UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Music-Emotion: Towards automated real-time recognition of affective states with a wearable Brain-Computer Interface

Michele Romani

M.Sc. Thesis in Human-Computer Interaction & Design
February 2022

Supervisors:
dr. M. Poel
Data Management & Biometrics Group

dr. ir. D. Reidsma
Human Media Interaction Group

Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Preface

In 2019, pushed by the continuous unsatisfactory feelings that haunted me, I quit my consulting job as software developer and abandoned a stable situation to pursue a Master's degree in Human-Computer Interaction. I had the strong desire to expand my knowledge and unleash my creativity, but I also wanted to dedicate my efforts to something I truly cared about, that could be meaningful for me and for others. When I stumbled upon Brain-Computer Interfaces, and later Affective Computing, something clicked. I could finally draw a line connecting my technical background in Computer Science with my interests in humanistic subjects like philosophy, psychology and human learning - in other words: Cognitive Science, the study of the mind and its processes using technological means. The passionate people I met in these two years gave an essential contribution in shaping the direction of my studies, and finally a fortuitous encounter on a flight from Paris to Milan in January 2020 set the basis for what later became my graduation project, the core of this research.

The last decade has seen a wave of renewed attention to the individual needs of people, from the fundamental ones like health and education, up to hobbies, creativity and passions. I think we struggle to better understand and take care of ourselves, because it is inherently hard to keep control over our mind and body. Funnelling our energies on what we think really matters feels tiring, and often we push ourselves over limits we are not aware of, with critical risks for our mental health. Like many others, I have always seen computers as an extension of the human brain, not as a substitute tool for it. In my vision, technologies like Artificial Intelligence are not here to replace humans, but to augment human intellect and help people express their true potentials by taking away part of the effort we would need to put on boring or hard tasks. Technology needs the capability understand us so we can use it to better shift the focus on ourselves, and this is the great challenge that Affective Computing and Brain-Computer Interfaces can help us facing in the years to come, possibly disrupting society like many other great technological innovations did in the past. Such epochal changes are unpredictable and frightening, but consciously embracing them in advance will reduce the risk of collateral damage caused by misusing technology. With this project, I took my first step into these innovative fields and I

am determined to responsibly design technologies that can improve each individual's life and, consequently, society itself in the years to come.

I need to acknowledge many people for their direct or indirect contributions that made this project possible. First of all my expanded family, including relatives and close friends, that were always supportive and fueled me with love regardless the physical distance. A special thanks goes to Mannes Poel, that guided my learning process for more than a year and got me passionate about BCI. Another special thanks goes to the crew of the Innovation Lab at myBrain Technologies: Giuseppe Spinelli, whose random encounter on a flight created this beautiful opportunity, Xavier Navarro-Sune, that weekly mentored and reviewed my progresses together with Giuseppe, and Yohan Attal that always found the time to share insightful ideas and comments despite being busy in running a company. I also wanna thank all the other colleagues at myBrainTechnologies that welcomed me and helped in many organizational steps. Finally, heartfelt thanks to my university colleagues and friends, from Université Paris-Saclay and University of Twente, because in the worst moments we stayed together and cheered each other up, and in the best moments we shared our passions and enjoyed our adventures with a light mind as young people should always do.

I wish you a good reading,

Michele

"Il corpo faccia ciò che vuole, io sono la mente." - R.L. Montalcini

Summary

This research set out to investigate the feasibility of performing Emotion-Recognition using Melomind, a wearable neural interface manufactured by myBrainTechnologies. Melomind is capable of recording EEG signals, that can be processed using machine learning algorithms in the form of a classification task of the emotional dimensions of valence and arousal.

This study introduces the fields of Brain-Computer Interfaces and Affective Computing, the perception of the market, the leading companies producing wearable neural devices for non-clinical applications and the relevance of studying emotions using music, from both the perspectives of market demand and enhancing the user experience.

The goal of this research was to evaluate Melomind's capabilities for a future real-time application that can be used to perform Emotion-Recognition. In order to do so, the Valence-Arousal model by James Russel was used as metric for the dimensions of emotions, then several models of emotional correlates in brain activity were evaluated to define what features of the EEG would be more suitable for the task.

The relevant related work was reviewed and studied to provide a methodological framework for the machine learning task that could be adapted to the constraints imposed by the limited hardware of the Melomind. An experimental protocol was designed around the inherent advantages of wearable technologies to collect a dataset with continuous labelling of emotions on the Valence-Arousal coordinate system. Possible biases caused by listening conditions, data labelling tools, emotional interference, multiple cognitive tasks and external factors were taken into account and the protocol was tested during a pilot week with employees of myBrainTechnologies prior to the real experiment.

Data were collected using a robust protocol in two different conditions for music listening: eyes-open with a labelling task and eyes-closed solely. Data were then processed using a lightweight automated preprocessing pipeline and two types of features were extracted from the Power Spectra Density of the EEG signal: neuromarkers and frequency-band specific spectral properties calculated in the Theta, Alpha and Beta bands of the EEG signal. Features dimensionality was reduced

through features extraction using Principal Component Analysis and the classification task was performed with subject-dependent strategy. The problem was simplified into two separate binary classifications tasks for valence and arousal, and two supervised learning algorithms were tested: Support-Vector Machines and Multi-Layer Perceptron. The hyper-parameters were tuned using GridSearch to select the configuration that yielded the highest Matthews Correlation Coefficient score for each participant, a coefficient that is gaining popularity in machine learning research thanks to its higher reliability.

All models were then trained and tested using 5-fold leave-one-block-out cross-validation that produced two cross-validated scores on the training datasets: CV accuracy and CV MCC. Then, models were further tested on a completely unseen split of data that produced two more scores: test accuracy and MCC. Results were collected and the two classification methods were compared with each other and then with the comparable related work.

Some models showed promising classification results, reaching 80% accuracy in arousal classification and 75% accuracy in valence classification with both SVM and MLP. MCC scores confirmed an average positive learning capability of the models, although many models ended up overfitting or underfitting. The average classification results did not meet the initial expectations and are below many of the related studies, suggesting that the adopted lightweight pre-processing, the limited hardware of the Melomind or a combination of both are hindering the classification task and are not yet suitable for real-time Emotion-Recognition.

The final discussion covers the current challenges of real-time Emotion-Recognition reported by this and related studies and delves into possible improvement of the emotional self-reporting, the features selection, the artifacts cleaning process and the requirements to move from subject-dependent classification to subject independent-classification.

In the conclusion, some considerations are raised from answering the research questions and then an improved artifact cleaning approach is recommended for a follow-up study using the same dataset, that could give further insights on the development of a wearable affective Brain-Computer Interface using Melomind.

Contents

List of acronyms

AC	Affective Computing
AI	Artificial Intelligence
ASR	Artifact Subspace Reconstruction
AuPP	Automated Pre-processing Pipeline
AWI	Approach-Withdrawal Index
BCI	Brain-Computer Interface
BCIs	Brain-Computer Interfaces
BVP	blood-volume pressure
CAR	common average reference
CNN	Convolutional Neural Networks
CNS	central nervous system
CV	cross-validated
CV MCC	cross-validated MCC
DFT	Discrete Fourier Transform
EA	Experimental Annotation
EC	eyes-closed
ECG	Electrocardiography
EDA	electro-dermal activity
EEG	Electroencephalography
EMG	Electromyography

EO eyes-open

EOG Electrooculography

EP evoked potentials

ER Emotion-Recognition

ERPs event-related potentials

EVI Emotion Valence Index

FD Fractal Dimension

FFT Fast Fourier Transform

FMTI Frontal Midline Theta Index

fMRI Functional Magnetic Resonance Imaging

fNIRS Functional Near-Infrared Spectroscopy

GBDT Gradient-Boosting Decision Tree

GSR Galvanic Skin Response

HR heart rate

ICA Independent Component Analysis

LOBO leave-one-block-out

LOSO leave-one-subject-out

MCA Multimedia Content Analysis

MCC Matthews Correlation Coefficient

MEG Magnetoencephalography

MLP Multi-Layer Perceptron

PANAS Positive And Negative Affect Scales

PCA Principal Component Analysis

PNS peripheral nervous system

PPG photoplethysmogram

PSD Power Spectra Density

QC Quality Checker

QIRem Quality Index Removal

SAM Self-Assessment Mannikin

SASI Spectral Asymmetry Index

SPT SignalProcessingToolbox

SFS sequential features selection

STFT Short-Time Fourier Transform

SVM Support Vector Machine

VA valence-arousal

Chapter 1

Introduction

The evolution of technology is inherently bound to the evolution of society and human desires. In recent years the focus of the technology-mediated services has shifted from mere functionalism to become more aesthetically, functionally, socially, and interactively pleasurable. The most successful multimedia creation and distribution companies offer customized recommending services and then aggregate correct predictions between users sharing similar taste or preferences to improve their offer: an experience as tailored as possible to users' individual needs. Understanding users' behavior and emotions is not only very profitable for companies that want to continuously engage their users, but also a popular topic among researchers and designers that thrive to better understand the human mind to enhance the quality of human-computer interactions. It is also becoming a necessity for the end users themselves, who are not satisfied anymore by tinkering with technology but want the interaction to be flexible and seamlessly usable in the daily life. Recent applications and services offer the possibility for people to monitor their body, mind, and health through continuous collection of physiological signals from wearable sensors, for example to keep track of good sport habits, sleep quality, stress level and more. But it is also possible to infer affective states from clues in the recorded brain activity. Given the increasing interest of researchers and companies in the affective field, the more and more frequent use of physiological and behavioral clues to assess mental states will keep growing until technologies of daily use will be standardly designed with brain-reading capabilities. The human brain is the central and most important organ of our body because it is where our consciousness, our "self", resides and it is the command center of all vital functions. And yet, it is also the one we understand least, despite the ongoing research. We rightfully assume emotions originate in the brain, but we can only observe their physiological responses and we can only qualitatively support these observations through inherently imprecise self-assessment tools. Trying to find a correlation between the physiological response and the self-assessed mental state or emotion is not as simple as correlating fac-

