, C4, and P3 (Figure 4.3 D), differences between pleasant and neutral conditions are significant between 3 to 5 Hz. On average across hemispheres, the negative index indicates increased left-sided activity for unpleasant responses. Pleasant responses show no lateralized activity with values close to zero (1 - 3 Hz) and marginally negative values in higher frequencies (3.5 - 9 Hz). There is no significant difference of conditions neutral and pleasant on either channel pair. If not otherwise noted, there is no significant difference between unpleasant and neutral conditions.

The topographical distribution of power is schematically plotted in Figure 4.4 across gross frequency bands delta (1 - 4 Hz), theta (5 - 7 Hz), alpha (8 - 12 Hz), and beta (13 - 25 Hz). Figure 4.4 A shows the power distribution for unpleasant minus neutral, whilst the distribution of pleasant minus neutral is shown in the same figure in B. The resolution is suboptimal due to the number of channels, i.e. 8.

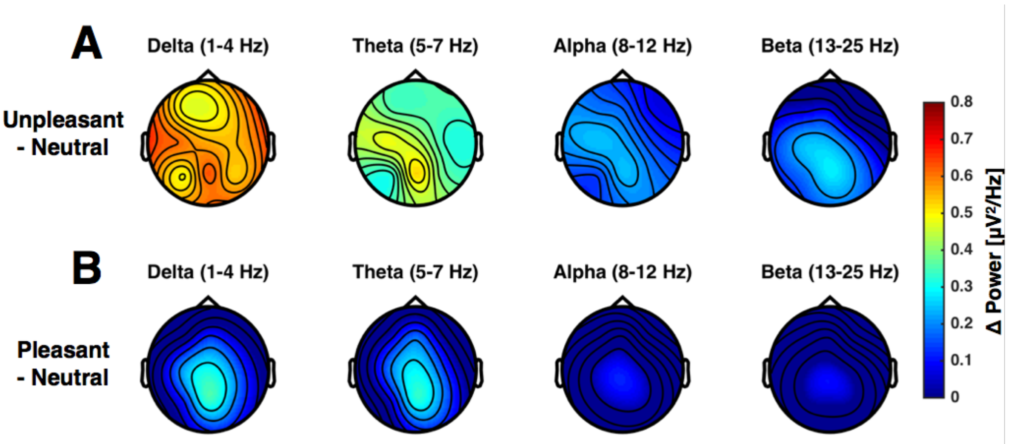


Figure 4.4.: Scalp topography plots of grand-average spectral differences for unpleasant minus neutral (top) and pleasant minus neutral (bottom) valence categories for different frequency bands.

To estimate class separability, R^2-values between valence conditions across all channels between 1 to 9 Hz were computed. R^2-values of unpleasant vs. neutral and unpleasant vs. pleasant were small on all channels without clear patterns (data not shown) indicating poor class separability by classification. The R^2-values for valence conditions pleasant vs. neutral are shown in Figure 4.5. Highest R^2-values are observed on Pz at 3 Hz. Also at Pz, R^2-values are high between 1 to 8 Hz as compared to the other channels.

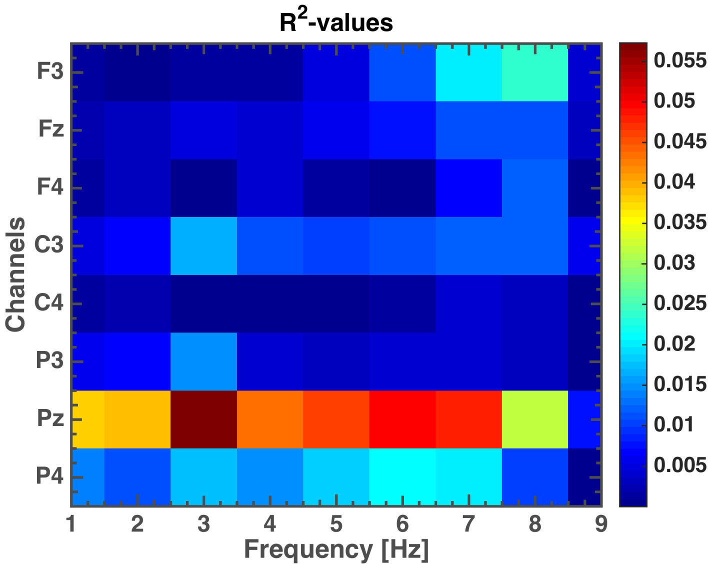


Figure 4.5.: R^2-values across all subjects per frequency and channels of conditions pleasant and neutral.

4.5.3. Classification

Classification performance in AUC-value obtained by 10-fold cross-validation on the complete dataset of all subjects for conditions unpleasant vs. neutral is 0.54, for unpleasant vs. pleasant it is 0.50, and for pleasant vs. neutral it is the best performance of 0.84. Class distributions were 464 trials for unpleasant, 2411 for neutral, and 2247 for pleasant. Classes were balanced to the smallest amount of trials, to have an even amount of trials. Table 4.3 depicts all performance measures in cross-subject classification results.

Table 4.3.: Classification performance on complete dataset with all subject data concatenated in three binary classification problems. Categories are abbreviated as follows: unpleasant ’-’, neutral ’0’, and pleasant ’+’. Classes are balanced.

	’-’ vs. ’0’	**’-’ vs. ’+’**	**’+’ vs. ’0’**
Accuracy	52.26 %	50,11 %	76,61 %
AUC-value	0.54	0.50	0.84
F1-score	0.12	0.00	0.76

Since classification of pleasant vs. neutral emotional conditions yielded the highest classification performance on the complete dataset, cross-subject classification has been further investigated for pleasant vs. neutral states. Table 4.4 depicts cross-subject classification results with their respective significance levels and class ratios. The classification of pleasant vs. neutral conditions in a cross-subject approach with LOSOE led to significantly above chance performances. The group average AUC-value is 0.65 ± 0.14.

Cross-subject classification performances, respective individual significance levels obtained by permutation tests at $p = 0.05$ are summarized in Figure 4.6. Individual significance levels of classification performance were exceeded by 16 subjects, 8 subjects performed below chance, and 1 subject performed above chance yet not above individual significance. Class ratios in the testing set were on average 66 ± 12 %.

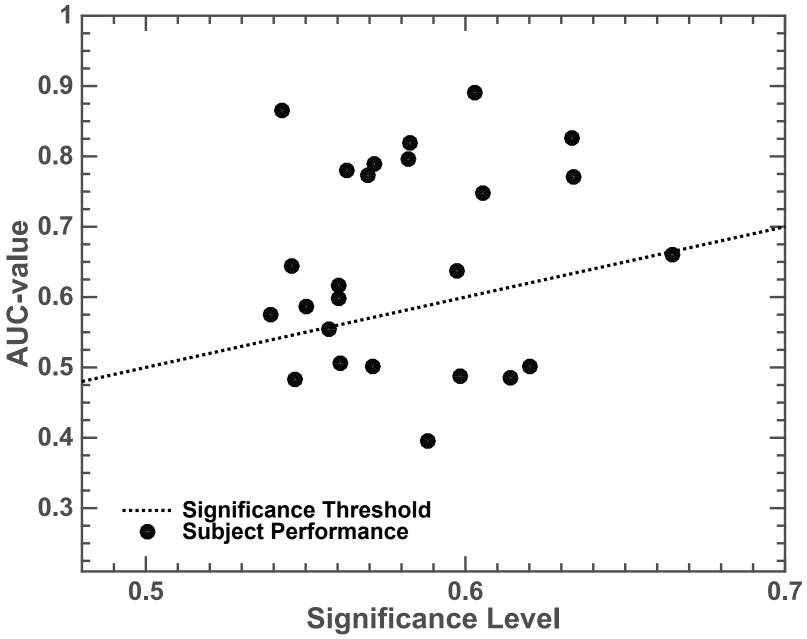


Figure 4.6.: Cross-subject classification performances in AUC-values against their respective significance levels. Each point represents performance of one subject obtained in leave-one-subject-out-estimation. The dotted line represents the significance threshold. Classes for training the classifier model were balanced to have an even amount of trials across conditions. Individual significance levels of classification performance were exceeded by 16 subjects, 8 subjects performed below chance, and 1 subject performed above chance yet not above individual significance.

Table 4.4.: Classification performance of AUC-values obtained by leave-one-subject-out-estimation for each dataset for the classification of pleasant vs. neutral conditions as well as class ratios within the testing set. An AUC-value of 1 means perfect classification where 0.5 is chance level. Bold values indicate significant AUC-values. The third column shows the AUC-value significance level at $p = 0.05$ obtained by permutation tests for each dataset. Ratios are larger classes divided by the total amount of trials in both classes combined. Training data were balanced to have an even amount of trials across conditions.

Participant	**AUC-value**	**Sig. Level**	**Class Ratios [%]**
01	**0.77**	0.63	63.64
02	**0.78**	0.56	52.42
03	**0.75**	0.61	71.60
04	**0.58**	0.54	58.78
05	0.49	0.60	60.55
06	**0.64**	0.55	61.34
07	0.50	0.62	57.81
08	0.48	0.55	53.78
09	0.66	0.66	52.94
10	**0.83**	0.63	92.77
11	**0.89**	0.60	89.02
12	0.55	0.56	62.77
13	0.50	0.57	84.32
14	**0.60**	0.56	58.02
15	**0.80**	0.58	66.00
16	**0.59**	0.55	66.49
17	**0.87**	0.54	69.19
18	**0.64**	0.60	75.89
19	**0.82**	0.58	55.77
20	**0.77**	0.57	59.35
21	0.50	0.56	57.84
22	0.48	0.61	75.25
23	**0.79**	0.57	74.49
24	0.40	0.59	84.53
25	**0.62**	0.56	56.44
Mean	**0.65** ± 0.14	0.58 ± 0.03	66.44 ± 11.69

4.6. Discussion

The present study investigated electrophysiological data of a naturalistic infant-parent interaction emotion induction paradigm recorded from preverbal infants in order to develop an automatic affect recognition system for sensory or mentally deprived caretakers and a child. Spectral power of infants' emotional responses across hemispheres showed significant differences in emotional valence (unpleasant, neutral, and pleasant). Based on these differences we could use machine learning methods to train a classifier that successfully discriminated emotional valence on unseen data.

4.6.1. Design and Procedure

The own parent is a realistic stimulus of vital importance, as compared to standardized audio-visual stimuli especially in an infant population. Particularly, motion [183] and the infants' attention towards faces [184] are of high relevance. Using an interaction-based approach, the degree of meaningful emotional experiences is substantially increased in such a screnario as opposed to standardized stimuli. Furthermore, the attentional focus of infants is ensured by the familiarity of the stimulus, i.e. the own parent, as compared to standardized stimuli delivered by a screen or speakers. At the same time, the well-defined nature of standardized stimuli approaches is missing. Trial numbers vary substantially due to manual triggering and success rates.

Employing an offline behavioral scoring scheme based on the analysis of video recordings, emotional segments were identified in the EEG. This approach ensured that EEG data consisted of meaningful emotional information. The amount of trials between unpleasant and pleasant conditions varies strongly due to ethical constraints regarding the induction of unpleasant emotional experiences in infants. Suitable emotion induction paradigms have to be designed to not interfere with the infant's well-being and at the same time to ensure a sufficient number of unpleasant emotional trials.

4.6.2. Neurophysiological Measures

The data shows increased frontal left-sided activation for pleasant than for unpleasant and higher frontal right-sided activation for unpleasant than for pleasant emotional responses in the theta (4 - 6 Hz) and alpha (6 - 9 Hz) band. These findings are in line with previous results in infants which indicate differences in frontal activation asymmetry between certain positive and negative emotions [74, 185], specifically in the theta and alpha band (up to 12 Hz) [76]. Neutral responses in comparison to unpleasant and pleasant showed the smallest left-sided activation in relation to valence. Power spectra of neutral and pleasant responses were significantly different in the theta and alpha band confirming values as features for classification. Over central and parietal electrodes, lateralization is opposite to frontal electrodes with increased right-sided activation for pleasant than for unpleasant responses in the delta, theta, and alpha band. Pleasant and neutral as well as unpleasant and neutral responses were significantly different in the theta band at central electrodes also suggesting power as feature for classification. Furthermore, significant differences between neutral and pleasant responses in the upper theta and low alpha band at parietal electrodes shows the usefulness of power as a feature for classification. Activity in those bands is relevant for emotional processing [78, 66] Furthermore, [186] reported an increase in theta during pleasurable stimulation in infants. The potential separability of neutral and pleasant emotional valence by spectral power features in classification is supported by positive R^2-values in the theta (electrode C3, P3, Pz, and P4) and alpha band (electrode F3, Fz, F4, C3, P3, Pz, P4). Since differences between unpleasant and neutral responses were only significant centrally in the theta band, and overall small R^2-values lacked clear patterns, the classification of unpleasant vs. neutral responses by spectral power in machine learning is questionable. The data do not suggest the classification of negative vs. positive valence by spectral power because there are no significant differences between unpleasant and pleasant conditions.

4.6.3. Classification of EEG

Classification on a complete balanced dataset of all subjects yielded excellent above chance classification performance in the binary classification of pleasant vs. neutral conditions (0.84 AUC-value) by spectral power features. As already suggested by the feature analysis, the classification of unpleasant vs. neutral conditions was slightly above chance (0..54 AUC-value) and the classification of unpleasant vs. pleasant was at chance (0.50 AUC-value). A within-subject classification was not feasible due to the small number of trials obtained. (Note: This is an important point since there has to be a critical amount of trials per class for successful classification. A cross-subject classification is always the stronger argument since the variance between subjects in classifier model is included.)

A within-subject was not feasible due to the paradigm design. In the offline pre-processing, emotionally relevant original trial segments were further partitioned into equally long sub-trials of 2 seconds. To conduct proper within-subject classification, the testing set must consist of sub-trials from one original trial not used for training the classifier model. This hard constraint for proper classification practice could not be maintained in the majority of datasets. Assuming that the dataset was not partitioned as outlined, the classifier model may operate on known data. Simply put, training and testing data overlap rendering any results futile. To give a general example: if an original trial consists of slow drifts or a baseline shift, segmented sub-trials also contain these drifts or shift. If a classifier is evaluated in a 10-fold cross-validation, the dataset containing slow drifts or a baseline shift is partitioned into training and test data. In this case the classifier model is likely to learn the slow drifts or baseline shift, rather than relevant features. Eventually during testing, the model recognizes slow drifts or baseline shift of sub-trials and outputs superior prediction performance due to the link of sub-trials in the training and test set. For a visualization, please see the procedure design and trial structure in Figure 4.1.

Thus, a cross-subject approach was chosen on pleasant vs. neutral conditions. Reported cross-subject classification performances are a strong argument for successful generalization of spectral features. With regard to the application of an affect recognition classifier, good cross-subject classification performance indicates the best strategy since a trained classifier potentially performs well on an unseen dataset.

There are no comparable numbers for this population, because affect recognition studies conducted in adults deviated in paradigm design. Studies employing standardized stimuli for emotion induction report between 55 % and 62 % accuracy in the classification between two emotional states [163, 167]. Those studies investigated small trial sizes in a within-subject approach which leads to methodological challenges in machine learning regarding chance levels and significance of results [156, 158]. The here presented cross-subject approach with LOSOE allowed for a sufficiently large trial size in training the classifier model. Class balancing and AUC-values ensure a chance level of 0.5 when assessing classification performance. In the emotional recall study by [164] which used also a non-standardized emotion induction paradigm, the authors report on average 80 % accuracy for the classification of two states (calm vs. positive) by time-frequency features and mutual information. However, the authors did not report a validation of the electrophysiological features. Their results are based on a within-subject approach. The variance between EEG datasets of subjects is substantial. Our cross-subject approach shows better generalization of spectral features, which we statistically validated before classification. Furthermore, permutation tests showed the significance of classification performances in 16 subjects. The code used for feature selection and machine learning has been successfully used for affect classification in EEG [18] and was made publicly available (https://github.com/dthettich/BSClassify).