tual measures, because of the uncertainty of the factors involved. Modern wearable Brain-Computer Interfaces (BCIs), mostly represented by EEG-capable devices, are still limited in their functionalities and design. Recorded signals are often affected by noise or artifactual information and the user experience is so heavily hindered that most companies are reluctant in investing into them, and more research and optimization is needed before pushing them to the general public. Self-assessing emotions is also non-trivial because it requires a strong understanding of one's perceived emotions, and this perception has great variability from individual to individual. Creating models able to generalize through all the subjective differences is complicated, especially with performances that enable designers to create enjoyable user experiences for everyone. Thus, we enter in a challenging and almost paradoxical situation: on one hand the goal is to find a common approach to exploit generic behavioral and physiological patterns, on the other hand it is also necessary to account for individual differences to offer the customized experience that users desire. This research takes an extra step into the challenge by evaluating what could be the classification performances of an affective Brain-Computer Interface (BCI) system for emotion-aware recommendations using a wearable device. It delves on relevant insights on the main problematics that researchers and designers will face in the years to come when classifying brain emotions, and finally discusses possible solutions and future developments towards online Emotion-Recognition.

1.1 Motivation

In 2016, the American research and advisory information firm Gartner published their yearly hype cycle for emerging technologies , positioning "Brain-Computer Interfaces" and "Affective Computing" on the growing slope of the "Peak of Inflated Expectations", both with an estimated period of about 10 or more years before mainstream adoption (see Figure 1.1). In the versions published over the last 5 years, "Affective Computing" completely disappeared from the hype cycle and "Brain-Computer Interfaces" slowly climbed the slope before disappearing as well. This trend suggests that both fields reached the peak of the inflated expectations and fell down the "Through of Disillusionment" slope as the technology failed to meet the expectations of users, researchers, and investors. Nevertheless, some innovative companies and startups stepped up to the challenge and started a new innovation cycle by developing wearable neural sensors and slowly pushing again BCI towards mainstream adoption. For example, NextMind[1] released in early 2021 a wearable sensor for active control of multimedia applications and games. Muse[2] instead developed sev-

[1] https://www.next-mind.com/

[2] https://choosemuse.com/

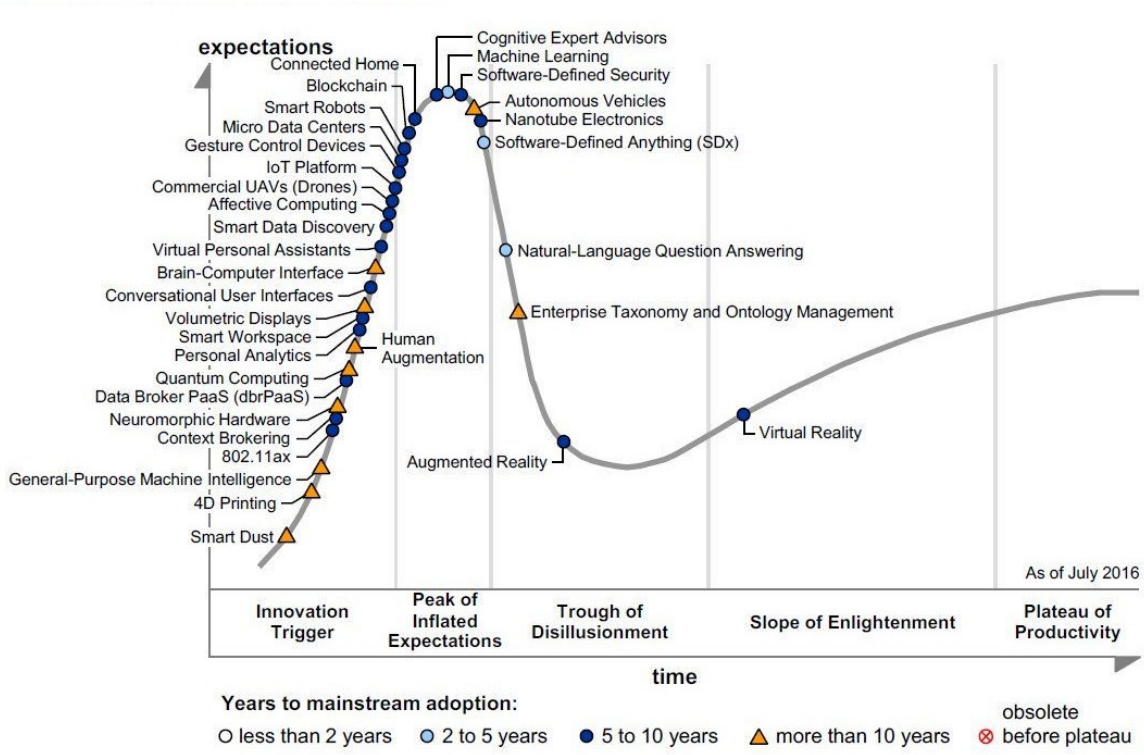

Figure 1.1: Gartner hype cycle for emerging technologies in 2016, from Gartner[1]

eral wearable BCIs for neurofeedback during guided meditations, recently updated to be used for sleep tracking. Enophone produces EEG-capable headphones and recently made a partnership with BrainFlow to integrate their SDK and build affective applications for music listening [2]. Melomind[3] , the device used in the current research, is another EEG capable device with headphones and an application for neurofeedback training. Ontbo[4] already promises an application with their headphones to generate music playlists that can alter the level of motivation, relaxation, stress, and concentration in the brain activity. Major brands like Valve, Tobii and OpenBCI are collaborating to bring together the world of virtual reality and BCI [3], and Facebook is developing their own neural interfaces after acquiring the neurotech startup CTRL-Labs [4]. In the academic world, we can find very recent papers that focused on affective music recommendations, for example Chang et al. [5] proposed a recommending system to suggest the appropriate stress-relief music to the users based on the inferred stress level in their EEG, while Abdul et al. [6] designed an emotion-aware system to correlate implicit emotional user tags and musical features. This reignited interest in affective applications and the development of a new generation of BCIs supports the relevance in researching and developing now the methods and tools that will be used to design the affective systems of tomorrow. Given this

[3]https://www.melomind.com/en/product/melomind-en/
[4]https://ontbo.com/en/

context, music is one of the best elicitors for the field of Emotion-Recognition for a multitude of reasons:

- It is a proven powerful elicitor that arouses instinctive physiological reactions in the human body.
- It has been used for centuries to convey emotional meaning, and now more than ever even across cultures, social classes, and age groups.
- The physiological emotional response is agnostic of the stimulus, thus findings on music-elicited Emotion-Recognition are theoretically transferrable to other applications.
- The market of music recommending systems is flourishing and very competitive, with many companies taking part in the technological development of such systems.

The music-emotion experience is very personal and influenced by internal factors, such as tempo and pitch of a song, as well as external factors such as memories, context, or correlated events associated to a past pleasurable or unpleasurable experience. The proposal of this study is a novel approach to the Emotion-Recognition task using Melomind, a wearable and consumer-oriented BCI. The goal is to investigate the feasibility and the performances of the classification task in realistic listening conditions to evaluate the future development of wearable BCIs equipped with online Emotion-Recognition for daily use. Instead of focusing on the correlation between the musical features and the event-related potentials (ERPs) in the brain, the approach of this research is to use emotion-labelled songs as elicitors to study the spectral properties of frequency bands associated with emotions in the Electroencephalography (EEG) signal. These spectral properties can be translated into features and used for the computation of neural biomarkers, called "neuromarkers" in this research, to be fed into a classifier for the classification of the emotional valence and arousal based on known patterns in the brain activity. To support the collected physiological data, participants have been asked to continuously self-assess their emotions in a coordinate system representing the emotional dimensions of valence and arousal. The research questions of this study try to fulfil the design requirements of exploring the performances of the Emotion-Recognition task using a wearable EEG headset, considering the disadvantages and the advantages of this specific technology. The dry electrodes of the Melomind in pair with the headphones form factor allow for a very quick and relatively comfortable setup, enabling the researcher to focus on the task and overall shorten the experimental sessions. Consequently, it was possible for a single researcher to collect data from 45 subjects over 15 days. This technology has also limited recording capabilities;

thus, the quality and quantity of data is lower than what could be obtained with standard EEG lab equipment featuring 32 wet electrodes headcaps. Furthermore, the position of the Melomind electrodes only allows recording signals from the frontal and/or the parietal regions of the cerebral cortex, limiting considerably the area of study. The data has been collected through an experimental phase as result of a collaboration between the University of Twente and myBrainTechnologies, the company that manufactures Melomind. The research was approved by the Ethics Committee Computer & Information Science and the Dean of the EEMCS faculty following the regulations in force at the University of Twente, with reference number RP 2021-43.

1.2 Research questions

To evaluate this novel approach, this study aims to answer the following main research question:

RQ: *"What are the accuracy and MCC scores of subject-dependent classification of music-elicited emotional valence and arousal in the EEG signal using SVM and MLP algorithms with Melomind?"*

The mains research question was then extended by the following sub questions to support possible design choices for a real-time music recommending system based on brain activity.

SRQ1: *"What are the most relevant selected Power Spectral Density features to perform the Emotion-Recognition using SVM and MLP algorithms with Melomind?"*

SRQ2: *"What is the best classification strategy applicable to the current software and hardware capabilities of Melomind using SVM and MLP algorithms?"*

1.3 Report organization

To answer these research questions, the main models for self-assessment of emotions have been studied, together with the existing models relating the two main dimensions of emotional valence and arousal to brain activity (see Chapter 2). The most compelling related work in classification of music-elicited emotions has been reviewed (see Chapter 3) and used as methodical foundation. An experiment was designed to collect data in two listening conditions and a processing pipeline was implemented to extract features and to train classification algorithms that could pro-

duce comparable results with the related work (see Chapter 4). The results were then collected (see Chapter 5) and discussed with a view to possible future developments (see Chapter 6).

Chapter 2

Background

In this chapter some background on Affective Computing and Brain-Computer Interfaces is provided. Then the circumplex model of affect is introduced together with the models for emotional correlates in the brain. Finally, the neuromarkers used as principal features for classification are presented.

2.1 Affective Computing

All technologies that support the expression and the processing of human affective behaviours fall under the name of Affective Computing (AC), a relatively new branch of computer science named after Rosalind Picard's work [7], that thoroughly described the practical methods and the ethical implications of building computers that can understand and express human emotions through the processing of behavioural, physiological, or conversational data. Building a very sophisticated Artificial Intelligence (AI) that mimics the human behaviour and can understand and replicate human emotions is still very far from the current state of art technology. Yet, the idea raises compelling questions on how we could interact with such entities and opens many the ethical implications that are valid already for the intelligent systems being built nowadays. In fact, concrete applications able to perform Emotion-Recognition tasks are already on the market and in recent years have become a matter of great interest for researchers working in both academia and the industry. Emotion-Recognition (ER) is the task of recognizing human emotions by inferring them from different clues in the data. For example, from metadata collected from the usage of a software system; from data collected using wearable sensors such as accelerometers and gyroscopes; from photos and videos of facial expressions using computer vision; from text and voice samples processed using natural language processing techniques; from physiological measurements such as heart rate, dermal activity, and of course also brain activity.

The rapidly growing branch of AI, mostly represented by machine learning and deep learning algorithms, further fuels the development and the improvement of systems that can perform ER, by offering powerful techniques that can leverage the quantity of data to build fast and reliable systems more suitable for real-time tasks. Because of the nature of these data, it is possible to infer very sensitive behavioural and affective information from the users, exploitable by companies for commercial uses and marketing proposals, evoking the unpleasant dreads of an Orwellian society where powerful corporations know exactly what people are thinking, feeling, desiring, or fearing and manipulate their emotions for "evil" purposes. Cases like the "Facebook emotional manipulation study" [8] already demonstrate the interest and the ease for companies in inferring and manipulating their users' emotions and getting away with little or no consequences. Thus, philosophers and scholars already advocate for researchers and designers to design technology in a socially responsible behavior perspective [9], so that users and companies are not tempted to misuse technology but rather use it to improve society. In the context of this research, a company building an intelligent system that can offer affective user experiences must ward the users' control over their data and utilize these data with consensus and in respect of privacy laws, for example with anonymization of sensitive information and transparent guidelines to make the users aware on how, where and when their data will be used. Even a music recommending system can raise serious ethical concerns: social functionalities might reveal sensitive information of a user to their network of friends, or groups of users might be targeted by affective promotional advertisement that has a negative impact on their emotional state.

In 2003, Picard also addressed with several criticisms the main challenges [10] faced by designers engaged in building machines with affective abilities. Some are relevant for the current study, in particular regarding the ability of sensing and recognizing emotions: Picard argued that the range of means, and modalities of emotion expression is very broad and hardly accessible, unlikely to be feasible in the near future. Another similar criticism regarded the accuracy of recognizing an individual's emotional state from the available data and the difficulty in the articulation and assessment of one's own feelings. While these criticisms still hold true, the technological developments of the last two decades of wearable sensing devices, including wearable BCIs, greatly contributed to the collection of data that can be leveraged for affective computation, probably beyond Picard's expectations. Brain activity, blood volume pressure, movement recorded through accelerometers, and heart rate have been used in several experiments for the ER task and it has been proven possible by all these means, with different degrees of precision.The most relevant criticism probably regards the cognitive modeling of affective data. The progresses in psychological interpretation of idiosyncratic processes that characterize emotional

responses are little and often supported by data collected in highly artificial lab environments. Multiple models for emotion assessments coexist in the emotion research community and there is still disagreement on what type of mechanisms mediate the effects of emotion. These models inevitably represent stylized stereotypes of emotional responsiveness and do not exactly correspond to the behavior and feelings in real people. While this study does not delve in evaluating the goodness of the emotional models, the use of a wearable device and the automated processing of data towards the realization of an online system are an attempt to simulate less artificial conditions. Further discussion about the subjective experiences of the participants (see Chapter 6) gives further insights on the difficulties that arise in building affective systems from self-reported emotional responses.

2.2 Wearable Brain-Computer Interfaces

A BCI often referred to as brain-machine interface, is a system that creates a pathway of direct communication between a brain and a computer. Research in BCI dates back to 1973 when Jacques Vidal named the field in his paper "Towards Brain-Computer communication", and since then many technologies have been used to build this "bridge" between the human brain and machines; in their overview paper, Nicolas-Alonso and Gomez-Gil provide a good summary of the state-of-the-art BCIs [11]. In short, a first categorization can be made between invasive or partially invasive BCIs, implanted directly in the brain or on the skull, and non-invasive BCIs that can be easily placed on the scalp. The main advantage of the first two categories resides in the greater amount and quality of data that is possible to collect. However, surgeries to implant electrodes can have negative outcomes for the subject including the formation of scar tissues, rejection of the electrodes or even worse infection. Consequently, invasive BCIs are now mostly used in the experimental medical field where there is a necessity for a high temporal and spatial resolution to treat the patient's conditions. Non-invasive BCIs instead often sacrifice spatial resolution and cannot effectively capture high frequencies in the brain signal due to the dispersion caused by the skull thickness. In addition, they are affected by a higher presence of artefacts caused by environmental factors and muscular movements. Nevertheless, the easy and safe setup made non-invasive BCIs the preferred choice for researchers and now the majority of published BCIs work involves this type of BCIs. The main technologies for non-invasive BCIs are the following:

- **EEG**: can record electrical brain activity from the scalp in the form of time series. Direct brain activity with good temporal resolution and bad spatial resolution.

- **Functional Magnetic Resonance Imaging (fMRI)**: can record brain activity by detecting changes in the blood flow. Structural brain activity with bad temporal resolution but good spatial resolution.
- **Functional Near-Infrared Spectroscopy (fNIRS)**: can record brain activity based on neuro-vascular coupling. Indirect brain activity through oxygenation, bad temporal resolution, and good spatial resolution.
- **Magnetoencephalography (MEG)**: can record brain activity through the magnetic fields produced by the electrical currents in the brain. Direct brain activity with very good temporal resolution and spatial resolution is slightly better than EEG.

Among these technologies, EEG has the best cost/capabilities compromise and eventually became the most popular for BCIresearchers to study evoked potentials (EP) and ERPs, thanks to the relatively cheap cost compared to MEG, the good temporal resolution compared to fMRI and fNIRS and the good support provided by standardized software libraries for recording, streaming, and processing data like MNE[1] , EEGLab[2] , OpenVibe[3] and LabStreamingLayer[4]. Except for some very rare cases, most of the portable BCIs are based on EEG.