In comparison to the auditory affect classification study conducted in healthy adults (see Chapter 3), affect classification in infant EEG between neutral and pleasant performs better. However, there was a significant above chance classification of unpleasant and neutral states in that study. The superiority of affect classification in infant EEG stems likely from the quality of emotional EEG due to infant-parent interaction. Less cultural learning in infants in conjunction with the realistic and vital parent stimulus account for the success of

emotional EEG data available for classification.
Physiological results as well as substantial above chance cross-subject classification performance for pleasant vs. neutral conditions attribute for the success of emotion classification in the data analyzed.
The results support the possibility of the construction of a simple, non-invasive automatic emotional valence classification system in a brain-computer interface (BCI) context [10, 118] even for very small children with the age between 4 and 6 months.

4.7. Conclusion

Neurophysiological data of preverbal infants recorded non-invasively during emotional interactions of infant-parent pairs show increased left-sided frontal activation for pleasurable and increased right-sided frontal activation for non-pleasurable stimulation. Significant differences between unpleasant and neutral as well as pleasant and neutral responses in the power spectra lateralization in the alpha and theta band confirms power spectra as features for classification. Successful classification of pleasant vs. neutral emotional responses in the EEG is demonstrated. In a cross-subject classification approach, on average AUC-values of 0.65 for the classification of pleasant vs. neutral responses were obtained employing power spectral features and a linear support vector machine classifier. The results of this realistic, everyday life emotion induction approach strengthens our vision of an automatic and non-invasive emotional detection possibility in very small infants. Particularly in social interactions between a child a a severely impaired or sensory deprived caretaker such a system may significantly improve quality of interaction and care.

5 DEAP Classification & Comparison

The relatively young field of affective computing lacks standardized datasets for the comparison of different methods. Such benchmark datasets are common in the field of brain-computer interfacing [187, 188] or the machine learning community, e.g. the MNIST dataset for hand writing recognition in computer vision [189]. Nonetheless, for the detection of human affective states such datasets were not available. Koelstra et al. [167] were one of the first to release a dataset comprising physiological signals for the study of human affective states.

Classification of low vs. high valence will be conducted based on EEG features. As a main focus, the influence of class sizes on classification performances will be outlined and discussed. Finally due to their paradigm setup similarity, machine learning results will be compared and discussed with results from the auditory affect induction and classification study outlined in Chapter 3.

5.1. Background

DEAP is a database for emotion analysis using physiological signals. It has been released by Koelstra et al. in 2012 [167]. The authors aimed at providing "a multimodal dataset for the analysis of human affective states" [167].

DEAP consists of data recorded from peripheral physiological signals including GSR and

ECG as well as EEG as a central physiological signal. These measures were recorded whilst participants watched forty one-minute excerpts from music videos. Participants rated their experiences in terms of arousal, valence, liking, dominance, and familiarity.

Only recently, Ringeval et al. [190] have released a challenge for the detection of physiological states from peripheral signals. However, the focus of the present work lies in the detection of affective states from the EEG. Therefore, the dataset by Koelstra et al. is well suited for the application of the methodology developed within the course of the present thesis.

5.2. Material & Methods

The following descriptions of material and methods (Sections 5.2.1 to 5.2.4.) are mainly after [167] with some supplementary information. A complete dataset description by the authors can be found at `http://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html` or of course in the article [167].

5.2.1. Participants

The dataset comprises 32 participant data (16 female). Participants were between 19 and 37 years of age (mean age 26.9 years). Prior to participation, participants gave informed consent and filled out a questionnaire.

5.2.2. Stimuli Selection

Initially, 120 music videos were selected as stimuli. One half was automatically and one half was manually selected. The automatic selection was based on music piece tags by a website offering music streaming. Therefore, emotional meaningful descriptive words were chosen (e.g. ’aggressive’ or ’depressing’) in order to retrieve music videos.

Subsequently, the valence-arousal space was subdivided into four quadrants of low arousal / low valence (LALV), low arousal / high valence (LAHV), high arousal / low valence (HALV), as well as high arousal / high valence (HAHV). For each quadrant, 15 videos were selected automatically and 15 manually, respectively, resulting in 120 videos.

In order to extract emotional meaningful information within these videos, the authors proposed an affective highlighting algorithm to determine 60-second excerpts. Stimuli were then further curated via a web-based subjective assessment. Final stimuli selection was conducted such that only videos were selected that lie in the outermost corner of each quadrant. That resulted in 40 music videos with 60 seconds each.

5.2.3. Paradigm Design

Participants were seated in front of 17"-inch screen 1 meter away. The experiment started with a two minute fixation baseline. Subsequently, the 40 music videos were presented with the following additional steps. Firstly, the current trial number was displayed for two seconds. Secondly, a five second fixation baseline was recorded. Thirdly, the 60 second music video was presented. Fourthly, the self-assessment for arousal, valence, liking, and dominance was obtained (see [167] Figure 5). Valence, arousal, and dominance rating were obtained by a Likert-like scale from 1 to 9 as in the auditory affect induction study (see Chapter 4). For liking, three solutions, thumb-down, thumb-horizontal, and thumb-up were shown. After 20 trials, participants were allowed a short break. Following the break, electrode placement and proper conductivity were checked. Participants then completed the second 20 videos with equal steps.

In the following work presented here, participants' valence ratings were employed as labels for classification based on threshold values. The lower threshold was 3.825 and the upper threshold was set at 5.95 as derived from quantiles. For each participant, trials were divided by these thresholds into low and high valence trials. For comparison, this approach was similar to the one introduced in the auditory affect induction and classification study with unpleasant or pleasant trials (see Chapter 3).

5.2.4. Setup & Pre-processing

The experiments were performed under controlled conditions. Physiological signals were recorded with a Biosemi ActiveTwo system (`http://www.biosemi.com/`). The EEG was recorded at a sampling rate of 512 Hz using 32 active AgCl electrodes placed according to the international 10-20 system [49]. Electrode sites were Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, P4, P8, PO4, and O2. Thirteen peripheral physiological signals were also recorded. These will not be outlined further for peripheral physiological signals are not within the scope of the present work.

Out of the box, the DEAP dataset made available was further pre-processed in its MATLAB variant by the providers of DEAP. Therefore, the authors downsampled the data to 128 Hz, EOG artifacts were removed as in [167], a bandpass filter from 4.0 to 45.0 Hz was applied, and the data were averaged to the common reference. Trials consisted of 60 seconds where a 3 second pre-trial relative to stimulus-onset baseline was removed.

5.2.5. Classification

The following is own work and based on the classification apparatus developed. Classification of low vs. high valence was evaluated based on time- and frequency domain features employing a setup similar to the one described in Chapter 3, Section 3.3.2.

Feature Selection & Extraction

Although [167] did not report affect related effects in the time domain, features in the time domain were extracted from channels Pz, Cz, Cp2, Cp6, Cp5, and Cp1 from 0 s to 6 s relative to stimulus-onset. For comparison, channels and time frame are chosen due to the results reported in the auditory affect induction and classification study (see Chapter 3).

[167] reported significant affect related effects in the frequency domain and also employed those as EEG features for classification between low and high valence. Therefore, frequency features were also computed from the 60 s trials by the method of Burg [59] in 1 Hz frequency bins from 1 to 50 Hz with a model order of 8. Frequency features were taken from all channels.

To reduce the number of features, R^2-values between data and labels were computed for each feature and the features with the highest R^2-values were used for classification [131]. Only the 100 best scoring features were retained for training the classifier model and predictions.

Support Vector Machine

As classifier, a support vector machine (SVM) with a linear kernel (C=1) using the libSVM implementation [153] was employed. SVMs have been proven to be well suitable for brain state classification especially in the field of BCI research [123]. Label predictions as well as prediction probabilities [152, 151] were obtained. Due to the relatively small number of instances per class, a 5-fold cross-validation was employed to compute all performance measures.

Performance Measures

To assess classification performance, three measures were investigated: (i) classification accuracy, (ii) area under the curve (AUC) values, and (iii) F1-scores.

5.3. Results

Classification results of low vs. high valence for unbalanced as well as for balanced classes are depicted in the following. Firstly, classification performances employing time domain features of channels Pz, Cz, Cp2, Cp6, Cp5, and Cp1 in the first six seconds of stimulus-onset are shown in Figure 5.1. Secondly, classification of frequency domain features from 1 to 50 Hz of all channels is reported. Furthermore for each domain, the differences in classification performance between non-balanced classes and balanced classes are reported.

5.3.1. Time Domain Classification

The time domain classification of low vs. high valence for unbalanced as well as for balanced classes is shown in Figure 5.1.

Accuracy is shown in Figure 5.1 A, AUC-value in Figure 5.1 B, and F1-score in Figure 5.1 C. The group average and standard deviation for each metric is shown in Table 5.2. Class ratios are depicted in Figure 5.1 D. The mean of class ratios measured by the bigger sized class is 55.06 ± 25.44 %. In terms of absolute numbers, there are on average 9.91 ± 4.53 trials in class one and 17.53 ± 3.46 trials in class two. Class distributions are significantly different (two-tailed t-test, $p < 0.001$). A complete overview of number of trials per class and dataset is shown in Table 5.1.

Table 5.1.: DEAP number of trials for each dataset after thresholding into low and high valence. Thresholds were set at 3.825 and 5.95 for the separation.

Id.	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
low val.	17	9	4	20	8	4	9	8	8	10	11	15	12	16	10	17	7	3	12	7	8	15	4	11	15	12	5	13	12	3	7	5
high val.	18	20	16	14	22	24	25	15	13	13	16	17	13	19	17	11	18	17	16	16	18	15	19	16	15	24	22	22	20	15	18	17

Group average classification performances for time domain features are shown in Table 5.2 for the measures accuracy, AUC-value, and F1-score. Differences between non-balanced and balanced classification performances are significant for accuracy and F1-score (two-tailed t-test, $p < 0.01$ corrected for multiple comparisons).

If a chance level of 50 % for accuracy is assumed, a two-tailed t-test ($p < 0.01$) yields significant group average classification for this performance measure. However, the chance level is not at set 50 % due to non-balanced class sizes. In this case, permutation tests are necessary to correctly estimate a chance level.

Table 5.3 depicts average classification chance levels for non-balanced and balanced

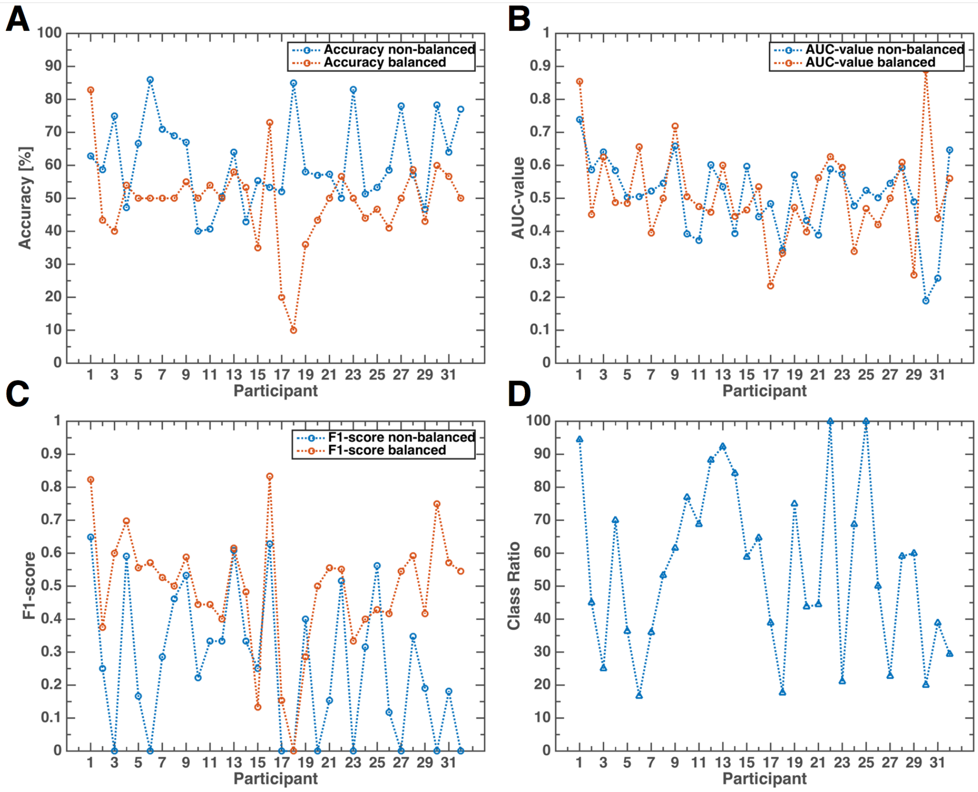


Figure 5.1.: Time domain classification performances of DEAP dataset participants for performance measures accuracy (**A**), AUC-value (**B**), and F1-score (**C**). Class ratios in percent (**D**).

Table 5.2.: Time domain average classification performances for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score.

	Non-balanced	**Balanced**
Accuracy	61.15 ± 12.93 %	48.89 ± 13.02 %
AUC-value	0.51 ± 0.12	0.51 ± 0.14
F1-score	0.26 ± 0.22	0.49 ± 0.18

classes obtained by 100 permutation tests ($p = 0.5$). Permutation results at $p = 0.5$ equal the average group chance level. Chance levels obtained for non-balanced and balanced classes are significantly different for accuracy (two-tailed t-test, $p < 0.01$) and F1-score (two-tailed t-test, $p < 0.0001$).

Table 5.3.: Time domain average classification chance levels obtained by permutation tests ($p = 0.5$) for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score.

	Non-balanced	**Balanced**
Accuracy	62.31 ± 13.99 %	49.52 ± 12.95 %
AUC-value	0.48 ± 0.12	0.46 ± 0.16
F1-score	0.20 ± 0.23	0.52 ± 0.17

Table 5.4 contrasts classification results of each participant for non-balanced and balanced classes against their respective individual significance level ($p = 0.05$) for time domain features. Bold values of individual classification performances indicate when they exceed individual significance levels.

For non-balanced classes, two datasets exceed significance in accuracy, nine in AUC-value, and five in F1-score. Notably, only dataset 19 yielded significance in accuracy and AUC-value, yet not in F1-score for the non-balanced classes. For balanced classes, one dataset yielded significance in accuracy, nine in AUC-value, and one in F1-score. The intersection where classification performances exceed significance across non-balanced and balanced classes in the same measure yields three cases (i.e. datasets 01, 15, 22). Only dataset 01 exceeded significance in all three performance measures in the balanced classes classification.

Differences between individual significance thresholds are significant between non-balanced and balanced classes for the measures AUC-value and F1-score (two-tailed t-test, $p < 0.01$).

A table in similar layout yet with individual classification performances against chance level ($p = 0.5$) can be found in the Appendix, Table C.1.

Figure 5.2 shows line plots of individual classification performance (depicted in blue) in contrast to their according significance thresholds (depicted in red). Figures 5.2 A and B show these results for non-balanced and balanced classes in accuracy, Figures 5.2 B and E for AUC-value, and Figures 5.2 C and F for F1-score.

Accuracy and AUC-value significance thresholds mainly are between 50 % to 85 %, and 0.0 to 0.8, respectively for non-balanced and balanced classes. F1-score significance thresholds in the non-balanced condition (Figure 5.4 C) oscillate between 0 and 0.7. In contrast in the balanced condition, F1-scores of significance oscillate comparable to AUC-values.

Evaluating the correlation between individual performance measures. In the non-balanced condition, the correlation between individual accuracy, AUC-value, and F1-score is only significant for accuracy and F1-scores, however negatively correlated ($r = -0.48$, $p < 0.01$). In the balanced condition, the correlation between all performance measures is significant ($p < 0.001$). Accuracy and AUC-value correlate with $r_1 = 0.66$, Accuracy and F1-score with $r_2 = 0.89$, and finally AUC-value with F1-score with $r = 0.66$.

Table 5.4.: Time domain average classification performances for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score, as well as their corresponding individual significance thresholds ($p = 0.05$). Values for accuracy are given in percent. Bold values indicate when individual performance exceed the individual significance level.