Figure 2.1: An example of standard EEG headset

Standard EEG-based BCIs require the subjects to wear a head-cap (Fig. 2.1) with 16, 32 or 64 electrodes placed over the scalp following the standard 10-20 system[5]. These electrodes usually require the displacing of conductive gel to obtain a

[1] https://mne.tools
[2] https://sccn.ucsd.edu/eeglab/index.php
[3] http://openvibe.inria.fr/
[4] https://github.com/sccn/labstreaminglayer
[5] https://www.evokedpotential.com/international-10-20-system.html

stable and qualitative signal. While this type of setup is acceptable for researchers in experimental environments, it is clearly impossible to imagine a daily adoption from users from an usability perspective. This problem is nowadays partly addressed by wearable EEG-capable headsets like Emotiv Epoc, Neurosky Mindwave, Muse Headband, NextMind, Melomind (Fig. 2.2) and many others, that mostly feature EEG capabilities with 2 up to 16 soft dry electrodes, and do not require conductive gel to capture the electric signal of the brain. While the quality of the signal is not usually as good and complete as the standard EEG-based devices, the setup of these portable BCIs is seamless and often requires just a simple calibration making them suitable for both researchers and consumers. The experimental and analytical phases of this study have been conducted as part of a research project funded by myBrainTechnologies, a startup based in Paris that designed Melomind, a BCI that includes a real-time auditory neurofeedback application to induce a relaxation state. Apart from its designed purpose, Melomind is a fully featured EEG-capable headset that comes in two versions, standard (Fig. 2.2a) and Q+ (Fig. 2.2b).

(a) *(b)*

Figure 2.2: a) Melomind with 2 dry electrodes (on the flexible antennas), and 2 textile electrodes on the cushions.
b) Melomind Q+ with 4 dry electrodes (on the flexible antennas), and 2 textile electrodes on the cushions.

The standard version used in this experiment features Bluetooth headphones with 2 textile reference electrodes and 2 dry electrodes for recording that can be placed on the parietal area in correspondence of the P3/P4 electrodes or in the frontal area in correspondence of the AF3/AF4 electrodes on the 10-20 system. The Q+ version (Fig. 2.2.b) was still under development at the time of the experiment, and it is identical to the standard version except that it features 4 dry electrodes for

simultaneous recording on the frontal and parietal areas, an accelerometer, and a photoplethysmogram (PPG) sensor to capture heart rate. Melomind records EEG signals at a sampling rate of 250 Hz and the acquisition application uses an algorithm called *QualityChecker* to help the researchers visually estimate the quality of the recorded signal in real-time. Both the EEG signal and the quality values are saved together with other metadata in a *.json* file at the end of the recording. The peculiar design of Melomind allows it to play music with very good sound quality while simultaneously recording EEG signals, thus making it a good choice for experiments and applications based on auditory stimuli.

2.3 A circumplex model of affect

Defining what exactly is an emotion or how to measure emotions is non-trivial and has been a matter of several studies by psychologists and cognitive scientists over the last century. Many explanations and definitions have been given resulting in a wide plethora of emotional models, sometimes discordant with each other. Recently, the circumplex model of affect [12] proposed by James Russel in 1980, often referred to as the valence-arousal (VA) model (Fig. 2.3), has become very popular due to its simplicity yet its efficiency in representing emotions in a 2-dimensional coordinate system. In the VA model, the valence of an emotion intended as a spec-

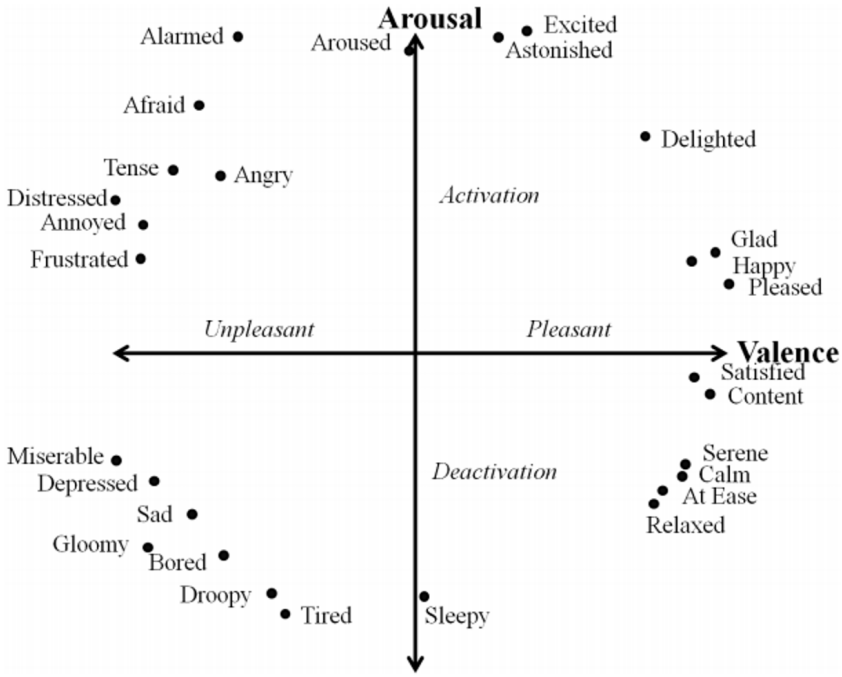

Figure 2.3: Valence-Arousal model taken from J.Russell [12]

trum between an unpleasant and a pleasant feeling represented on the X-axis of a Cartesian coordinate system, while the arousal of emotion intended as the physiological and psychological level of activity is represented on the Y-axis of the coordinate system. In the 4 quadrants originated by the intersection of the valence and arousal axes, Russel positioned a total of 28 basic and complex emotions through an experiment involving hundreds of participants that produced a similarity matrix geometrically representing the relations between these 28 words. While the subjective perception of where emotions should be placed might differ from the model's labelling, the VA model is still widely used because it simplifies the visualization of these emotional relations and can be easily integrated into tools for self-assessment of emotions. Russel's model is far from being perfect and other models tried to address some of the limitations, for example the Self-Assessment Mannikin (SAM) [13] allows the self-assessment of emotions in a 3-dimensional scale defined by valence, arousal and dominance while the Positive And Negative Affect Scales (PANAS) [14] were developed to reflect the extent to which a person feels enthusiastic, active, and alert using two 10-item mood scales. However, the VA model remains a popular choice among researchers for the classification of music-elicited emotions, and, for this reason and the previously mentioned simplicity, it was chosen as main emotion self-assessment tool. Some of its limitations are also discussed in the scope of this study (see Chapter 6).

2.4 Models of emotional correlates in brain activity

The connection between music and emotions has been a matter of great interest for researchers as music has shown its great potential for evoking a wide variety of emotions. Musicians and composers intuitively know from their experience what combination of notes, chords, tempo, or pitch is likely to evoke a specific emotion and this information is valuable both in terms of artistic expression and commercial use. Correlates between emotions and the physiological effects elicited on the human body have been studied thoroughly with the intent to define a standard model for the analysis and processing of emotional information. In 2001, Schmidt and Trainor [15] conveniently reviewed the main models for analysis of emotional valence and arousal (referred to as intensity) in the brain activity measured through EEG. The first model reviewed is from Davidson [16], which introduced the concept of organizing emotions around approach-avoidance tendencies in the brain. According to Davidson's model, emotions are differentially lateralized in the frontal region of the brain. In particular, the left frontal area is involved in the experience of positive emotions and the right frontal region is involved in the experience of negative emotions. The second model reviewed, brought up by Dawson [17] and Schmidt [18], states

that the pattern of absolute activation in the frontal region may reflect the intensity (arousal) of the affective experience. They supported this model with numerous studies that brought data evidence of such patterns. The third model, proposed by Heller [18], considers both dimensions and argues that frontal and right parietal-temporal regions are involved in the processing of emotions. According to his model, the frontal region modulates the emotional valence similarly to the model proposed by Davidson. Heller further states that the right parietal region is involved in the modulation of autonomic and behavioural arousal, i.e. higher levels of parietal-temporal activity are associated with high levels of arousal. In summary, these studies bring evidence that asymmetrical frontal EEG activity may reflect the valence of emotion experienced, while frontal and parietal absolute activation may reflect the intensity of the emotional experience. The resulting models lay the foundations for the analysis of emotions regardless of the stimulus; the application of these models to the analysis of music-elicited emotions is further discussed (see Chapter 3).

2.5 Spectral features and emotional neuromarkers

The power spectrum of the EEG signal can be decomposed and classified into five main frequency bands associated to specific cognitive processes:

- **Delta waves** (0.5hz to 3hz): associated to deepest meditation states and dreamless sleep.
- **Theta Waves** (3hz to 8hz): associated to learning, working memory and intuition.
- **Alpha Waves** (8hz to 12hz): the most prominent in brain activity, it is associated with the resting state of the brain, as well as coordination and alertness.
- **Beta Waves** (12hz to 33hz): another commonly found wave in the brain activity, it is associated with active cognitive tasks, attention, and movement.
- **Gamma Waves** (25hz to 100h): associated to high level cognitive processing tasks, senses, and perceptions.

For this study, the most sensitive frequency bands for the study of emotions, according to literature, were considered: Theta, Alpha and Beta. These frequency bands have been often encountered in the analysis of music-elicited emotions and used in the ER task. The approach of this study was to calculate bio-markers from the brain activity, referred to as “neuromarkers”, to represent differential and rational measurements supported by the models of emotional correlates. These neuromarkers are a

convenient mathematical expression of the theories of frontal asymmetrical activity and regional absolute activation in the brain activity, and some equivalent version can be found in the related studies with similar or different names. The computed neuromarkers are the following:

- **Approach-Withdrawal Index (AWI)**: asymmetry index representing the approach-avoidance tendencies of Alpha waves in the frontal area of the brain and computed as difference of Alpha power between the two frontal electrodes after normalization against the baseline using the following formula

$$AW = (Pow_{\alpha}AF4 - Pow_{\alpha}AF3)$$

- **Frontal Midline Theta Index (FMTI)**: index representing absolute increasing/decreasing activation in the Theta activity on the Fz channel compared to the baseline period, it is inferred between the frontal electrodes after normalization against the baseline using the following formula

$$FMT = Mean(Pow_{\theta}AF4, Pow_{\theta}AF3)$$

- **Spectral Asymmetry Index (SASI)**: index representing the balance of Theta and Beta frequency band power, computed on both frontal electrodes after normalization against the baseline using the following formula

$$SASI(AF3) = \frac{(Pow_{\beta}AF3 - Pow_{\theta}AF3)}{(Pow_{\beta}AF3 + Pow_{\theta}AF3)}$$

$$SASI(AF4) = \frac{(Pow_{\beta}AF4 - Pow_{\theta}AF4)}{(Pow_{\beta}AF4 + Pow_{\theta}AF4)}$$

AWI and SASI, or equivalent measurements, have been reported to reflect the emotional valence [15], [19]. FMTI, or equivalent measurements, has been reported to reflect the level of appreciation [20], the emotional valence [21] and the arousal dimension of emotions [22]. A multitude of reviewed studies [23], [24], [25], [26], [27] (see Chapter 3) make use of equivalent measurements as features for classification of emotional valence and arousal. Zhao et al. [22] investigated the use of asymmetries and mid-line absolute power in the Theta, Alpha and Beta frequency bands to classify discrete emotion within the same emotional spectrum and were able to successfully classify four emotions with appreciable precision. To strengthen the effectiveness in the classification of emotional valence and arousal, the neuromarkers were supported by an extra set of features of the EEG signal, extracted in the same frequency bands of interest: raw power, skewness, kurtosis, standard deviation, ratio, and relative spectral difference (see Chapter 4).

Chapter 3

Related work

In this chapter, the most relevant studies on the classification of music-elicited emotions are reviewed. In 2013 T. Eerola and J. Vuoskoski [28] reviewed and categorized 251 studies related to music and emotions in terms of approaches, emotion models and stimuli. However, most of them are not comparable in terms of methods with the current study and the list is not updated with the most recent findings in the field of ER. Therefore, the selection of studies reported below features some very well-known foundational ones that provide the theoretical framework to approach the selection of relevant features and the analysis of emotions using physiological signals, and some more recent ones, not included in the review paper from Eerola and Vuoskoski, focused on the classification of music-elicited emotions using machine learning algorithms. They have been ordered chronologically to emphasize the methodological progresses and their contribution to the design of the current experimental protocol and the classification methods have been highlighted where appropriate. However, to underline the novelty of this research, only one reviewed study features the use of a wearable EEG headset with 8 dry electrodes and another one artificially simulates the use of wearable device by picking 2 frontal electrodes from a bigger dataset recorded with 32 wet electrodes.

3.1 Classification of music-elicited emotions

L.A. Schmidt and L.J. Trainor [15] were the first investigators that reviewed all the existing regional brain activation/emotion models and tried to systematically verify their validity for the analysis of music-elicited emotions. To do so, they designed an experiment selecting 4 orchestral excerpts that were pre-rated to represent the following classes:

1. Intense-unpleasant emotion: fear
2. Intense-pleasant emotion: joy

3. Calm-pleasant emotion: happy

4. Calm-unpleasant emotion: sad

They hypothesized that:

- "Asymmetric frontal activation reflects emotional valence"
  - Greater relative left frontal EEG activity for joy and happy musical pieces
  - Greater relative right frontal EEG activity for fear and sad musical pieces
- "Regional brain activation reflects emotional intensity"
  - A significant main effect for the intensity of affective musical excerpts on overall frontal EEG activity is characterized by a frontal pattern that would distinguish across valence as predicted by Davidson, Schmidt, and Dawson.
  - Right parietal activity would distinguish the intensity of the affective musical excerpts across valence as predicted by Heller.

Then, they recruited 59 participants (30 females) right-handed undergraduates of psychology between 18 and 34. Their EEG signal was recorded continuously for 60 seconds for each musical excerpt. The data was pre-processed and cleaned from artefacts, then analysed using Discrete Fourier Transform (DFT) with a Hanning window of 1-second width and 50% overlap.

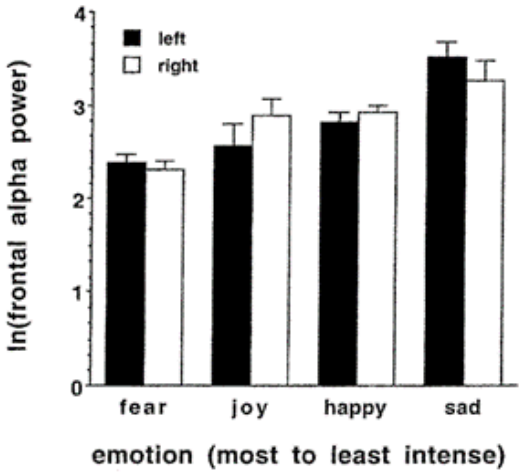

Figure 3.1: Valence by hemisphere interaction showed differences among the four musical excerpts on the left and right frontal EEG alpha power. Taken from [15]

Frequency-band specific power in the alpha band (8-13Hz) was derived from the DFT output over the complete power spectrum. ANOVA analysis on valence by hemisphere interaction revealed a consistent behaviour between the left frontal EEG activity with positive valence songs and between the right frontal EEG activity with negative valence songs (Fig. 3.1). According to their findings, they could confirm that asymmetrical frontal activation indeed distinguished the valence of musical emotion: subjects exhibited greater relative left frontal EEG activity for musical excerpts with positive valence and vice versa on the right side for musical excerpts with negative valence (Fig. 3.2), confirming their first hypothesis. Furthermore, the results showed that musically induced emotions elicit the same frontal brain regions as emotions induced through other means, which validates musical stimuli as good emotional elicitors for agnostic emotion classification. Intensity by hemisphere analysis reported main effects on intensity, but without interaction: subjects showed significantly greater activity in the frontal region as the affective stimuli became more intense.

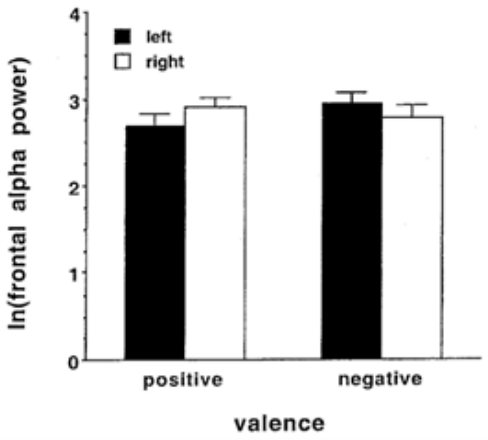

Figure 3.2: Valence by hemisphere interaction.Alpha power is inversely related to activity, positive left is accordingly lower for positive emotions and negative right is lower for negative emotion. Taken from [15]

The frontal EEG activity decreased from the presentation of the fear to the joy to the happy to the sad excerpts and it is consistent with the behavioural rating of intensity. Since they only used auditory stimuli, the lack of parietal differences might be due to the lack of external focus from environmental stimuli.

Valence by intensity analysis showed that musical excerpts with higher intensity and positive valence elicited significantly higher activity compared with the opposite combination, low intensity and negative valence (Fig. 3.3). This study is relevant for the analysis of music-elicited emotions because it provides a validation for the models proposed by Davidson, Fox and Heller that state the approach-withdrawal

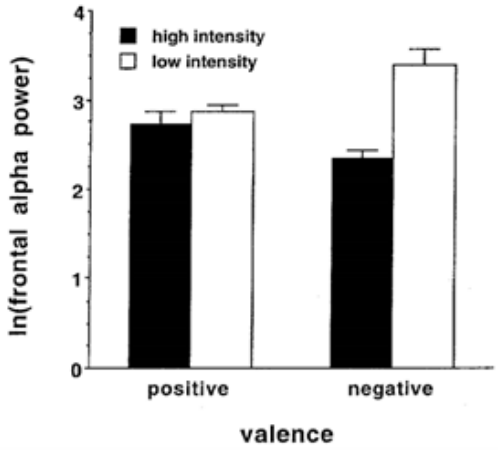

Figure 3.3: Valence by intensity interaction. Greater power (lower activity) in the low-intensity excerpts than in higher intensity excerpts, with a more extreme difference for negative valenced excerpts. Taken from [15]

tendencies in the frontal EEG activity: positive emotions are processed in the left anterior region of the brain, negative emotions are processed in the right anterior region of the brain. Regarding the intensity, or arousal dimension, of the emotions, the results are consistent with the models by Davidson, Dawson and Schmidt that correlate absolute frontal activation with the intensity of the emotional experience. However, in contrast with Heller's model, they did not find relevant differences in the right parietal activity, possibly because of their experimental setup.