	Non-balanced						**Balanced**					
	Accuracy [%]		**AUC-value**		**F1-score**		**Accuracy [%]**		**AUC-value**		**F1-score**	
Id.	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**
01	62.86	68.57	**0.74**	0.68	0.65	0.67	**82.86**	68.10	**0.85**	0.72	**0.82**	0.67
02	58.67	72.67	0.59	0.77	**0.25**	0.20	43.33	80.00	0.45	0.86	0.38	0.80
03	75.00	80.00	0.64	0.81	0.00	0.00	40.00	80.00	0.62	0.81	0.60	0.89
04	47.14	68.57	0.58	0.69	0.59	0.77	54.00	75.33	0.49	0.63	0.70	0.85
05	66.67	73.33	0.50	0.52	**0.17**	0.00	50.00	63.33	0.48	0.64	0.56	0.67
06	86.00	86.00	0.51	0.59	0.00	0.00	50.00	70.00	**0.66**	0.62	0.57	0.67
07	70.95	76.67	0.52	0.71	**0.29**	0.20	50.00	50.00	**0.40**	0.35	0.53	0.62
08	69.00	70.00	**0.55**	0.50	**0.46**	0.33	50.00	50.00	0.50	0.51	0.50	0.53
09	67.00	77.00	0.66	0.86	0.53	0.62	55.00	80.00	0.72	0.79	0.59	0.80
10	40.00	65.00	0.39	0.63	0.22	0.60	50.00	60.00	0.51	0.59	0.44	0.64
11	40.67	63.33	0.37	0.62	0.33	0.38	54.00	64.00	0.48	0.69	0.44	0.61
12	50.48	62.86	**0.60**	0.60	0.33	0.54	50.00	63.33	0.46	0.59	0.40	0.54
13	**64.00**	60.00	0.54	0.58	0.61	0.65	58.00	63.00	0.60	0.64	0.62	0.71
14	42.86	68.57	0.39	0.73	0.33	0.67	53.33	72.86	0.45	0.77	0.48	0.76
15	55.33	56.00	**0.60**	0.40	0.25	0.32	35.00	50.00	**0.46**	0.43	0.13	0.64
16	53.33	64.67	0.44	0.53	0.63	0.77	73.00	77.00	**0.54**	0.50	0.83	0.87
17	52.00	76.00	0.48	0.72	0.00	0.25	20.00	80.00	0.23	0.85	0.15	0.80
18	85.00	85.00	0.34	0.36	0.00	0.00	10.00	30.00	**0.33**	0.22	0.00	0.50
19	**58.00**	53.33	**0.57**	0.51	0.40	0.43	36.00	58.00	0.47	0.54	0.29	0.62
20	57.00	70.00	0.43	0.63	0.00	0.00	43.33	60.00	0.40	0.57	0.50	0.53
21	57.33	73.33	0.39	0.74	0.15	0.22	50.00	73.33	0.56	0.77	0.56	0.75
22	50.00	50.00	**0.59**	0.49	0.52	0.56	56.67	56.67	**0.63**	0.53	0.55	0.61
23	83.00	83.00	0.57	0.75	0.00	0.00	50.00	60.00	**0.59**	0.50	0.33	0.60
24	51.33	60.00	**0.48**	0.47	0.32	0.42	44.00	51.00	0.34	0.45	0.40	0.42
25	53.33	53.33	**0.52**	0.49	**0.56**	0.55	46.67	53.33	0.47	0.50	0.43	0.56
26	58.57	64.29	0.50	0.51	0.12	0.32	41.00	62.00	0.42	0.66	0.42	0.67
27	78.00	81.33	0.55	0.63	0.00	0.00	50.00	80.00	0.50	0.72	0.55	0.80
28	57.14	62.86	0.59	0.62	0.35	0.38	58.67	77.33	0.61	0.85	0.59	0.81
29	46.67	62.86	0.49	0.54	0.19	0.42	43.00	54.00	0.27	0.53	0.42	0.52
30	78.33	85.00	0.19	0.69	0.00	0.00	60.00	60.00	**0.89**	0.39	0.75	0.75
31	64.00	76.00	0.26	0.71	0.18	0.25	56.67	63.33	0.44	0.64	0.57	0.62
32	77.00	77.00	**0.65**	0.51	0.00	0.00	50.00	80.00	0.56	0.82	0.55	0.75
Mean	61.15	69.58	0.51	0.61	0.26	0.33	48.89	64.56	0.51	0.62	0.49	0.67

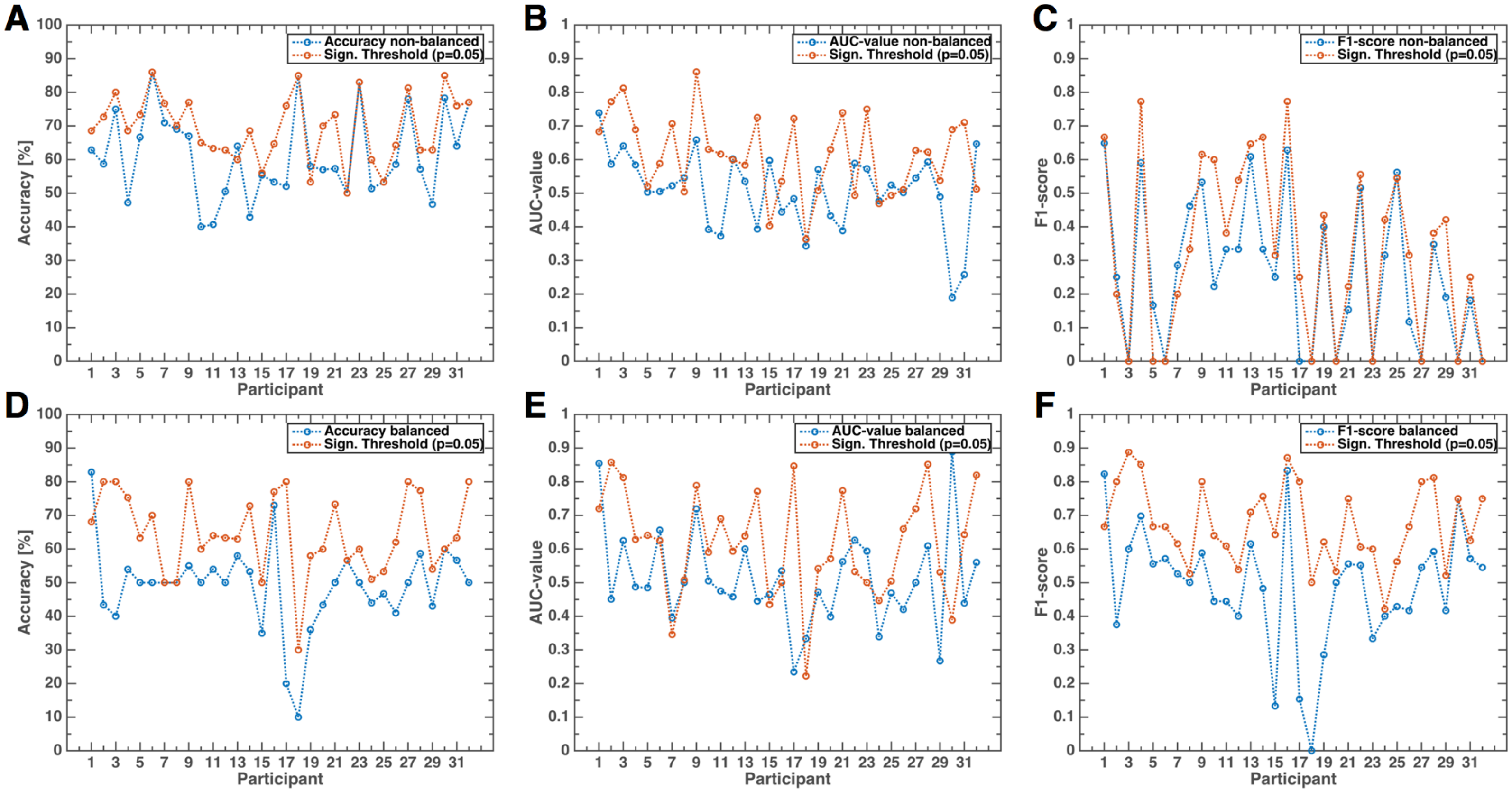


Figure 5.2.: Time domain classification performance of DEAP dataset participants and their significance levels ($p = 0.05$) obtained in permutation test. The top row shows both measures for non-balanced classes in accuracy (**A**), AUC-value (**B**), and F1-score (**C**). The bottom row shows measures for balanced classes in accuracy (**D**), AUC-value (**E**), and F1-score (**F**).

5.3.2. Frequency Domain Classification

The frequency domain classification of low vs. high valence for unbalanced as well as for balanced classes is shown in Figure 5.3.

Accuracy is shown in Figure 5.3 A, AUC-value in Figure 5.1 B, and F1-score in Figure 5.1 C. The group average and standard deviation for each metric is shown in Table 5.2. Class ratios are depicted in Figure 5.3 D. The distributions of classes for non-balanced data are the same as mentioned before.

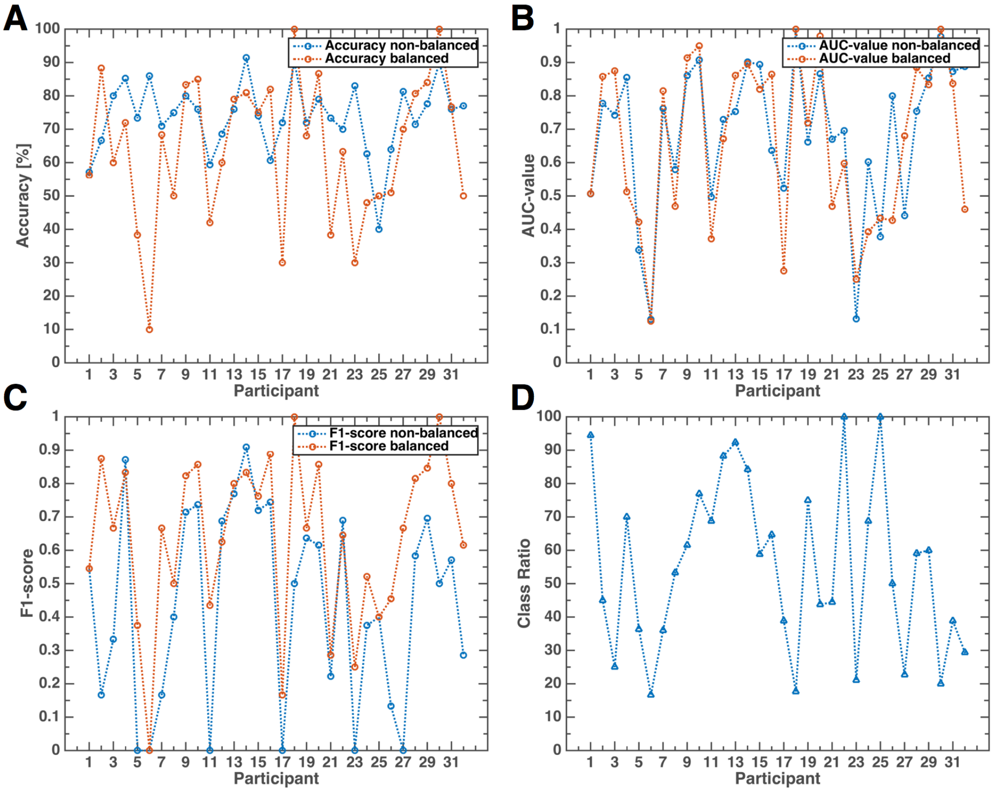


Figure 5.3.: Frequency domain classification performances of DEAP dataset participants for performance measures accuracy (**A**), AUC-value (**B**), and F1-score (**C**). Class ratios in percent (**D**).

Group average classification performances are shown in Table 5.5 for the measures accuracy, AUC-values, and F1-score. Group averages are significantly above chance for balanced classes across all performance measures (right-tailed t-test, $p < 0.01$). Differences between non-balanced and balanced classification performances are significant for accuracy and F1-score (two-tailed t-test, $p < 0.01$ corrected for multiple comparisons).

Table 5.6 depicts average classification chance levels for non-balanced and balanced classes obtained by 100 permutation tests ($p = 0.5$). Permutation results at $p = 0.5$ equal

Table 5.5.: Frequency domain average classification performances for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score.

	Non-balanced	**Balanced**
Accuracy	73.74 ± 10.68 %	64.29 ± 21.68 %
AUC-value	0.69 ± 0.22	0.66 ± 0.25
F1-score	0.44 ± 0.29	0.64 ± 0.25

the average group chance level. Chance levels obtained for non-balanced and balanced classes are significantly different for accuracy (two-tailed t-test, $p < 0.001$) and F1-score (two-tailed t-test, $p < 0.0001$).

The expected chance level in the balanced case is 50 %, 0.5, and 0.5 for the respective performance measure. Computed chance levels from permutation tests ($p = 0.05$) are not significantly different than the ones expected (two-tailed t-test, $p < 0.001$) for frequency domain features. It is to note that the AUC-value chance level is 0.45 for the non-balanced and balanced condition with a standard deviation of approximately 0.05.

Table 5.6.: Frequency domain average classification chance levels obtained by permutation tests ($p = 0.5$) for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score.

	Non-balanced	**Balanced**
Accuracy	64.67 ± 12.00 %	47.70 ± 8.95 %
AUC-value	0.45 ± 0.04	0.45 ± 0.05
F1-score	0.15 ± 0.21	0.48 ± 0.11

Table 5.7 contrasts classification results of each participant for non-balanced and balanced classes against their respective individual significance level ($p = 0.05$) for frequency domain features. Bold values of individual classification performances indicate when they exceed individual significance levels.

For non-balanced classes, fifteen datasets exceed significance in accuracy, 19 in AUC-value, and fourteen in F1-score. Notably, twelve datasets (i.e. 04, 09, 10, 11, 13, 14, 15, 19, 20, 28, 29, and 30) exceeded individual significance in all three performance measures.

For balanced classes, ten datasets yielded significance in accuracy, twelve in AUC-value, and ten in F1-score. Nine datasets (i.e. 02, 09, 10, 13, 14, 20, 28, 29, and 30) exceed significance in all three performance measures.

The intersection of datasets where classification performances exceed significance across non-balanced and balanced classes in all measures is 09, 10, 13, 14, 20, 28, and 29. The average number of trials of these datasets is 11.14 in class ’low valence’ and 16.57 in class ’high valence’. Thus, approximately 21 trials were used for classifier training and 5 for testing in the non-balanced case. For balanced classes, these number amount to 18 for training and 4 for testing. On average for non-balanced classes,

Differences between individual significance thresholds are significant between non-balanced and balanced classes for the measures AUC-value and F1-score (two-tailed t-test, $p < 0.001$).

A table in similar layout yet with individual classification performances against chance level ($p = 0.5$) can be found in the Appendix, Table C.2.

Figure 5.4 shows line plots of individual classification performance (depicted in blue) in contrast to their according significance thresholds (depicted in red). Figures 5.4 A and B show these results for non-balanced and balanced classes in accuracy, Figures 5.4 B and E for AUC-value, and Figures 5.4 C and F for F1-score.

Accuracy and AUC-value significance thresholds mainly oscillate between 60 % to 90 %, and 0.65 to 0.9, respectively for non-balanced and balanced classes. F1-score significance thresholds in the non-balanced condition (Figure 5.4 C) oscillate between 0 and 0.7. In contrast in the balanced condition, F1-scores of significance oscillate comparable to AUC-values.

Evaluating the correlation between individual performance measures. In the non-balanced condition, the correlation between individual accuracy, AUC-value, and F1-score is only significant for AUC-values and F1-scores ($r = 0.64$, $p < 0.001$). In the balanced condition, the correlation between all performance measures is significant ($p < 0.0001$). Accuracy and AUC-value correlate with $r_1 = 0.94$, Accuracy and F1-score with $r_2 = 0.98$, and finally AUC-value with F1-score with $r = 0.92$.

Table 5.7.: Frequency domain average classification performances for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score, as well as their corresponding individual significance thresholds ($p = 0.05$). Values for accuracy are given in percent. Bold values indicate when individual performance exceed the individual significance level.