In 2009, Lin et al. [23] recorded 26 participants to perform Emotion-Recognition of four emotional states representing the 4 quadrants of the VA space: joy, pleasure, sadness, and anger. They proposed the listening of pre-labelled emotional music and then collected the discrete self-reported labels from their subjects to be used for classification. After removing motion artifacts with visual inspection, they extracted the frequency-band specific power in the Theta, Alpha, Beta and Gamma bands using Short-Time Fourier Transform (STFT) and then derived the power of each EEG component across time over 32-channels. They calculated 12 asymmetry indexes (ASM12) as the difference in power from 12 symmetric electrode pairs for at total 60 features over five EEG components. Support Vector Machine (SVM) was used to classify the data in three different configurations:

- "all-together": multi-class in one step with Crammer's optimization formulation
- "one-against-one": binary classification for K(K-1)/2 classifiers, then the test prediction is decided with max wins strategy.
- "model-based": two-level nested binary classifiers, one for valence and one for

arousal, then the results are aggregated.

The they then proceeded with a subject-dependent strategy for classification. They obtained the best performance with the “one-against-one” scheme, respectively 94.86 % (1.76) accuracy for valence and 94.43% (2.12) for arousal and showed that the binary classification strategy is consistently more reliable than multi-class classification. However, there are no reports about the distribution of the self-reported labels, meaning we do not know if the datasets were balanced or unbalanced between the four emotional states.

In 2010, a remarkable comparison of modern methods can be found in a subsequent study from Y. Lin et al. [24], which developed a systematic framework for optimization of EEG-based emotion recognition. According to the asymmetry and regional activation theories, they extracted a set of spectral features (PSD30) from Delta, Theta, Alpha, Beta and Gamma frequency-band specific power. Subsequently, they derived several asymmetry indexes by power subtraction (DASM12) or division (RASM12) between 12 symmetrical pairs of 24 electrodes placed over the frontal, central and parietal areas of the brain and lastly, they also used the individual spectra of these 24 electrodes (PSD24). They then proceeded with automatic feature selection to improve the accuracy of the classification of four emotional states, namely joy, anger, sadness, and pleasure, testing subject-dependent classification with both SVM and Multi-Layer Perceptron (MLP). After F-score ranking all the features by performance, they identified which features were subject-independent to the whole dataset. Finally, they repeated the experiment lowering the number of electrodes and features, and they compared the performance of both subject-independent and subject-dependent features.

According to the classification results, differential asymmetry features (DASM12) yielded better accuracy than rational asymmetry features (RASM12), Furthermore, DASM12 significantly improved classification performances compared to PSD24, even if they were derived from the same electrodes, suggesting that hemispheric power asymmetry is more discriminating in the measurement of mental states. These differential asymmetries were also subject-independent, meaning that their performance was consistent across subjects. In addition, further experiments on electrodes reduction proved the classification performance to be quite comparable despite the lower number of features, and only dramatically declined when the number of features was reduced below 10 (see Fig. 3.4). This study provides useful insights on the performances of features based on hemispheric asymmetry. However, it is hard to compare with other works or the current study, because of the decision to report classification performances of 4 emotional classes created from the aggregated dimensions in the valence-arousal space.

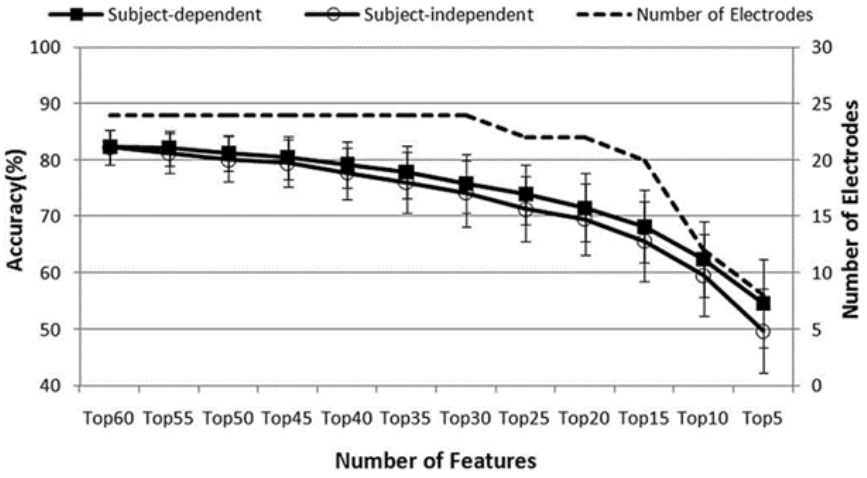

Figure 3.4: Comparison of average accuracy of subject-dependent and independent features and the number of electrodes for subject-independent features. Taken from [24]

In 2013, Koelstra et al. [27] created one of the biggest public available datasets, both in terms of participants and variety of signals collected, to investigate the emotion recognition task using physiological signals. They investigated the role of emotions in communication and realized that most systems fail in interpreting the human emotional vocabulary, they are not able to identify emotional states and use them accordingly. According to the authors, the goal of affective computing is to fill this gap and synthesize emotional responses. Affective characteristics are features that can describe multimedia content and can be associated with implicit emotional tags. These tags could then be used to improve the performance of recommendation and retrieval systems in understanding the user's taste and then recommend content that matches the current emotion. They adopted a three-dimensional model that adds third dimension to Russel's valence-arousal model: dominance. Arousal ranges from inactive to active, valence ranges from unpleasant to pleasant and finally dominance ranges from a "helpless and weak" feeling to an "empowered" feeling. They utilized SAM [13] for the self-reporting task of discrete emotions. Physiological signals seem to carry emotional information; they comprise signals from the central nervous system (CNS) and peripheral nervous system (PNS) as well (see Fig. 3.5). They are available to be used for emotion assessment but were not in the main scope of the experiment. Music videos were used as visual stimuli, 32 participants (16 females) took part in the experiment and their EEG and peripheral physiological signals were recorded while they were watching 40 selected stimuli. They were asked to rate each video in terms of arousal, valence, like/dislike, dominance, and

familiarity and for 22 of them the frontal face video was also recorded. After several steps of semi-automatic selection of 120 stimuli and a manual selection for the rest, 40 final stimuli were selected, considering an equal distribution between the 4 quadrants/classes that can be identified in the valence-arousal space:

- **LALV** : low arousal / low valence
- **LAHV**: low arousal / high valence
- **HALV**: high arousal / low valence
- **HAHV**: high arousal / high valence

For each selected music video, they extracted a 1-minute segment and used an affective highlighting algorithm by Soleymani et al [29]. Between each segment, there were 55 seconds of overlap, content features were extracted and provided the input for the regressors. The emotional highlight score was computed with the following equation:

$$e_i = \sqrt{a_i^2 + v_i^2}$$

Where *a* is the arousal, *v* is the valence and e_i is the i-th segment of emotional highlight. For each video, the segment with the highest score was extracted for the experiment.

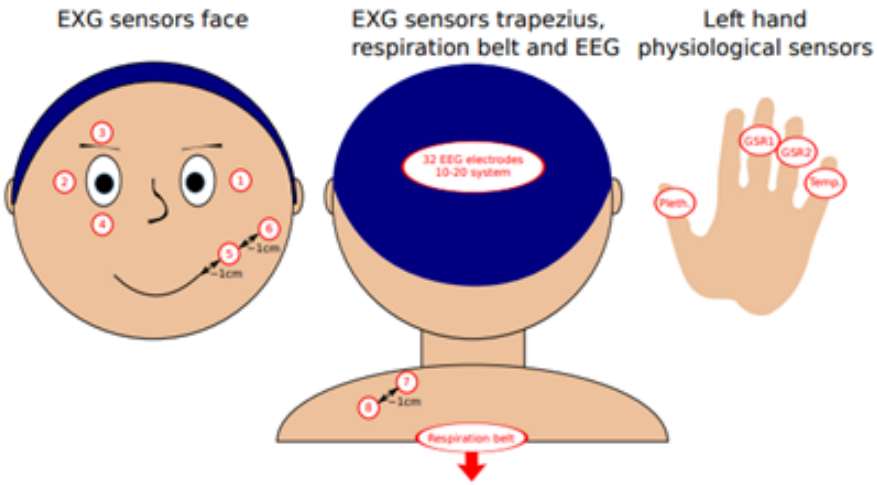

Figure 3.5: Placement of physiological sensors to record EOG, EMG, GSR, BVP, temperature and respiration. Taken from [27].

The experiment included 2 minutes recording of the baseline, then the 40 videos were presented in 40 trials, each consisting of:

- 2-second screen displaying the current trial number

- 5-second baseline recording
- 1-minute display of the music video
- Self-assessment for arousal, valence, liking and dominance

The stimuli from the four conditions, in general, elicited the target emotion (see Fig. 3.6) and high-arousing conditions worked particularly well.