	Non-balanced						**Balanced**					
	Accuracy [%]		**AUC-value**		**F1-score**		**Accuracy [%]**		**AUC-value**		**F1-score**	
Id.	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**
01	57.14	65.71	0.51	0.70	0.55	0.62	56.19	70.48	0.51	0.74	0.55	0.71
02	66.67	72.67	**0.78**	0.74	0.17	0.31	**88.33**	83.33	**0.86**	0.84	**0.88**	0.82
03	80.00	85.00	0.74	0.84	0.33	0.40	60.00	80.00	0.88	0.88	0.67	0.80
04	**85.24**	64.76	**0.86**	0.64	**0.87**	0.75	72.00	72.00	0.51	0.69	0.83	0.83
05	73.33	76.67	0.34	0.70	0.00	0.22	38.33	68.33	0.42	0.69	0.38	0.71
06	86.00	86.00	0.13	0.64	0.00	0.00	10.00	70.00	0.12	0.75	0.00	0.67
07	70.95	76.67	**0.76**	0.74	0.17	0.20	68.33	71.67	**0.81**	0.78	0.67	0.74
08	**75.00**	70.00	0.58	0.68	**0.40**	0.36	50.00	63.33	0.47	0.66	0.50	0.67
09	**80.00**	67.00	**0.86**	0.62	**0.71**	0.36	**83.33**	70.00	**0.91**	0.77	**0.82**	0.71
10	**76.00**	69.00	**0.91**	0.70	**0.74**	0.50	**85.00**	75.00	**0.95**	0.76	**0.86**	0.74
11	59.33	66.67	0.50	0.75	0.00	0.43	42.00	69.00	0.37	0.74	0.43	0.70
12	**68.57**	67.62	**0.73**	0.70	**0.69**	0.64	60.00	70.00	0.67	0.70	0.62	0.67
13	**76.00**	68.00	**0.75**	0.71	**0.77**	0.69	**79.00**	71.00	**0.86**	0.71	**0.80**	0.69
14	**91.43**	62.86	**0.90**	0.68	**0.91**	0.52	**80.95**	64.76	**0.90**	0.69	**0.83**	0.65
15	**74.00**	67.33	**0.89**	0.70	**0.72**	0.35	75.00	80.00	0.82	0.84	0.76	0.80
16	60.67	64.67	**0.64**	0.64	0.74	0.77	82.00	82.00	**0.86**	0.78	0.89	0.89
17	72.00	72.00	0.52	0.65	0.00	0.25	30.00	70.00	0.28	0.76	0.17	0.71
18	90.00	90.00	**1.00**	0.86	0.50	0.50	100.00	100.00	1.00	1.00	1.00	1.00
19	**72.00**	65.33	**0.66**	0.65	**0.64**	0.44	**68.00**	67.00	**0.72**	0.68	0.67	0.67
20	**79.00**	74.00	**0.87**	0.71	**0.62**	0.36	**86.67**	76.67	**0.98**	0.76	**0.86**	0.77
21	**73.33**	73.33	0.67	0.68	0.22	0.22	38.33	76.67	0.47	0.82	0.29	0.78
22	**70.00**	70.00	0.70	0.70	0.69	0.69	63.33	66.67	0.60	0.69	0.65	0.67
23	83.00	87.00	0.13	0.74	0.00	0.40	30.00	90.00	0.25	1.00	0.25	0.86
24	62.67	66.67	0.60	0.66	0.38	0.40	48.00	68.00	0.39	0.66	0.52	0.69
25	40.00	70.00	0.38	0.71	0.40	0.69	50.00	70.00	0.43	0.68	0.40	0.69
26	63.93	71.79	**0.80**	0.66	0.13	0.29	51.00	63.00	0.43	0.66	0.45	0.64
27	81.33	81.33	0.44	0.72	0.00	0.00	70.00	70.00	0.68	0.78	0.67	0.73
28	**71.43**	65.71	**0.75**	0.62	**0.58**	0.32	**80.67**	73.33	**0.88**	0.79	**0.81**	0.74
29	**77.62**	69.52	**0.85**	0.71	**0.70**	0.50	**84.00**	71.00	**0.83**	0.69	**0.85**	0.72
30	**90.00**	85.00	**0.98**	0.78	**0.50**	0.40	**100.00**	80.00	**1.00**	0.89	**1.00**	0.80
31	76.00	76.00	**0.87**	0.73	**0.57**	0.33	76.67	80.00	0.84	0.84	**0.80**	0.77
32	77.00	77.00	**0.89**	0.80	0.29	0.33	50.00	80.00	0.46	0.84	0.62	0.80
Mean	73.74	72.67	0.69	0.70	0.44	0.41	64.29	73.85	0.66	0.77	0.64	0.74

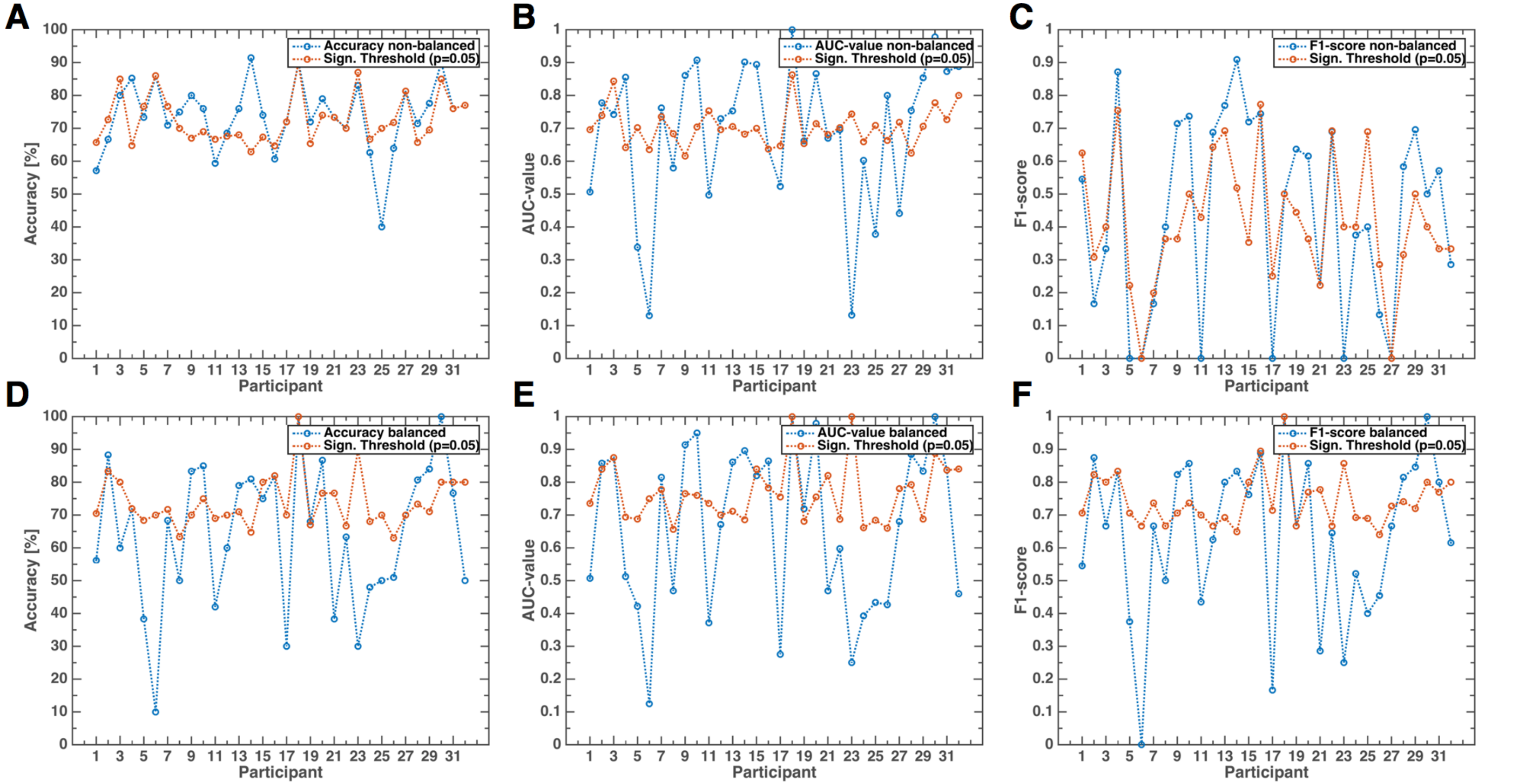


Figure 5.4.: Frequency domain classification performance of DEAP dataset participants and their significance levels ($p = 0.05$) obtained in permutation test. The top row shows both measures for non-balanced classes in accuracy (**A**), AUC-value (**B**), and F1-score (**C**). The bottom row shows measures for balanced classes in accuracy (**D**), AUC-value (**E**), and F1-score (**F**).

5.4. Discussion

The DEAP dataset released by Koelstra et al. [167] is a dataset for the study of affective states in central and peripheral physiological signals. Time domain and frequency domain features of two affective states, i.e. low and high valence, of the central nervous system were classified using a support vector machine classifier. Results were investigated for differences of non-balanced and balanced classes in three performance measures. Performances were significantly different between class conditions.

Group averages of classified frequency domain features were significantly above chance for all performance measures in balanced classes (64.29 % accuracy, 0.66 AUC-value, and 0.64 F1-score). Koelstra et al. reported 57.60 % in group average accuracy and 0.56 in F1-score when non-balanced classes were classified in Fisher's linear discriminant analysis.

Although trial sizes were small, the results show successful discrimination between low and high valence using the machine learning pipeline introduced with EEG frequency domain features.

5.4.1. Time Domain Classification

According to the expectation for balanced classes, group averages for time domain classification were approximately at chance levels (accuracy: 48.89 ± 13.02 %, AUC-value: 0.51 ± 0.14, and F1-score: 0.49 ± 0.18). For non-balanced classes, chance levels derived by permutation tests for accuracy (62.31 %) are higher than the initial classification accuracy (61.15 %). Thus Thus, there is no evidence for a systematic effect in time domain data exploitable for successful classification of low and high valence. Furthermore, the initial consideration of AUC-values [154] being the most robust performance measure in this analysis across non-balanced and balanced classes is supported.

5.4.2. Frequency Domain Classification

Non-balanced (except for F1-scores) and balanced classes yielded above chance performance at group level. Performances were significantly different between class size conditions. For non-balanced classes, accuracy and AUC-values were at 74.74 % and 0.69, respectively. However, these results are difficult to interpret due to the small number of trials per class as well as skewed class sizes. Chance level derived by permutation tests in non-balanced classification for accuracy with 64.67 % was significantly higher than 50 % as was expected by the class ratios ($\frac{10}{28}$ to $\frac{18}{28}$). F1-scores for non-balanced classes were below chance with 0.44 and the derived chance level from permutation tests was at 0.15. As outlined in Section 2.7.4, F1-scores are computed from the TP, FP, as well as FN, however

they do not take into account the TN which renders this performance measure challenging to interpret in certain circumstances for it strongly depends on the definition of the TP, FP, and FN [155]. The results presented are superior to Winkler et al. (2010) [94] who classified frontal spectral features in negative vs. positive emotional states induced by pictures with 56 % accuracy. In comparison Koelstra et al. [167] reported an F1-score of 0.564 when classifying between low and high valence in non-balanced classes with Fisher's linear discriminant analysis [144] and frequency domain features. As features, they employed averages across gross frequency bands (theta: 4-7 Hz, alpha: 8-13 Hz, beta: 14-29 Hz, and gamma: 30-47 Hz) as well as inter-hemispheric indices. Performances achieved with the classification apparatus developed in the present thesis are superior to the results from [167]. Gupta and Falk (2015) [] introduced graph theoretical features for the classification of the DEAP dataset and reported a performance increase up to 66 % when classifying valence. The authors did not address class-imbalance. Technically, the performances achieved with our approach of 73.74 % are clearly higher. As outlined throughout the present work, it is generally advisable to only refer to the balanced classes classification performance.
Sample size, especially across classes, is a critical factor in the validation of classification performance. At the same time, the number of samples per class a crucial success factor in training a machine learning model. As a rule of thumb, there should be at least 40 instances per class available for classification [191]. Across classification results and participants, AUC-values are the most robust measure besides accuracy in terms of outliers.

5.5. Conclusion

The DEAP dataset released by Koelstra et al. [167] is a dataset for the study of affective states in central and peripheral physiological signals. Classification of spectral frequency domain features with R^2-value feature selection and SVM between low and high valence is superior to the results reported by Koelstra et al. employing an LDA classification approach. Group averages of classified frequency domain features were significantly above chance for all performance measures in balanced classes (64.29 % accuracy, 0.66 AUC-value, and 0.64 F1-score). Koelstra et al. reported 57.60 % in group average accuracy and 0.56 in F1-score when non-balanced classes were classified in Fisher's linear discriminant analysis. Although trial sizes were small, the results show successful discrimination between low and high valence using the machine learning pipeline introduced with EEG frequency domain features.

6 Discussion

To date, classic human-computer interaction (HCI), as the interaction between humans and computing systems, lacks affect as a communication channel. The relatively young field of affective computing seeks to also incorporate psychophysiological information about the inner state of an individual into classic HCI [10, 118]. Recent technological advancements in computing soft-, and hardware, as well as in the recording of physiological signals have led to an increased interest in the automatic extraction and interpretation of psychophysiological information.

The present thesis focuses on the classification of three affective states (i.e. unpleasant, neutral, and pleasant) from the electroencephalogram (EEG) in healthy adults, motor-impaired individuals with cerebral palsy, as well as preverbal infants. Affective states are derived from valence following the dimensional emotion model [16, 17]. Therefore, a machine learning framework has been developed for MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States) and made publicly available online (`https://github.com/dthettich/BSClassify`). The framework is based on fast feature selection and reduction by R^2-values [131] as well as a linear support vector machine classifier [153].

A machine learning pipeline common to brain-computer interface research is maintained throughout analyses [123]. As features, only statistically significant correlates of affect known to the neuroscience literature are employed from the time- and frequency domain.

6.1. Auditory Affect Induction and Classification

Neural responses to a selection of emotion-laden sounds from the International Affective Digitized Sounds 2nd Edition (IADS-2) database [160] are investigated in a healthy as well as in a motor-impaired population of individuals with cerebral palsy. The paradigm design focuses on the maximization of trials available with simultaneously avoiding confounding factors such as habituation. For each emotional condition, 40 trials are available. Sounds are suitable stimuli for participants where visual fixation is absent [40]. Individual participant ratings correlated strongly with literature ratings ($r = 0.81$, $p < 0.001$).

As a time domain correlate of affect, the late positive potential (LPP) is identified over central midline electrodes in the healthy population. The LPP is an event-related potential (ERP) with a positive deflection approximately 400 ms post stimulus-onset with variable length. Amplitudes are more positive relative to affective content of stimuli. Results are comparable to [54, 55].

Time domain features of affect have not been employed for classification of affective states in the EEG. Following the machine learning pipeline outlined, above chance group average classification of unpleasant vs. pleasant (53.39 % accuracy and 0.54 AUC-value) as well as neutral vs. pleasant (53.32 % accuracy and 0.54 AUC-value) is achieved in the healthy population. In comparison, Koelstra et al. [167] report significant group average classification in a similar population of low vs. high valence (i.e. unpleasant vs. pleasant) of 57.60 % accuracy and 0.56 F1-score. However, classes in that study are not balanced, thus scaling up the chance level. Furthermore, features are based on frequency domain spectra and stimuli are in total 40 music video clips of 60 s.

Inter-hemispheric frontal frequency domain correlates of affect are only present in trend, yet are not significant.

For the first time, affect classification is conducted in a motor-impaired population with cerebral palsy. This population consists of potential target users who benefit of affective brain-computer interfaces. In the comparably small sample dataset size of $n = 4$, time domain LPP differences are not significant, yet visible as a trend. Although meticulous care is taken when applying artifact rejection techniques [130], movement artifacts render the feasibility of analyses difficult in this population [40, 192, 193]. Classification of affective states is also conducted employing time domain features of cerebral palsy data. However, results are not useful since good practice in the classification of brain state differences is the statistical testing of such, before classification [172]. They are included as an example that potentially contaminated EEG may yield above chance classification performance.

There are several limitations to the current study. The advantages of auditory stimuli from the IADS-2 have been outlined. Yet, their effectiveness as a tool for substantial emotion induction, especially at the low or high end of the valence scale in a healthy population, is open to debate. For example, a handful of healthy participants verbally communicated after the debriefing that the experienced sounds where not really strong when compared to scenes from a current movie in the cinema.