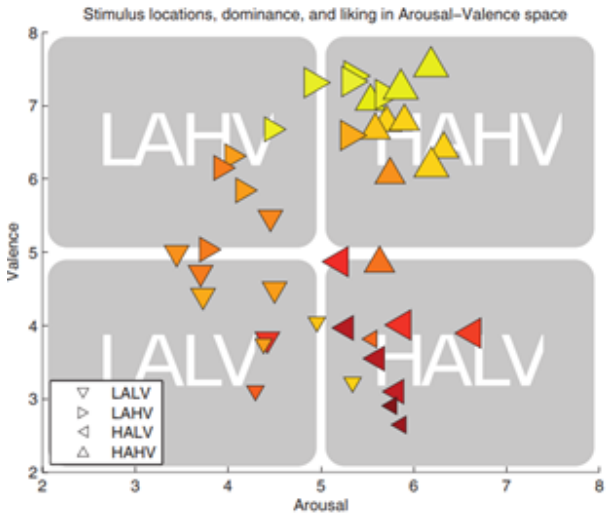

Figure 3.6: Mean location of the stimuli on the Valence-Arousal space for the 4 classes. Liking is colour-coded as red for low and bright yellow for high, while dominance is size-coded with small symbols for low and big symbols for high.Taken from [27].

Emotions with strong valence and low arousal were instead more difficult to elicit. They also observed a high positive correlation between liking and valence, and between dominance and valence, meaning that people liked music that gave them a positive feeling or feeling of empowerment. Medium positive correlations were observed between arousal and dominance and between arousal and liking. Familiarity correlated with liking and valence moderately and positively. They found negative correlations for arousal in the Theta, Alpha and Gamma band, with the central Alpha power decreasing for higher arousal, that matched the findings of their previous study. There is also an inverse relationship between Alpha power and the general level of arousal. Valence instead showed the strongest correlation with EEG signals in all the frequency bands. In Theta and Alpha frequency bands, an increase in valence led to an increase in power. For the Beta frequency, there are a central decrease and an occipital and right temporal increase of power, associated with positive emotional self-induction and external stimulation. Liking correlates could be found in all frequency bands, for Theta and Alpha power they showed an increase

over the left frontocentral cortices (see Fig. 3.7). In summary, their observed correlations partially concur with their previous study and other studies that explore the neurophysiological correlates of affective states.

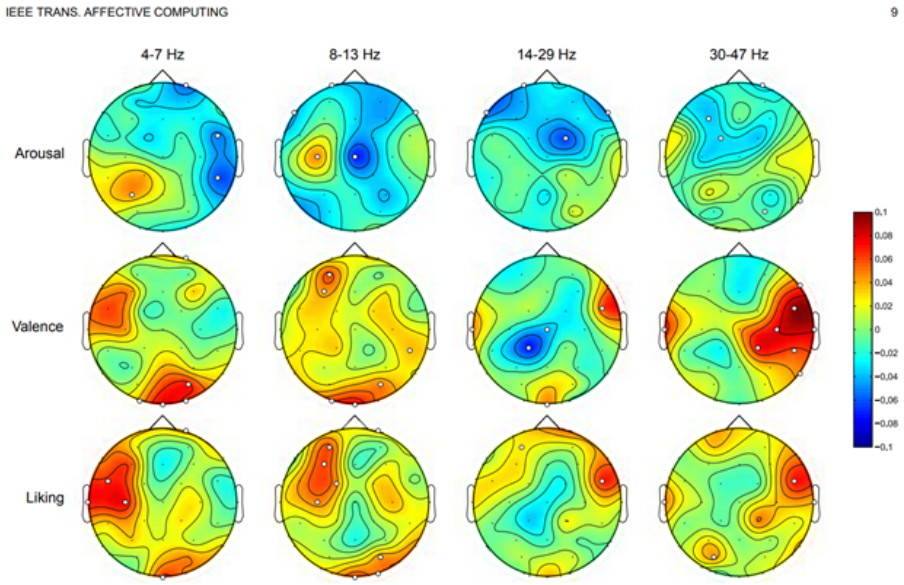

Figure 3.7: Mean correlations overall participants of valence, arousal, and rating with power in Theta, Alpha, Beta and Gamma bands. Highlighted sensors are significantly correlated with the ratings.Taken from [27].

To preprocess the EEG data, the signal was down sampled, high pass filtered using EEGLab and then eye artifacts were removed with blind source separation technique using the recorded Electrooculography (EOG). Classification was experimented in three modalities: EEG signals, peripheral physiological signals, and Multimedia Content Analysis (MCA), but only the first one is relevant for the current study. Power spectral features were extracted from the EEG signal in the Theta, Alpha, Beta, and Gamma bands, then the asymmetry was measured as difference in spectral power from symmetrical pairs of electrodes. In total they used 216 EEG features for single trial subject-dependent classification with leave-one-out cross validation scheme, where one stimulus was used as test set at each step of the cross validation. Many datasets where unbalanced in the distribution of valence and arousal labels, thus they used F1-score to assess the reliability of the accuracy scores. Using a gaussian naïve Bayesian classifier, they reported an average accuracy of 62% for arousal classification with F1-score of 0.583, and an average accuracy of 57.6% for valence classification with F1-score of 0.563. The results were compared against random guessing, class ration and default majority class guessing. The F1-score distribution was significantly higher than 0.5, indicating the models' capability to learn from EEG features despite unbalanced datasets and av-

erage accuracy lower than majority class guessing.

Reuderik, Mühl and Poel [21] investigated correlations between emotional valence, arousal, and dominance during game play to study affective correlates in a realistic and uncontrolled environment. To induce frustration, they utilized a game designed to ignore 15% of keyboard input for short periods, with the screen lagging to simulate an under-powered computer. They recruited 12 healthy users and asked them to play the game through a sequence of random permutations of two normal games and one frustrating game of 2 minutes each. After each game, the subjects were asked to self-report their current mental state using Self-Assessment Manikins. During the experimental session, they were recorded with a BioSemi ActiveTwo EEG system with 32 active electrodes placed at the extended locations of the 10-20 system. In addition, EOG were recorded to measure the influence of ocular artifacts and Electromyography (EMG) to record the finger movement used to control the game. The EEG data were high pass filtered to remove frequencies below 0.2Hz and notch filtered to remove power line noise, then the signal was corrected for eye movements using linear regression analysis subtraction. Finally, the data was re-referenced to the common average reference (CAR). For the feature extraction, they estimated the power in different frequencies using Welch's method for each experimental game session, then within each session they summed the log-power in the Alpha band for each electrode and subtracted the band power of electrodes on the left hemisphere from the corresponding electrodes on the right hemisphere. The obtained Alpha-asymmetry indexes for each sensor pair were used to find a correlate with valence. The results of the statistical correlations of self-reported valence, arousal and dominance confirmed correlates for both valence and arousal in the Theta, Delta, and Alpha frequency-band specific powers during the activity of gaming. The different affective dimension did not seem to be orthogonal, valence and dominance ratings were highly correlated, thus effects found in the EEG related to one affective dimension can be attributed to the other. Their conclusions on frequency-specific and localized emotional interpretation are valuable for the current study, despite the use of a video-game as elicitor. The asymmetry of Alpha power and fronto-central Theta power were validated for the measurement of emotional valence. Right frontal Alpha power and the absence of right parietal Delta power instead were instead indicative of the arousal dimension. They concluded stating that these effects and the stronger narrow-band effects could be used for automatic recognition of affect. The importance of this study lies in the validation of the asymmetry and regional absolute activation theories in a realistic scenario, which is of fundamental importance in the pursuit of real-time ER.

The data-oriented approach in the study from Thammasan et al. [26] in 2016 focused on considering emotional oscillations within a single music piece during EEG-based emotion recognition. They proposed a continuous emotion-recognition approach, including self-reporting and continuous emotion annotation using the VA model. After adopting two different approaches for information extraction, Fractal Dimension (FD) and Power Spectra Density (PSD), they discovered that FD slightly outperforms PSD in both arousal and valence subject-dependent classification, while having a higher correlation between the classification and self-reported emotions. FD is an alternative approach to the analysis of irregular time series, proposed by T. Higuchi [30], and it is now getting more popularity in affective computing research, because of a relative simplicity compared to PSD. Higher values of FD reflect the higher activity of the brain; PSD instead indicates the signal power in specific frequency ranges, and it's based on Fast Fourier Transform (FFT), used to decompose the EEG signal into the previously explained frequency ranges: Delta, Theta, Alpha, Beta, and Gamma. EEG-based emotion recognition is promising for potential applications like music therapy, multimedia tagging, and multimedia retrieval. However, previous studies did not consider emotion variation and usually adopted a single emotion annotation approach because the experiments were run with music excerpts shorter than a minute. Since emotions are subjective and the same piece of music can induce different emotions in listeners, they also gathered self-annotated emotion labels from them. This can overcome the problem that many studies have, i.e. the use of emotionally pre-labelled music pieces from standard libraries, where emotions are labelled by an expert or by other users. For the experimental session, they recorded 15 male participants between 22 and 30, all healthy students from Osaka University. The music collection was composed of 40 pieces in MIDI format, so that additional emotions contributions by lyrics could be eliminated. The 12 electrodes used were chosen for their location close to the frontal lobe, the part of the brain that is crucial for emotion regulation: Fp1, Fp2, F3, F4, F7, F8, Fz, C3, C4, T3, T4, and Pz. Each song length was on average two minutes, then followed by 16 seconds of silent rests to allow the participant to mitigate the effects from the previous song when starting the next one. After the listening session, the participants listened to the same songs and annotated their perceived emotions by clicking on the corresponding points in the valence arousal emotion space. For feature extraction, they applied a sliding window segmentation that could analyse temporal data to track emotional fluctuation. A window of 1000 samples was used, equivalent to 4 seconds. To perform emotion classification, emotional tagging in one window was set to a high or positive value if the number of positive instances was greater than the number of negative instances. Using feature extraction with the two approaches, they obtained 12 features from FD value calculation and 60 features with PSD. Since