Although the paradigm is designed to maximize trials when compared with other studies [169], the amount of trials available is still small when compared to other classification problems. For example, when classifying user input in a regular P300 speller BCI, approximately 180 trials are available in total for one symbol.

In conclusion, the paradigm design with maximized number of trials, statistically validated features, as well as validated and comparable classification results of this study are a step towards the right direction in affective computing.

6.2. Affect Classification in Infants

The vision behind affect classification in preverbal infants is to provide an emotional communication channel between the child on a sensory deprived or severely impaired caretaker. Electrophysiological data of a novel emotion induction paradigm for infants up to 6-months of age are investigated. In that paradigm, emotions are induced by the interaction of infant-adult pairs in different scenarios. The adult is a meaningful stimulus to an infant, when compared to standardized stimuli [21]. Furthermore, with standardized audio-visual stimuli, it is more difficult for infants to maintain target focus.

Frontal inter-hemispheric differences in EEG power relative to stimulus valence are reported in the literature [78, 67]. Furthermore, hedonic theta is stated in relation to pleasurable stimulation [73]. In general, infant EEG is found to be shifted to the left in terms of frequency bands [51], i.e. neural firing generally occurs less frequent when compared to adult EEG. In the present data, frontal EEG asymmetries between hemispheres for unpleasant to neutral as well as pleasant to neutral are validated in the range 3 to 6 Hz.

For the first time, automatic affect classification in preverbal infant EEG was conducted. Our results support the possibility of the construction of a simple, non-invasive automatic emotional valence classification system in a brain-computer interface (BCI) context [10, 118] even for very small children with the age between 4 to 6 months of age. In a cross-subject classification approach by leave-one-subject-out estimation (LOSOE), significant above chance performances are obtained using frequency domain features (62.62 % accuracy, 0.65 AUC-value, and 0.60 F1-score). Comparable machine learning studies regarding affect classification of infant EEG are not available. However, there are two studies applying machine learning to data of infant-adult interaction in different contexts. Shami et al. [194] report results of different classification methods of emotions in adult-infant speech.

Messinger et al. [195] conduct behavior prediction by adult-infant face-to-face interaction and machine learning in order to design smart agents for interaction.
Compared to classification results of the auditory affect induction and classification study outlined in Chapter 3, infant classification performances are greatly higher. The own parent is a strong and relevant stimulus to a preverbal infant. Thus, it is suggested to address the design of more relevant and personalized emotional stimuli for the study of affect also in adults in an affective computing context. The LOSOE accounts for a strong feasibility in the classification of neutral vs. pleasant emotional states in the present infant EEG data. Training set data are balanced in order to avoid shifted baselines of chance levels.
Only neutral and pleasant conditions are classified since the amount of unpleasant trials is rather small compared with the other two. The topic of inducing unpleasant states in infants has to be taken seriously, let alone due to ethical reasons. Thus, the scenarios for this condition are designed such that the adult expresses their concern about their infant. This already resulted in behaviorally observable changes of expression. Yet the success of emotion induction is supported by results of physiological analyses and classification. The results of this realistic, everyday life emotion induction approach strengthens our vision of an automatic and non-invasive emotional detection possibility in very small infants. Particularly in social interactions between a child and a severely impaired or sensory deprived caretaker such a system may significantly improve quality of interaction and care. Including peripheral physiological measures in addition to the EEG and/or adding other non-invasive central nervous system measures such as portable near-Infrared spectroscopy (NIRS) will likely greatly improve classification performance.

6.3. DEAP Classification & Comparison

The DEAP dataset released by Koelstra et al. [167] is a dataset for the study of affective states in central and peripheral physiological signals. Time domain and frequency domain features of two affective states, i.e. low and high valence, of the central nervous system are classified. Koelstra et al. report significant above chance classification of these two conditions using statistically validated spectral features. In the comparison of classifying non-balanced and balanced classes, it is shown that non-balanced classes lead to spurious performances. Precisely in the classification of time domain features of non-balanced classes, an accuracy of 62.31 % is obtained. Nonetheless, AUC-values are robust to such imbalances. As a main strategy however, balanced classes are recommended for the production of comparable results across classification studies. Permutation tests are highly recommended in any case [196, 197, 157]. By classifying frequency domain features, the machine learning apparatus employed here outperforms the results by Koelstra et al. ([167], Table 7). The machine learning cascade developed in this thesis clearly shows potential and success in the classification of affective states in the EEG. Nonetheless, spectral features were computed from 60 s EEG thus increasing the likelihood of actual "emotional" EEG being present for classification. In comparison to the auditory affect induction study conducted in healthy and individuals with cerebral palsy, the lower performance there is likely due to the different paradigm design especially with the smaller time frame of 1.4 s and features employed for classification. The cerebral palsy data furthermore showed artifacts which added substantial noise and variance to the data likely contaminating the target features, i.e. ERP amplitude in the late positive potential.

6.4. Conclusion and Future Directions

The present work evaluates the feasibility of affect classification in the electroencephalogram of the affective states unpleasant, neutral, and pleasant. In the context of affective computing, two proof-of-concept studies in three different populations are introduced: healthy adults, motor-impaired cerebral palsy individuals, and preverbal infants.

Current research on the automatic classification of affective states from electrophysiological signals lacks comparability of results due to significant differences in paradigm design, methodology, as well as machine learning approaches [169]. Yet, alone the experimental setup for emotion induction comprises a vast parameter space and possible confounding factors [172].

Throughout the present work, all studies are designed for an offline classification analysis with a common machine learning apparatus in order to establish comparable results. The results suggest a significant above chance group classification of two affective states in a healthy and a preverbal infant population. Physiological correlates of affect in a cerebral palsy population, as a real user target group for affective computing systems, could not be validated statistically, however is present in trend.

Sample and therefore class size are identified as a key parameter for the success of classification in machine learning. As outlined, it is of utmost importance to follow correct classification practices, e.g. balancing classes or performance measures, to allow for a fair comparison of results. Therefore, the following measures should always be named when reporting results: number of classes, number of samples per class, number of features, performance computation (e.g. cross-validation), chance level (directly computed [158] or computed by permutation tests), and machine learning approach.

The quality of "emotional" EEG available for analysis is a starting point for future research. Especially the identification of meaningful individualized stimuli seems key for future affective computing research in order to establish solid and high-quality data. Affective computing is an interdisciplinary endeavour where neuroscientists, psychologists, and computer scientists must closely work together and correspond vividly such that field-specific expertise is combined in order to foster great results. For the future, it is desirable to establish further open datasets of affective physiological data as well as to publish analysis and machine learning code. As a next step, the combination of additional physiological measures, for example from the periphery, should be investigated in affective computing. The results presented in this thesis strengthen the vision of an automatic affect recognition system by means of physiology augmenting brain-computer interfaces by the ability to identify and communicate users' inner states of affect.

A Appendix

A.1. IADS-2 Ids, Valence, and Arousal Values

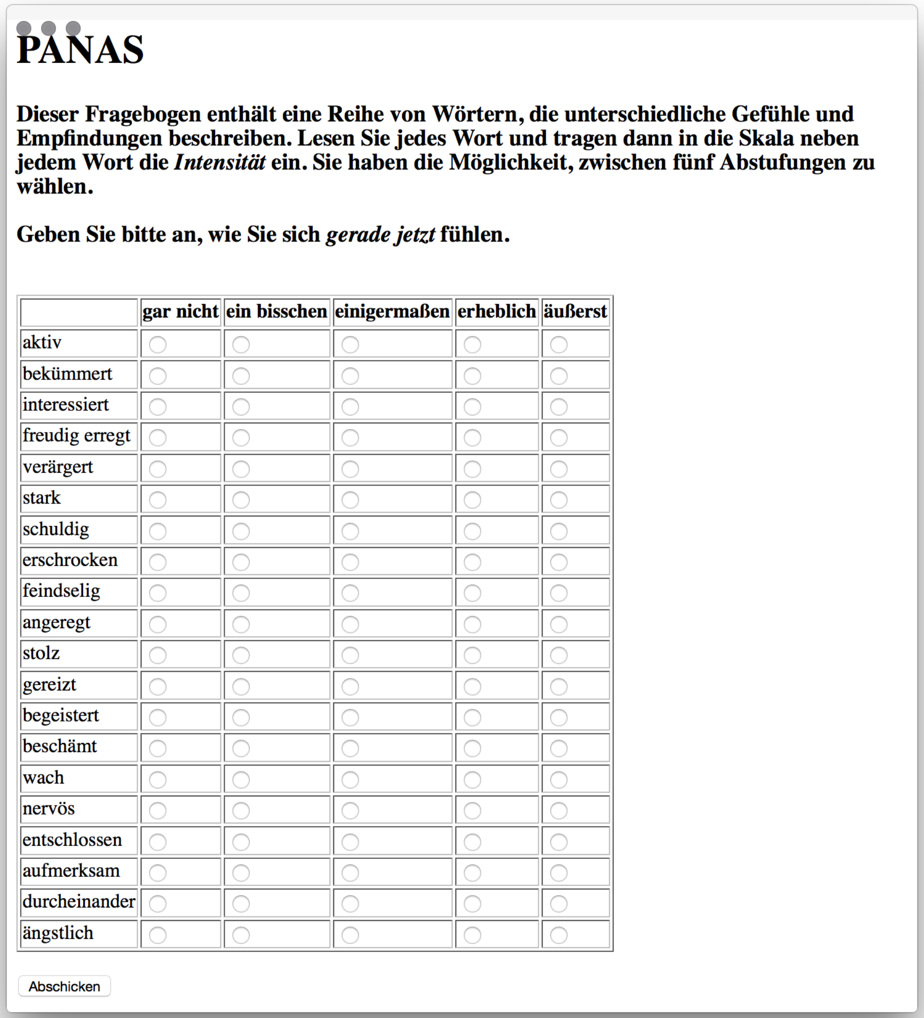
For the auditory affect induction and classification study outlined in Chapter 3, sounds from the International Affective Digitized Sounds 2nd Edition (IADS-2) database [160] were employed. Table A.1 denotes IADS-2 sound ids and corresponding valence/arousal values in the respective valence category.

Table A.1.: IADS-2 sound ids and respective valence/arousal values for each emotional category.

Unpleasant			**Neutral**			**Pleasant**		
Id.	**Val.**	**Aro.**	**Id**	**Val.**	**Aro.**	**Id**	**Val.**	**Aro.**
106	1.57	5.68	102	4.52	2.88	110	6.31	3.36
115	1.68	6.07	120	4.52	4.03	112	6.62	3.36
244	1.68	6.31	170	4.63	4.12	151	6.81	4.18
255	1.93	6.39	246	4.68	4.35	172	6.82	4.46
260	2.01	6.57	262	4.72	4.41	200	6.84	4.47
276	2.04	6.59	322	4.83	4.42	202	6.94	4.51
278	2.04	6.82	358	4.83	4.60	220	6.94	4.95
279	2.06	6.87	364	4.83	4.60	226	6.97	5.42
284	2.08	6.91	368	4.86	4.65	311	7.00	5.87
286	2.16	7.03	373	4.88	4.65	360	7.12	5.89
288	2.34	7.05	376	4.95	4.65	365	7.20	6.00
289	2.42	7.08	410	5.01	4.75	716	7.28	6.03
296	2.44	7.10	425	5.09	4.79	726	7.40	6.32
420	2.61	7.27	627	5.09	4.87	809	7.44	6.44
424	2.65	7.39	698	5.15	4.91	810	7.51	6.85
624	2.71	7.77	700	5.18	4.97	811	7.64	7.10
703	2.82	7.88	701	5.19	5.15	813	7.65	7.12
711	2.89	7.95	722	5.20	5.41	815	7.67	7.13
712	3.08	7.98	723	5.26	5.62	817	7.78	7.15
719	3.37	7.99	728	5.31	5.89	820	7.90	7.54
Mean	2.34	7.04	**Mean**	4.94	4.69	**Mean**	7.19	5.71

A.2. German Version of PANAS Questionnaire

To assess whether participants had substantial deviations from their current moods in the auditory affect induction and classification study outlined in Chapter 3, a German version of the Positive Affect Negative Affect Scale (PANAS) [173, 174] questionnaire had to be filled out by participants. Participant ratings were only in the range of "gar nicht", "ein bisschen", or "einigermaßen". Thus, there were no substantial deviations to a standard baseline constitution.



PANAS

Dieser Fragebogen enthält eine Reihe von Wörtern, die unterschiedliche Gefühle und Empfindungen beschreiben. Lesen Sie jedes Wort und tragen dann in die Skala neben jedem Wort die *Intensität* ein. Sie haben die Möglichkeit, zwischen fünf Abstufungen zu wählen.

Geben Sie bitte an, wie Sie sich *gerade jetzt* fühlen.

	gar nicht	ein bisschen	einigermaßen	erheblich	äußerst
aktiv	○	○	○	○	○
bekümmert	○	○	○	○	○
interessiert	○	○	○	○	○
freudig erregt	○	○	○	○	○
verärgert	○	○	○	○	○
stark	○	○	○	○	○
schuldig	○	○	○	○	○
erschrocken	○	○	○	○	○
feindselig	○	○	○	○	○
angeregt	○	○	○	○	○
stolz	○	○	○	○	○
gereizt	○	○	○	○	○
begeistert	○	○	○	○	○
beschämt	○	○	○	○	○
wach	○	○	○	○	○
nervös	○	○	○	○	○
entschlossen	○	○	○	○	○
aufmerksam	○	○	○	○	○
durcheinander	○	○	○	○	○
ängstlich	○	○	○	○	○

Figure A.1.: German version of the PANAS realized in web browser form.

A.3. Healthy Individual Chance Levels Obtained by Permutation Tests

Table A.2 shows healthy individual chance levels obtained by permutation tests at $\alpha = 0.5$ of time domain features form the auditory affect induction and classification study outlined in Chapter 3. The notation with α stems from the one introduced in Chapter 2.7.4. In the present thesis, the α-notation for individual significance at different levels is also used as the more known p-notation.

Table A.2.: Healthy individual chance levels of classification at significance threshold $\alpha = 0.5$ obtained by permutation tests for the performance measures accuracy, AUC-value, and F1-score based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 in 100 iterations. Columns indicate classes of respective binary classification problems (’-’ unpleasant, ’0’ neutral, ’+’ pleasant). Classes are balanced with 40 instances each.

	’-’ vs. ’0’			**’-’ vs. ’+’**			**’+’ vs. ’0’**		
Participant	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**
S01	51.25 %	0.51	0.50	48.75 %	0.49	0.49	51.25 %	0.51	0.51
S02	50.36 %	0.50	0.49	48.57 %	0.49	0.47	50.00 %	0.50	0.49
S03	48.75 %	0.49	0.49	51.25 %	0.51	0.52	50.00 %	0.50	0.51
S04	48.75 %	0.49	0.49	48.75 %	0.49	0.50	46.25 %	0.46	0.47
S05	51.25 %	0.51	0.51	48.75 %	0.49	0.50	50.00 %	0.50	0.48
S06	50.00 %	0.50	0.51	48.75 %	0.49	0.50	50.00 %	0.50	0.52
S07	51.25 %	0.51	0.49	50.00 %	0.50	0.51	47.50 %	0.47	0.49
S08	48.75 %	0.49	0.49	47.50 %	0.47	0.48	48.75 %	0.49	0.49
S09	50.00 %	0.50	0.51	51.25 %	0.51	0.51	50.00 %	0.50	0.51
S10	50.00 %	0.50	0.49	50.00 %	0.50	0.50	48.75 %	0.49	0.48
S11	48.75 %	0.49	0.49	48.75 %	0.49	0.48	50.00 %	0.50	0.50
S12	48.75 %	0.49	0.48	48.75 %	0.49	0.49	48.75 %	0.49	0.49
S13	50.00 %	0.50	0.49	48.75 %	0.49	0.51	50.00 %	0.50	0.50
S14	48.75 %	0.49	0.51	50.00 %	0.50	0.52	47.50 %	0.47	0.49
S15	48.75 %	0.49	0.49	48.75 %	0.49	0.49	50.00 %	0.50	0.49
S16	50.00 %	0.50	0.51	51.25 %	0.51	0.51	50.00 %	0.50	0.52
S17	48.75 %	0.49	0.48	48.75 %	0.49	0.50	50.00 %	0.50	0.51
S18	50.00 %	0.50	0.49	48.75 %	0.49	0.50	50.00 %	0.50	0.51
S19	48.75 %	0.49	0.49	48.75 %	0.49	0.49	51.25 %	0.51	0.51
S20	50.00 %	0.50	0.51	50.00 %	0.50	0.50	48.75 %	0.49	0.49
S21	50.00 %	0.50	0.52	48.75 %	0.49	0.49	50.00 %	0.50	0.49
S22	48.75 %	0.49	0.49	50.00 %	0.50	0.49	48.75 %	0.49	0.49
S23	47.50 %	0.47	0.49	48.75 %	0.49	0.50	50.00 %	0.50	0.51
Mean	49.53 %	0.50	0.50	49.29 %	0.49	0.50	49.46 %	0.49	0.50

A.4. Cerebral Palsy Individual Chance Levels Obtained by Permutation Tests

Table A.3 shows cerebral palsy individual chance levels obtained by permutation tests at $\alpha = 0.5$ of time domain features form the auditory affect induction and classification study outlined in Chapter 3.

Table A.3.: Cerebral palsy individual chance levels of classification at significance threshold $\alpha = 0.5$ obtained by permutation tests for the performance measures accuracy, AUC-value, and F1-score based on time domain EEG data of channels Cz, Pz, Cp1, Cp2, Cp5, and Cp6 in 100 iterations. Columns indicate classes of respective binary classification problems ('-' unpleasant, '0' neutral, '+' pleasant). Classes are balanced with 40 instances each.

	'-' vs. '0'			**'-' vs. '+'**			**'+' vs. '0'**		
Participant	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**	**Accuracy**	**AUC**	**F1-Score**
S01	51.25 %	0.51	0.46	50.00 %	0.50	0.33	53.75 %	0.54	0.50
S02	56.25 %	0.56	0.52	48.75 %	0.49	0.43	50.00 %	0.50	0.45
S03	55.18 %	0.55	0.59	48.04 %	0.48	0.55	70.36 %	0.70	0.73
S04	51.25 %	0.51	0.50	51.25 %	0.51	0.49	47.50 %	0.47	0.47
Mean	53.48 %	0.53	0.52	49.51 %	0.50	0.45	55.40 %	0.55	0.54

B
Appendix

B.1. Infant Affect Classification Performance Measure Comparison

Cross-subject classification performances obtained by LOSOE, respective individual significance levels obtained by permutation tests at $p = 0.05$, as well as class ratios are shown in Figure 4.6 for the three performance measures: accuracy, AUC-value, and F1-score. The average accuracy is 62.62 ± 11.09 %. The average AUC-value is 0.65 ± 0.14. The average F1-score is 0.60 ± 0.16. Class ratios in the testing set are on average 66 ± 12 % as compared to the larger class.

An overview when individual significance thresholds are exceeded for the respective performance measure are shown in Table B.1.

In total, 15 subjects exceed individual significance levels in all performance measures (i.e. subjects 1, 2, 3, 4 9, 14, 15, 16 17 19, 20, 23, and 25). Subjects 5 and 8 exhibit no significant classification at all. Subject 10 shows significance only in AUC-value. Subjects where only F1-scores exceed the significance threshold are 13, 21, and 24.

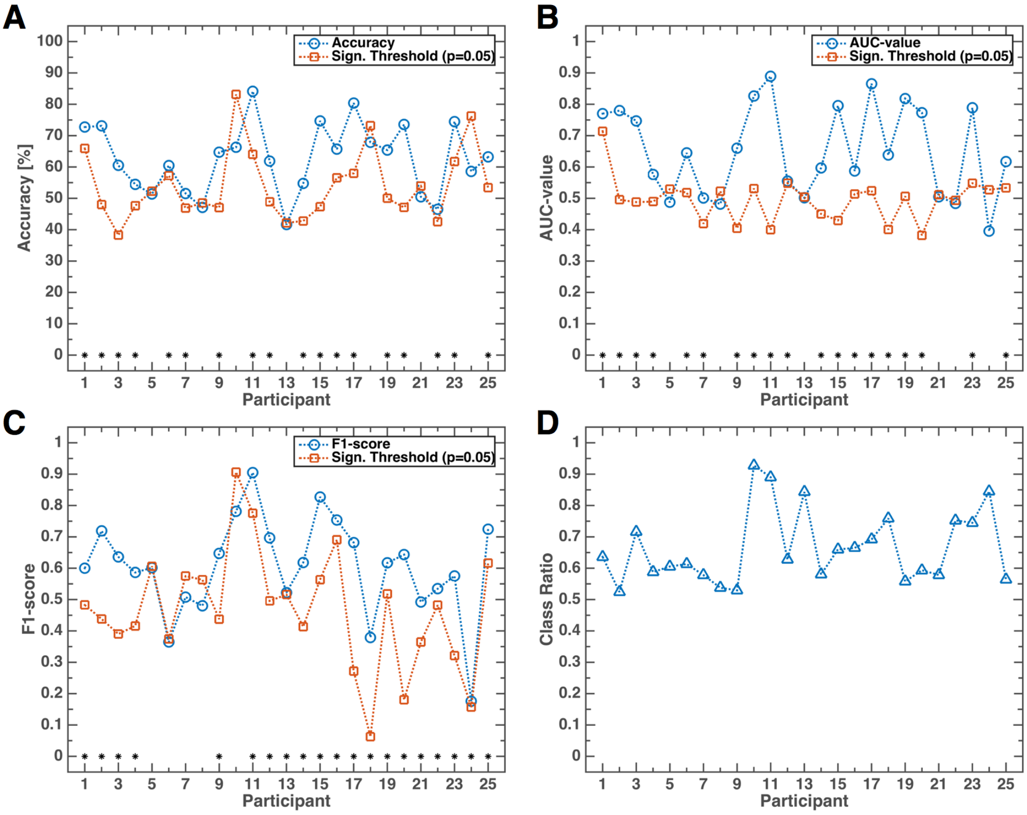


Figure B.1.: Cross-subject classification performances depicted in blue: accuracies (**A**), AUC-values (**B**), F1-scores (**C**), as well as class ratios of the testing set (**D**) of binary classification of pleasant vs. neutral conditions using features from the frequency domain 1-9 Hz. Permutation results at $p = 0.05$ are depicted in red. Asterisks at the 0 mark indicate if performance measures exceed the 5 % significance level. Classes for model training are balanced.

Table B.1.: Per subject individual significance exceeded at $p = 0.05$ for performance measures accuracy, AUC-values, and F1-scores. Distance values indicate disagreement between the three performance measures.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
Accuracy	*	*	*	*		*	*		*		*	*		*	*	*	*		*	*		*	*		*
AUC-value	*	*	*	*		*	*		*	*	*	*		*	*	*	*	*	*	*			*		*
F1-score	*	*	*	*					*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Distance	0	0	0	0	0	1	1	0	0	2	0	0	2	0	0	0	0	1	0	0	2	1	0	2	0

C
Appendix

C.1. DEAP Individual Chance Levels Obtained by Permutation Tests

Table C.1 contrasts individual chance levels at $p = 0.5$ as obtained by permutation tests as well as individual performances of time domain feature classification from the DEAP dataset analysis outlined in Chapter 5. Bold values indicate when individual performance exceeds individual chance levels.

Table C.2 similarly shows the same results yet for the classification of frequency domain features.

Table C.1.: DEAP Time domain average classification chance levels obtained by permutation tests ($p = 0.5$) for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score for each participant, as well as individual classification performances. Values for accuracy are given in percent. Bold values indicate when individual performance exceed the individual chance level.

	Non-balanced						**Balanced**					
	Accuracy [%]		**AUC-value**		**F1-score**		**Accuracy [%]**		**AUC-value**		**F1-score**	
Id.	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**
01	**62.86**	57.14	**0.74**	0.60	**0.65**	0.53	**82.86**	56.67	**0.85**	0.61	**0.82**	0.55
02	58.67	69.33	0.59	0.60	**0.25**	0.00	43.33	66.67	0.45	0.72	0.38	0.62
03	75.00	75.00	0.64	0.66	0.00	0.00	40.00	60.00	0.62	0.62	0.60	0.75
04	47.14	58.10	**0.58**	0.55	0.59	0.68	54.00	68.67	0.49	0.49	0.70	0.81
05	66.67	70.00	**0.50**	0.37	**0.17**	0.00	**50.00**	45.00	**0.48**	0.43	**0.56**	0.50
06	86.00	86.00	**0.51**	0.46	0.00	0.00	**50.00**	40.00	**0.66**	0.44	**0.57**	0.00
07	70.95	73.81	0.52	0.56	**0.29**	0.00	**50.00**	38.33	**0.40**	0.20	**0.53**	0.52
08	**69.00**	61.00	**0.55**	0.38	**0.46**	0.17	**50.00**	38.33	**0.50**	0.36	**0.50**	0.38
09	67.00	67.00	0.66	0.76	**0.53**	0.40	55.00	61.67	**0.72**	0.65	0.59	0.62
10	40.00	48.00	0.39	0.47	0.22	0.38	**50.00**	45.00	**0.51**	0.47	0.44	0.53
11	40.67	52.00	0.37	0.50	**0.33**	0.14	**54.00**	50.00	0.48	0.53	0.44	0.45
12	**50.48**	50.00	**0.60**	0.47	**0.33**	0.32	50.00	50.00	0.46	0.49	0.40	0.40
13	**64.00**	44.00	**0.54**	0.45	**0.61**	0.53	**58.00**	54.00	**0.60**	0.49	0.62	0.63
14	42.86	57.14	0.39	0.61	0.33	0.55	53.33	63.33	0.45	0.68	0.48	0.68
15	**55.33**	44.67	**0.60**	0.28	**0.25**	0.12	35.00	40.00	**0.46**	0.26	0.13	0.57
16	53.33	57.33	**0.44**	0.40	0.63	0.71	73.00	77.00	**0.54**	0.34	0.83	0.87
17	52.00	68.00	0.48	0.57	0.00	0.00	20.00	63.33	0.23	0.64	0.15	0.62
18	85.00	85.00	**0.34**	0.22	0.00	0.00	10.00	30.00	**0.33**	0.06	0.00	0.50
19	**58.00**	42.67	**0.57**	0.38	**0.40**	0.26	36.00	43.00	**0.47**	0.40	0.29	0.48
20	57.00	70.00	0.43	0.49	0.00	0.00	43.33	43.33	0.40	0.41	**0.50**	0.33
21	57.33	65.33	0.39	0.59	**0.15**	0.00	50.00	56.67	0.56	0.58	0.56	0.62
22	**50.00**	43.33	**0.59**	0.39	**0.52**	0.44	**56.67**	43.33	**0.63**	0.42	**0.55**	0.45
23	83.00	83.00	0.57	0.59	0.00	0.00	**50.00**	20.00	**0.59**	0.28	**0.33**	0.25
24	**51.33**	48.67	**0.48**	0.36	**0.32**	0.22	**44.00**	37.00	**0.34**	0.31	**0.40**	0.29
25	**53.33**	40.00	**0.52**	0.38	**0.56**	0.43	**46.67**	40.00	**0.47**	0.35	**0.43**	0.41
26	**58.57**	55.71	**0.50**	0.38	0.12	0.12	41.00	47.00	0.42	0.47	0.42	0.50
27	78.00	81.33	**0.55**	0.51	0.00	0.00	50.00	60.00	0.50	0.50	0.55	0.60
28	**57.14**	54.29	**0.59**	0.47	**0.35**	0.19	58.67	65.33	0.61	0.73	0.59	0.71
29	46.67	51.90	**0.49**	0.40	0.19	0.21	**43.00**	41.00	0.27	0.40	**0.42**	0.36
30	78.33	85.00	0.19	0.58	0.00	0.00	**60.00**	30.00	**0.89**	0.28	**0.75**	0.50
31	64.00	72.00	0.26	0.57	**0.18**	0.00	**56.67**	50.00	0.44	0.48	**0.57**	0.46
32	77.00	77.00	**0.65**	0.37	0.00	0.00	50.00	60.00	0.56	0.62	**0.55**	0.50
Mean	61.15	62.31	0.51	0.48	0.26	0.20	48.89	49.52	0.51	0.46	0.49	0.52

Table C.2.: DEAP Frequency domain average classification performances for non-balanced and balanced datasets for metrics accuracy, AUC-value, and F1-score, as well as their corresponding individual significance thresholds ($p = 0.5$). Values for accuracy are given in percent. Bold values indicate when individual performance exceed the individual significance level.

	Non-balanced						**Balanced**					
	Accuracy [%]		**AUC-value**		**F1-score**		**Accuracy [%]**		**AUC-value**		**F1-score**	
Id.	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**	**Ind.**	**Thresh.**
01	**57.14**	48.57	**0.51**	0.47	**0.55**	0.40	**56.19**	47.14	**0.51**	0.47	**0.55**	0.47
02	66.67	69.33	**0.78**	0.46	**0.17**	0.00	**88.33**	55.00	**0.86**	0.57	**0.88**	0.52
03	**80.00**	75.00	**0.74**	0.42	**0.33**	0.00	**60.00**	40.00	**0.88**	0.44	**0.67**	0.44
04	**85.24**	56.19	**0.86**	0.47	**0.87**	0.71	72.00	72.00	**0.51**	0.47	0.83	0.83
05	73.33	73.33	0.34	0.42	0.00	0.00	38.33	41.67	0.42	0.42	0.38	0.40
06	86.00	86.00	0.13	0.36	0.00	0.00	10.00	30.00	0.12	0.31	0.00	0.29
07	70.95	73.81	**0.76**	0.47	**0.17**	0.00	**68.33**	50.00	**0.81**	0.49	**0.67**	0.44
08	**75.00**	65.00	**0.58**	0.45	**0.40**	0.00	**50.00**	38.33	**0.47**	0.39	**0.50**	0.40
09	**80.00**	57.00	**0.86**	0.40	**0.71**	0.00	**83.33**	43.33	**0.91**	0.40	**0.82**	0.42
10	**76.00**	52.00	**0.91**	0.48	**0.74**	0.18	**85.00**	50.00	**0.95**	0.47	**0.86**	0.48
11	**59.33**	56.00	**0.50**	0.47	0.00	0.14	42.00	46.00	0.37	0.46	0.43	0.43
12	**68.57**	49.52	**0.73**	0.43	**0.69**	0.30	**60.00**	46.67	**0.67**	0.47	**0.62**	0.46
13	**76.00**	48.00	**0.75**	0.44	**0.77**	0.38	**79.00**	46.00	**0.86**	0.46	**0.80**	0.45
14	**91.43**	51.43	**0.90**	0.47	**0.91**	0.28	**80.95**	47.14	**0.90**	0.46	**0.83**	0.47
15	**74.00**	62.00	**0.89**	0.49	**0.72**	0.14	**75.00**	55.00	**0.82**	0.53	**0.76**	0.50
16	**60.67**	57.33	**0.64**	0.45	**0.74**	0.73	**82.00**	77.00	**0.86**	0.48	**0.89**	0.87
17	72.00	72.00	**0.52**	0.44	0.00	0.00	30.00	46.67	0.28	0.43	0.17	0.46
18	**90.00**	85.00	**1.00**	0.33	**0.50**	0.00	**100.00**	60.00	**1.00**	0.56	**1.00**	0.67
19	**72.00**	53.33	**0.66**	0.43	**0.64**	0.14	**68.00**	46.00	**0.72**	0.43	**0.67**	0.43
20	**79.00**	70.00	**0.87**	0.44	**0.62**	0.00	**86.67**	46.67	**0.98**	0.48	**0.86**	0.47
21	**73.33**	69.33	**0.67**	0.42	**0.22**	0.00	38.33	43.33	**0.47**	0.46	0.29	0.44
22	**70.00**	50.00	**0.70**	0.50	**0.69**	0.48	**63.33**	46.67	**0.60**	0.47	**0.65**	0.46
23	83.00	83.00	0.13	0.42	0.00	0.00	30.00	40.00	0.25	0.44	0.25	0.50
24	**62.67**	58.67	**0.60**	0.48	**0.38**	0.15	**48.00**	47.00	0.39	0.46	**0.52**	0.45
25	40.00	50.00	0.38	0.50	0.40	0.47	50.00	50.00	0.43	0.50	0.40	0.48
26	63.93	66.79	**0.80**	0.45	**0.13**	0.00	**51.00**	46.00	0.43	0.45	**0.45**	0.44
27	81.33	81.33	0.44	0.50	0.00	0.00	**70.00**	50.00	**0.68**	0.44	**0.67**	0.50
28	**71.43**	60.00	**0.75**	0.46	**0.58**	0.12	**80.67**	49.33	**0.88**	0.48	**0.81**	0.48
29	**77.62**	59.52	**0.85**	0.47	**0.70**	0.13	**84.00**	46.00	**0.83**	0.44	**0.85**	0.44
30	**90.00**	85.00	**0.98**	0.36	**0.50**	0.00	**100.00**	40.00	**1.00**	0.33	**1.00**	0.33
31	**76.00**	68.00	**0.87**	0.43	**0.57**	0.00	**76.67**	43.33	**0.84**	0.47	**0.80**	0.43
32	77.00	77.00	**0.89**	0.48	**0.29**	0.00	**50.00**	40.00	**0.46**	0.42	**0.62**	0.44
Mean	73.74	64.67	0.69	0.45	0.44	0.15	64.29	47.70	0.66	0.45	0.64	0.48

List of Figures

List of Tables

Bibliography

[1] Jacques J Vidal. Real-time detection of brain events in eeg. *Proceedings of the IEEE*, 65(5):633–641, 1977.

[2] Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

[3] Antonio R Damasio. *The feeling of what happens: Body, emotion and the making of consciousness*. Random House, 2000.

[4] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[5] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[6] Paul R Kleinginna Jr and Anne M Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379, 1981.

[7] William James. What is an emotion? *Mind*, 9:188–205, 1890.

[8] Walter B Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, pages 106–124, 1927.

[9] Stanley Schachter and Jerome Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379, 1962.

[10] Rosalind Wright Picard. Affective computing. *MIT Press, Cambridge, Mass.*, 1995.

[11] Charles Darwin. *The expression of the emotions in man and animals*. John Murray, London, 1872.

[12] Paul Ekman, Robert W Levenson, and Wallace V Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210, 1983.

[13] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[14] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[15] Paul Ekman. Facial expressions. *Handbook of cognition and emotion*, 16:301–320, 1999.

[16] Wilhelm Wundt. Über die psychische kausalität und das prinzip des psychologischen parallelismus. *Philosophische Studien*, (10):1–124, 1894.

[17] J. A. Russel. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1167 – 1178, 1980.

[18] Dirk Tassilo Hettich, Elaina Bolinger, Tamara Matuz, Niels Birbaumer, Wolfgang Rosenstiel, and Martin Spüler. Eeg responses to auditory stimuli for automatic affect recognition. *Frontiers in Neuroscience*, 10(244), 2016.

[19] Charles E Osgood. The nature and measurement of meaning. *Psychological bulletin*, 49(3):197, 1952.

[20] Anna Wierzbicka. Talking about emotions: Semantics, culture, and cognition. *Cognition & Emotion*, 6(3-4):285–319, 1992.

[21] Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. *Handbook of emotions*. Guilford Press, 2010.

[22] Teresa Farroni, Enrica Menon, Silvia Rigato, and Mark H Johnson. The perception of facial expressions in newborns. *European Journal of Developmental Psychology*, 4(1):2–13, 2007.

[23] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[24] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge university press, 1990.

[25] Lisa Feldman Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46, 2006.

[26] Zhihong Zeng, Maja Pantic, Glenn Roisman, Thomas S Huang, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[27] Charles Darwin, Paul Ekman, and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[28] John T Cacioppo, Louis G Tassinary, and Gary G Berntson. Psychophysiological science. *Handbook of psychophysiology*, 2:3–23, 2000.

[29] John T Cacioppo, Gary G Berntson, David J Klein, and Kirsten M Poehlmann. Psychophysiology of emotion across the life span. *Annual review of gerontology and geriatrics*, 17:27–74, 1997.

[30] Ulf Dimberg. Facial electromyography and emotional reactions. *Psychophysiology*, 1990.

[31] Gary E Schwartz, Paul L Fair, Patricia Salt, Michel R Mandel, and Gerald L Klerman. Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science*, 192(4238):489–491, 1976.

[32] Solange Akselrod, David Gordon, F Andrew Ubel, Daniel C Shannon, AC Berger, and Richard J Cohen. Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. *science*, 213(4504):220–222, 1981.

[33] Maurizio Codispoti, Vera Ferrari, and Margaret M Bradley. Repetitive picture processing: autonomic and cortical correlates. *Brain research*, 1068(1):213–220, 2006.

[34] Margaret M Bradley and Peter J Lang. Affective reactions to acoustic stimuli. *Psychophysiology*, 37(02):204–215, 2000.

[35] Maurizio Codispoti, Paola Surcinelli, and Bruno Baldaro. Watching emotional movies: Affective reactions and gender differences. *International Journal of Psychophysiology*, 69(2):90–95, 2008.

[36] Avram Goldstein. Thrills in response to music and other stimuli. *Physiological Psychology*, 8(1):126–129, 1980.

[37] Daniela Sammler, Maren Grigutsch, Thomas Fritz, and Stefan Koelsch. Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44(2):293–304, 2007.

[38] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30:261–261, 1993.

[39] Maurizio Codispoti, Margaret M Bradley, and Peter J Lang. Affective reactions to briefly presented pictures. *Psychophysiology*, 38(3):474–478, 2001.

[40] Yehya Mohamad, Dirk T Hettich, Elaina Bolinger, Niels Birbaumer, Wolfgang Rosenstiel, Martin Bogdan, and Tamara Matuz. *Detection and Utilization of Emotional State for Disabled Users*, pages 248–255. Springer, 2014.

[41] Stéphanie Khalfa, Peretz Isabelle, Blondin Jean-Pierre, and Robert Manon. Event-related skin conductance responses to musical emotions in humans. *Neuroscience letters*, 328(2):145–149, 2002.

[42] J LeDoux. Emotion circuits in the brain. 2003.

[43] Lisa Feldman Barrett, Batja Mesquita, Kevin N Ochsner, and James J Gross. The experience of emotion. *Annual review of psychology*, 58:373, 2007.

[44] Hedy Kober, Lisa Feldman Barrett, Josh Joseph, Eliza Bliss-Moreau, Kristen Lindquist, and Tor D Wager. Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage*, 42(2):998–1031, 2008.

[45] Kristen A Lindquist, Tor D Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, 35(03):121–143, 2012.

[46] Kristen A Lindquist, Ajay B Satpute, Tor D Wager, Jochen Weber, and Lisa Feldman Barrett. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. *Cerebral Cortex*, page bhv001, 2015.

[47] Lisa Feldman Barrett and Eliza Bliss-Moreau. Affect as a psychological primitive. *Advances in experimental social psychology*, 41:167–218, 2009.

[48] Hans Berger. Über das elektrenkephalogramm des menschen. *European archives of psychiatry and clinical neuroscience*, 98(1):231–254, 1933.

[49] Frank Sharbrough. American electroencephalographic society guidelines for standard electrode position nomenclature. *Clin Neurophysiology*, 8:200 – 202, 1991.

[50] H Aurlien, IO Gjerde, JH Aarseth, G Eldøen, B Karlsen, H Skeidsvoll, and NE Gilhus. Eeg background activity described by a large computerized database. *Clinical Neurophysiology*, 115(3):665–673, 2004.

[51] Peter J Marshall, Yair Bar-Haim, and Nathan A Fox. Development of the eeg from 5 months to 4 years of age. *Clinical Neurophysiology*, 113(8):1199–1208, 2002.

[52] Luis Carretié, José A Hinojosa, Manuel Martín-Loeches, Francisco Mercado, and Manuel Tapia. Automatic attention to emotional stimuli: neural correlates. *Human brain mapping*, 22(4):290–299, 2004.

[53] Jonas K Olofsson, Steven Nordin, Henrique Sequeira, and John Polich. Affective picture processing: an integrative review of erp findings. *Biological psychology*, 77(3):247–265, 2008.

[54] Harald T Schupp, Bruce N Cuthbert, Margaret M Bradley, John T Cacioppo, Tiffany Ito, and Peter J Lang. Affective picture processing: the late positive potential is modulated by motivational relevance. *Psychophysiology*, 37(2):257–261, 2000.

[55] B. N. Cuthbert, H. T. Schupp, M. M. Bradley, N. Birbaumer, and P. J. Lang. Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. *Biol Psychol*, 52(2):95–111, 2000.

[56] Kate E Briggs and Frances H Martin. Affective picture processing and motivational relevance: arousal and valence effects on erps in an oddball task. *International Journal of Psychophysiology*, 72(3):299–306, 2009.

[57] Charles A Nelson and Michelle de Haan. A neurobehavioral approach to the recognition of facial expressions in infancy. *The psychology of facial expression*, pages 176–204, 1997.

[58] Peter D Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[59] John Parker Burg. Maximum entropy spectral analysis. In *37th Annual International Meeting*. Society of Exploration Geophysics, 1967.

[60] Alan V Oppenheim, Ronald W Schafer, John R Buck, et al. *Discrete-time signal processing*, volume 2. Prentice-hall Englewood Cliffs, 1989.

[61] Mircea Steriade, David A McCormick, and Terrence J Sejnowski. Thalamocortical oscillations in the sleeping and aroused brain. *Science*, 262(5134):679–685, 1993.

[62] Mircea Steriade and Igor Timofeev. Neuronal plasticity in thalamocortical networks during sleep and waking oscillations. *Neuron*, 37(4):563–576, 2003.

[63] Gennady G Knyazev. Eeg delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience & Biobehavioral Reviews*, 36(1):677–695, 2012.

[64] Barbara E Jones. From waking to sleeping: neuronal and chemical substrates. *Trends in pharmacological sciences*, 26(11):578–586, 2005.

[65] Canan Başar-Eroglu, Erol Başar, Tamer Demiralp, and Martin Schürmann. P300-response: possible psychophysiological correlates in delta and theta frequency channels. a review. *International Journal of Psychophysiology*, 13(2):161–179, 1992.

[66] Michela Balconi and Claudio Lucchiari. Eeg correlates (event-related desynchronization) of emotional face elaboration: a temporal analysis. *Neuroscience letters*, 392(1):118–123, 2006.

[67] Michela Balconi and Uberto Pozzoli. Arousal effect on emotional face comprehension: frequency band changes in different time intervals. *Physiology & behavior*, 97(3):455–462, 2009.

[68] Manousos A Klados, Christos Frantzidis, Ana B Vivas, Christos Papadelis, Chrysa Lithari, Costas Pappas, and Panagiotis D Bamidis. A framework combining delta event-related oscillations (eros) and synchronisation effects (erd/ers) to study emotional processing. *Computational intelligence and neuroscience*, 2009:12, 2009.

[69] Wolfgang Klimesch. Memory processes, brain oscillations and eeg synchronization. *International Journal of Psychophysiology*, 24(1):61–100, 1996.

[70] Wolfgang Klimesch, Roman Freunberger, Paul Sauseng, and Walter Gruber. A short review of slow phase synchronization and memory: evidence for control processes in different memory systems? *Brain research*, 1235:31–44, 2008.

[71] Hiroaki Mizuhara and Yoko Yamaguchi. Human cortical circuits for central executive function emerge by theta phase synchronization. *Neuroimage*, 36(1):232–244, 2007.

[72] W Walter. Normal rhythms -— their development, distribution and significance. *Electroencephalography; a symposium on its various aspects.*, pages 203–227, 1950.

[73] Ernst Niedermeyer. The normal eeg of the waking adult. *Electroencephalography: Basic principles, clinical applications, and related fields*, page 167, 2005.

[74] Richard J Davidson and Nathan A Fox. Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants. *Science*, 218(4578):1235–1237, 1982.

[75] Carroll Ellis Izard. *Measuring emotions in infants and children*, volume 1. Cambridge University Press, 1982.

[76] Nathan A Fox and Richard J Davidson. Patterns of brain electrical activity during facial signs of emotion in 10-month-old infants. *Developmental Psychology*, 24(2):230, 1988.

[77] G Csibra, G Davis, MW Spratling, and MH Johnson. Gamma oscillations and object processing in the infant brain. *Science*, 290(5496):1582–1585, 2000.

[78] Ljubomir I Aftanas, Anton A Varlamov, Sergey V Pavlov, Viktor P Makhnev, and Natalya V Reva. Time-dependent cortical asymmetries induced by emotional arousal: Eeg analysis of event-related synchronization and desynchronization in individually defined frequency bands. *International Journal of Psychophysiology*, 44(1):67–82, 2002.

[79] LI Aftanas, AA Varlamov, SV Pavlov, VP Makhnev, and NV Reva. Affective picture processing: event-related synchronization within individually defined human theta band is modulated by valence dimension. *Neuroscience letters*, 303(2):115–118, 2001.

[80] Masaki Nishida, Jori Pearsall, Randy L Buckner, and Matthew P Walker. Rem sleep, prefrontal theta, and the consolidation of human emotional memory. *Cerebral Cortex*, 19(5):1158–1166, 2009.

[81] Wolfgang Klimesch, Paul Sauseng, and Simon Hanslmayr. Eeg alpha oscillations: the inhibition–timing hypothesis. *Brain research reviews*, 53(1):63–88, 2007.

[82] Robert J Barry, Adam R Clarke, Stuart J Johnstone, Christopher A Magee, and Jacqueline A Rushby. Eeg differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 118(12):2765–2773, 2007.

[83] Robert J Barry, Adam R Clarke, Stuart J Johnstone, and Christopher R Brown. Eeg differences in children between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 120(10):1806–1811, 2009.

[84] G Pfurtscheller, C Brunner, A Schlögl, and FH Lopes Da Silva. Mu rhythm (de) synchronization and eeg single-trial classification of different motor imagery tasks. *Neuroimage*, 31(1):153–159, 2006.

[85] Gert Pfurtscheller and Fernando H Lopes da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

[86] James A Coan and John JB Allen. Frontal eeg asymmetry as a moderator and mediator of emotion. *Biological psychology*, 67(1):7–50, 2004.

[87] G Wiedemann, P Pauli, W Dengler, W Lutzenberger, N Birbaumer, and G Buchkremer. Frontal brain asymmetry as a biological substrate of emotions in patients with panic disorders. *Archives of General Psychiatry*, 56(1):78–84, 1999.

[88] Elliot T Berkman and Matthew D Lieberman. Approaching the bad and avoiding the good: Lateral prefrontal cortical asymmetry distinguishes between action and valence. *Journal of Cognitive Neuroscience*, 22(9):1970–1979, 2010.

[89] Eddie Harmon-Jones and John JB Allen. Anger and frontal brain activity: Eeg asymmetry consistent with approach motivation despite negative affective valence. *Journal of personality and social psychology*, 74(5):1310, 1998.

[90] Charles S Carver and Eddie Harmon-Jones. Anger is an approach-related affect: evidence and implications. *Psychological bulletin*, 135(2):183, 2009.

[91] Eddie Harmon-Jones, Philip A Gable, and Carly K Peterson. The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update. *Biological psychology*, 84(3):451–462, 2010.

[92] Eddie Harmon-Jones, Philip A Gable, and Cindy Harmon-Jones. Individual differences in desire and approach motivation. *The Psychology of Desire*, 2015.

[93] Nathan A Fox and Richard J Davidson. Taste-elicited changes in facial signs of emotion and the asymmetry of brain electrical activity in human newborns. *Neuropsychologia*, 24(3):417–422, 1986.

[94] Irene Winkler, Mark Jäger, Vojkan Mihajlovic, and Tsvetomira Tsoneva. Frontal eeg asymmetry based classification of emotional valence using common spatial patterns. *World Academy of Science, Engineering and Technology*, 45:373–378, 2010.

[95] Stephen H Fairclough and Jenna S Roberts. Effects of performance feedback on cardiovascular reactivity and frontal eeg asymmetry. *International Journal of Psychophysiology*, 81(3):291–298, 2011.

[96] Willem J Kop, Stephen J Synowski, Miranda E Newell, Louis A Schmidt, Shari R Waldstein, and Nathan A Fox. Autonomic nervous system reactivity to positive and negative mood induction: The role of acute psychological responses and frontal electrocortical activity. *Biological psychology*, 86(3):230–238, 2011.

[97] Ronald N Goodman, Jeremy C Rietschel, Li-Chuan Lo, Michelle E Costanzo, and Bradley D Hatfield. Stress, emotion regulation and cognitive performance: The predictive contributions of trait and state relative frontal eeg alpha asymmetry. *International Journal of Psychophysiology*, 87(2):115–123, 2013.

[98] Stuart N Baker. Oscillatory interactions between sensorimotor cortex and the periphery. *Current opinion in neurobiology*, 17(6):649–655, 2007.

[99] Joscha T Schmiedt, Alexander Maier, Pascal Fries, Richard C Saunders, David A Leopold, and Michael C Schmid. Beta oscillation dynamics in extrastriate cortex after removal of primary visual cortex. *The Journal of Neuroscience*, 34(35):11857–11864, 2014.

[100] Yan Zhang, Yonghong Chen, Steven L Bressler, and Mingzhou Ding. Response preparation and inhibition: the role of the cortical sensorimotor beta rhythm. *Neuroscience*, 156(1):238–246, 2008.

[101] Andreas K Engel and Pascal Fries. Beta-band oscillations—signalling the status quo? *Current opinion in neurobiology*, 20(2):156–165, 2010.

[102] Andreas K Engel, Pascal Fries, and Wolf Singer. Dynamic predictions: oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, 2(10):704–716, 2001.

[103] Daniel Senkowski, Ulrich Pomper, Inga Fitzner, Andreas Karl Engel, and Andrej Kral. Beta-band activity in auditory pathways reflects speech localization and recognition in bilateral cochlear implant users. *Human brain mapping*, 35(7):3107–3121, 2014.

[104] Harry W Cole and William J Ray. Eeg correlates of emotional tasks related to attentional demands. *International Journal of Psychophysiology*, 3(1):33–41, 1985.

[105] Julie Onton and Scott Makeig. High-frequency broadband modulations of electroencephalographic spectra. *Frontiers in human neuroscience*, 3, 2009.

[106] Elise S Dan Glauser and Klaus R Scherer. Neuronal processes involved in subjective feeling emergence: Oscillatory activity during an emotional monitoring task. *Brain topography*, 20(4):224–231, 2008.

[107] Daniel Senkowski, Janine Kautz, Michael Hauck, Roger Zimmermann, and Andreas K Engel. Emotional facial expressions modulate pain-induced beta and gamma oscillations in sensorimotor cortex. *The Journal of Neuroscience*, 31(41):14542–14550, 2011.

[108] Pascal Fries. Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual review of neuroscience*, 32:209–224, 2009.

[109] Daniel Senkowski, Durk Talsma, Maren Grigutsch, Christoph S Herrmann, and Marty G Woldorff. Good times for multisensory integration: effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia*, 45(3):561–571, 2007.

[110] Daniel Senkowski, Till R Schneider, Frithjof Tandler, and Andreas K Engel. Gamma-band activity reflects multisensory matching in working memory. *Experimental brain research*, 198(2-3):363–372, 2009.

[111] Ole Jensen, Jochen Kaiser, and Jean-Philippe Lachaux. Human gamma-frequency oscillations associated with attention and memory. *Trends in neurosciences*, 30(7):317–324, 2007.

[112] Daniel Senkowski and Jürgen Gallinat. Dysfunctional prefrontal gamma-band oscillations reflect working memory and other cognitive deficits in schizophrenia. *Biological psychiatry*, 2015.

[113] H-Y Huang and P-C Lo. Eeg dynamics of experienced zen meditation practitioners probed by complexity index and spectral measure. *Journal of medical engineering & technology*, 33(4):314–321, 2009.

[114] Andreas Keil, Matthias M Müller, Thomas Gruber, Christian Wienbruch, Margarita Stolarova, and Thomas Elbert. Effects of emotional arousal in the cerebral hemispheres: a study of oscillatory brain activity and event-related potentials. *Clinical neurophysiology*, 112(11):2057–2068, 2001.

[115] LI Aftanas, NV Reva, AA Varlamov, SV Pavlov, and VP Makhnev. Analysis of evoked eeg synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. *Neuroscience and behavioral physiology*, 34(8):859–867, 2004.

[116] Drew B Headley and Denis Paré. In sync: gamma oscillations and emotional memory. *Frontiers in behavioral neuroscience*, 7, 2013.

[117] Lawrence J Hettinger, Pedro Branco, L Miguel Encarnacao, and Paolo Bonato. Neuroadaptive technologies: applying neuroergonomics to the design of advanced interfaces. *Theoretical Issues in Ergonomics Science*, 4(1-2):220–237, 2003.

[118] Stephen H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145, 2009.

[119] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan. Brain-computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

[120] Thorsten O Zander and Christian Kothe. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of neural engineering*, 8(2):025005, 2011.

[121] Gert Pfurtscheller, Brendan Z Allison, Clemens Brunner, Gunther Bauernfeind, Teodoro Solis-Escalante, Reinhold Scherer, Thorsten O Zander, Gernot Müller-Putz, Christa Neuper, and Niels Birbaumer. The hybrid bci. *Frontiers in neuroscience*, 4, 2010.

[122] Femke Nijboer, Jens Clausen, Brendan Z Allison, and Pim Haselager. The asilomar survey: Stakeholders' opinions on ethical issues related to brain-computer interfacing. *Neuroethics*, 6(3):541–578, 2013.

[123] Fabien Lotte, Marco Congedo, Anatole Lécuyer, and Fabrice Lamarche. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4, 2007.

[124] N. Birbaumer, A. Kübler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor. The thought translation device (ttd) for completely paralyzed patients. *Rehabilitation Engineering, IEEE Transactions on*, 8(2):190–193, 2000.

[125] N. Birbaumer. Slow Cortical Potentials: Plasticity, Operant Control, and Behavioral Effects. *The Neuroscientist*, 5(2):74–78, March 1999.

[126] M. Fabiani, G. Gratton, and Michael H. G. Coles. Event-related brain potentials. pages 53–84, 2000.

[127] M. Fabiani, G. Gratton, and E. Karis, D. andDonchn. Definition, identification and reliability of measurement of the p300 component of the event-related brain potential. *Advances in psychophysiology. GReenwich, CT:JAI Press;*, 1987.

[128] J. Polich. Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10):2128–2148, 2007.

[129] Ian Daly, Martin Billinger, Rafal Scherer, and Gernot Muller-Putz. On the automated removal of artifacts related to head movement from the eeg. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 21(3):427–434, 2013.

[130] Alois Schlögl, Claudia Keinrath, Doris Zimmermann, Reinhold Scherer, Robert Leeb, and Gert Pfurtscheller. A fully automated correction method of eog artifacts in eeg recordings. *Clinical neurophysiology*, 118(1):98–104, 2007.

[131] Martin Spüler, Wolfgang Rosenstiel, and Martin Bogdan. A fast feature selection method for high-dimensional meg bci data. In *Proceedings of the 5th Int. Brain-Computer Interface Conference, Graz, Austria*, pages 24–27, 2011.

[132] Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural engineering*, 4(2):R32, 2007.

[133] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012.

[134] G. Schalk, D. J. Mcfarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

[135] E. Donchin, K. M. Spencer, and R. Wijesinghe. The mental prosthesis: assessing the speed of a p300-based brain-computer interface. *Rehabilitation Engineering, IEEE Transactions on [see also IEEE Trans. on Neural Systems and Rehabilitation]*, 8(2):174–179, 2000.

[136] Adrian Furdea, Sebastian Halder, D.J. Krusienski, D. Bross, Femke Nijboer, Niels Birbaumer, and Andrea Kübler. An auditory oddball (p300) spelling system for brain-computer interfaces. *Psychophysiology*, 46, 2009.

[137] F. Nijboer, E. Sellers, J. Mellinger, M. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D. Krusienski, and T. Vaughan. A p300-based brain–computer interface for people with amyotrophic lateral sclerosis. *Clinical Neurophysiology*, 119(8):1909–1916, August 2008.

[138] Charles Kasiel Bliss. *Semantography (Blissymbolics): A Simple System of 100 Logical Pictorial Symbols, which Can be Operated and Read Like 1+ 2.* Semantography (Blissymbolics) Publications, 1978.

[139] Niels Birbaumer. Breaking the silence: brain–computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43(6):517–532, 2006.

[140] Martin Bax, Murray Goldstein, Peter Rosenbaum, Alan Leviton, Nigel Paneth, Bernard Dan, Bo Jacobsson, and Diane Damiano. Proposed definition and classification of cerebral palsy, april 2005. *Developmental Medicine & Child Neurology*, 47(08):571–576, 2005.

[141] Jackie Parkes, Melanie White-Koning, Heather O Dickinson, Ute Thyen, Catherine Arnaud, Eva Beckung, Jerome Fauconnier, Marco Marcelli, Vicki McManus, Susan I Michelsen, et al. Psychological problems in children with cerebral palsy: a cross-sectional european study. *Journal of Child Psychology and Psychiatry*, 49(4):405–413, 2008.

[142] Matthew J Bair, Rebecca L Robinson, Wayne Katon, and Kurt Kroenke. Depression and pain comorbidity: a literature review. *Archives of internal medicine*, 163(20):2433–2445, 2003.

[143] Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. No health without mental health. *The lancet*, 370(9590):859–877, 2007.

[144] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[145] L a Farwell, J M Martinerie, T R Bashore, P E Rapp, and P H Goddard. Optimal digital filters for long-latency components of the event-related brain potential. *Psychophysiology*, 30(3):306–15, May 1993.

[146] Andy Field. *Discovering Statistics Using SPSS (Introducing Statistical Methods series)*. Sage Publications Ltd, third edition edition, January 2009.

[147] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[148] John Platt et al. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods—support vector learning*, 3, 1999.

[149] John C Platt et al. Using analytic qp and sparseness to speed training of support vector machines. *Advances in neural information processing systems*, pages 557–563, 1999.

[150] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[151] Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C Weng. A note on platt's probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007.

[152] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[153] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[154] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31:1–38, 2004.

[155] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[156] Gernot Müller-Putz, Reinhold Scherer, Clemens Brunner, Robert Leeb, and Gert Pfurtscheller. Better than random: A closer look on bci results. *International Journal of Bioelectromagnetism*, 10(EPFL-ARTICLE-164768):52–55, 2008.

[157] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003.

[158] Etienne Combrisson and Karim Jerbi. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, 2015.

[159] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Instruction manual and affective ratings. *The center for research in psychophysiology, University of Florida*, 1999.

[160] Margaret M Bradley and Peter J Lang. The international affective digitized sounds (iads-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3*, 2007.

[161] Kazuhiko Takahashi. Comparison of emotion recognition methods from bio-potential signals. 40(2):90–98, 2004.

[162] Kazuhiko Takahashi. Remarks on svm-based emotion recognition from multi-modal biopotential signals. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pages 95–100. IEEE, 2004.

[163] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. *Multimedia content representation, classification and security*, pages 530–537, 2006.

[164] Guillaume Chanel, Joep JM Kierkels, Mohammad Soleymani, and Thierry Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607–627, 2009.

[165] Robert Horlings, Dragos Datcu, and Leon JM Rothkrantz. Emotion recognition using brain activity. In *Proceedings of the 9th international conference on computer systems and technologies and workshop for PhD students in computing*, page 6. ACM, 2008.

[166] Richard J Davidson. Cerebral asymmetry and emotion: Conceptual and methodological conundrums. *Cognition & Emotion*, 7(1):115–138, 1993.

[167] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1):18–31, 2012.

[168] Rishabh Gupta and Tiago H Falk. Affective state characterization based on electroencephalography graph-theoretic features. *7th Annual International IEEE EMBS Conference on Neural Engineering*, pages 577–580, 2015.

[169] Christian Mühl, Brendan Allison, Anton Nijholt, and Guillaume Chanel. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1(2):66–84, 2014.

[170] Sarunas J Raudys and Anil K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):252–264, 1991.

[171] Christian Mühl. *Toward affective brain-computer interfaces: exploring the neurophysiology of affect during human media interaction*. University of Twente, 2012.

[172] Anne-Marie Brouwer, Thorsten O Zander, Jan BF Van Erp, Johannes E Korteling, and Adelbert W Bronkhorst. Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in Neuroscience*, 9, 2015.

[173] David Watson, Lee A Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[174] Heinz W Krohne, Boris Egloff, Carl-Walter Kohlmann, and Anja Tausch. Untersuchungen mit einer deutschen version der “positive and negative affect schedule" (panas). *Diagnostica*, 1996.

[175] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 2010.

[176] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[177] Richard J Davidson. Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & Emotion*, 12(3):307–330, 1998.

[178] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. Motivated attention: Affect, activation, and action. *Attention and orienting: Sensory and motivational processes*, pages 97–135, 1997.

[179] Steven A Hillyard and Marta Kutas. Electrophysiology of cognitive processing. *Annual review of psychology*, 34(1):33–61, 1983.

[180] Richard J Davidson, Paul Ekman, Clifford D Saron, Joseph A Senulis, and Wallace V Friesen. Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology: I. *Journal of personality and social psychology*, 58(2):330, 1990.

[181] Charles T Leonard, Toshio Moritani, Helga Hirschfeld, and Hans Forssberg. Deficits in reciprocal inhibition of children with cerebral palsy as revealed by h reflex testing. *Developmental Medicine & Child Neurology*, 32(11):974–984, 1990.

[182] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[183] Teresa Farroni, Mark H Johnson, Margaret Brockbank, and Francesca Simion. Infants’ use of gaze direction to cue attention: The importance of perceived motion. *Visual cognition*, 7(6):705–718, 2000.

[184] Michael C Frank, Edward Vul, and Scott P Johnson. Development of infants’ attention to faces during the first year. *Cognition*, 110(2):160–170, 2009.

[185] Richard J Davidson and Nathan A Fox. Cerebral asymmetry and emotion: Developmental and individual differences. 1988.

[186] Robert L Maulsby. An illustration of emotionally evoked theta rhythm in infancy: Hedonic hypersynchrony. *Electroencephalography and Clinical Neurophysiology*, 31(2):157–165, 1971.

[187] Benjamin Blankertz, Klaus-Robert Müller, Gabriel Curio, Theresa M Vaughan, Gerwin Schalk, Jonathan R Wolpaw, Alois Schlögl, Christa Neuper, Gert Pfurtscheller, Thilo Hinterberger, et al. The bci competition 2003: progress and perspectives in detection and discrimination of eeg single trials. *Biomedical Engineering, IEEE Transactions on*, 51(6):1044–1051, 2004.

[188] Benjamin Blankertz, Klaus-Robert Müller, Dean J Krusienski, Gerwin Schalk, Jonathan R Wolpaw, Alois Schlögl, Gert Pfurtscheller, Jd R Millan, Michael Schröder, and Niels Birbaumer. The bci competition iii: Validating alternative approaches to actual bci problems. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 14(2):153–159, 2006.

[189] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits, 1998.

[190] Fabien Ringeval, Bjoern Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 1335–1336. ACM, 2015.

[191] Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models. *Analytica Chimica Acta*, 760:25 – 33, 2013.

[192] I Daly, F Aloise, P Aricó, J Belda, M Billinger, E Bolinger, F Cincotti, D T Hettich, M Iosa, J Laparra-Hernández, R Scherer, and G Müller-Putz. Rapid prototyping for hbci users with cerebral palsy. *Proceedings of the 5th International Brain-Computer Interface Meeting*, 2013.

[193] R Scherer, M Billinger, J Wagner, A Schwarz, D T Hettich, E Bolinger, M Lloria Garcia, J Navarro, and G Müller-Putz. Thought-based row-column scanning communication board for individuals with cerebral palsy. *Annals of Physical and Rehabilitation Medicine*, 2015.

[194] Mohammad Shami and Werner Verhelst. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49(3):201–212, 2007.

[195] Daniel M Messinger, Paul Ruvolo, Naomi V Ekas, and Alan Fogel. Applying machine learning to infant interaction: The development is in the details. *Neural Networks*, 23(8):1004–1016, 2010.

[196] Wenyu Jiang and Richard Simon. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29):5320–5334, 2007.

[197] Wenyu Jiang, Sudhir Varma, and Richard Simon. Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.