traditional methodologies neglected emotional changes over time, they decided to compare the continuous recognition with the traditional method by simulating it with a sliding window size expanded to the full length of the song. The “chance level” was introduced for annotated emotion to provide a benchmark to evaluate models since annotated data could be unevenly distributed in terms of positive/negative perception. In this way, the chance level is defined by the majority class of the training data (with 60% positive samples, the chance level would be 60%).

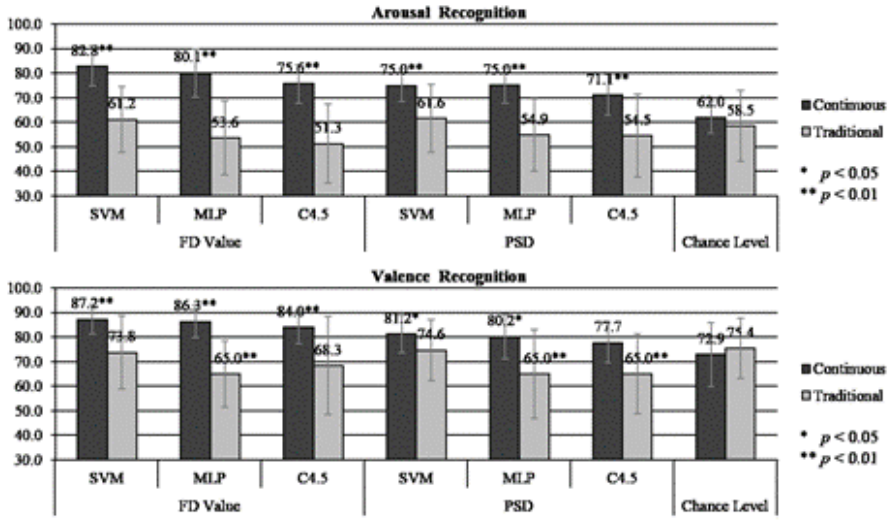

Figure 3.8: Average classification accuracy and standard deviation for valence and arousal across all subjects .Taken from [26].

The results (see Fig. 3.8) show that all the continuous approaches, regardless of the algorithm used, outperform the traditional method. Classification with FD features using SVM achieved the best relative result of 82.8% for arousal recognition with a chance level of 62%. Also in valence recognition, FD features proved to perform better with SVM and the highest accuracy of 87.2%. The general higher arousal correlation of FD value features might be the reason it performed better than PSD for arousal recognition, and similarly, they could achieve better results for valence recognition because of their slightly higher absolute correlations. The results obtained with PSD features are later compared with the results obtained in the current study, that also used features extracted from PSD (see Chapter 5.2).

Wu et al. [25] experimented valence classification on the DEAP dataset [27] in 2017 selecting a small subset of channels located in the frontal area according to the asymmetry theory, to simulate the use of a wearable device. They extracted spectral properties using FFT and calculated the following features:

- Entropy to measure the randomness of a signal in the Delta and Gamma fre-

quency bands.

- SASI to detect emotions based on a balance between power in Theta and Beta frequency bands as defined in Chapter 2.7.

- Emotion Valence Index (EVI) that reflects the hemisphere asymmetry of frontal Theta power similarly to AWI and calculated as:

$$EVI = \frac{10(\log_{10}(TF\theta PL) - \log_{10}(TF\theta PR)}{(\log_{10}(TF\theta PL) + \log_{10}(TF\theta PR)}$$

They also calculated differential asymmetry indexes based on the power in Beta (BASI), Delta (DASI) and Gamma (GASI) frequency bands, for a total of 68 features for classification. After building a multi-classifier system for subject-dependent classification, they obtained the best performance with a Gradient-Boosting Decision Tree (GBDT) classifier that scored a maximum valence classification accuracy of 76.34% and a mean accuracy of 75.15% using only two frontal electrodes, Fp1 and Fp2. They repeated the experiment with 4 and 6 frontal electrodes as shown in Figure 15 only slightly improving the performance. The subject-independent experiment with leave-one-subject-out scored sensibly worse, with an average accuracy of 61.82%, having highest accuracy of 91.67% and lowest accuracy of 21.43%. Furthermore, to consider the unbalance towards the positive valence class, the authors labelled each trial as positive if the valence score was higher than 7 and negative if the valence score was lower than 3, but it is not clear how the trial is labelled for a value in between nor what is the actual distribution of labels for each subject.

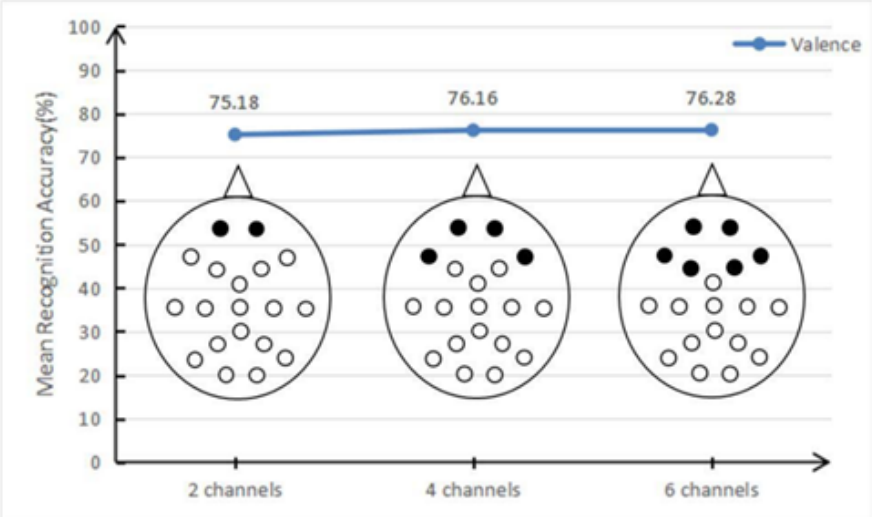

Figure 3.9: Mean classification accuracy with 3 different subsets of electrodes from DEAP. Taken from [25].

These results proves that classification is possible with a meagre amount of electrodes, but also that strategically selected frontal electrodes contain enough affective information to eventually perform the classification task using a wearable device and with a lower computational cost, thus more suitable for real-time classification. However, subject-independent classification seems to be highly affected by subject-dependent variations, thus less promising. The authors also did not mention their preprocessing strategy, and this fact is only explainable if they used the already preprocessed version of the DEAP dataset, made available by the authors in the same package with the raw data. This study is valuable for the investigation of Emotion-Recognition with a simulated wearable device but does not give further insights for a realistic real-time application.

Always in 2017, Thammasan et al. [31] presented a framework for adaptive multimodal recognition using a dataset collected recording the EEG, Electrocardiography (ECG) and Galvanic Skin Response (GSR) signals of 9 healthy subjects listening to music. This study is the only one based on a wearable EEG headset with 8 soft dry electrodes developed by IMEC. All the musical stimuli were selected based on previous studies and with statistically verified emotional ratings, simplified to 4 classes corresponding to the quadrants of the VA space. They also utilized a stratified music-selection approach, with a selection of 16 songs from the researchers themselves, and 8 songs subject-selected. They then proceeded with the experiment, collecting the VA labels with continuous annotation using the mouse and SAM [13] in a scale from 1 to 9 after each trial, followed by a rating of familiarity in a scale from 1 to 5 to verify the overall familiarity with the music. They also collected whether the subjects liked or not the song and their confidence in the self-rated annotations on a scale from 1 to 3. The PREP [32] preprocessing pipeline was run in EEGLab with standard filtering and Independent Component Analysis (ICA) decomposition to remove eye movements activity and muscular artifacts. The EEG features were extracted using multi-taper PSD to minimize the bias and make them more robust under stochasticity. PSD features were extracted in the Theta, Alpha, Beta and Gamma frequency bands, but no further differential or rational computation was applied. Classification was performed with SVM based on RBF kernel with subject-dependent strategy. For each subject, emotional valence and arousal were classified with leave-one-block-out cross-validation, then Matthews Correlation Coefficient (MCC) scores were calculated to give a more accurate representation of the classification performances with unbalanced classes. This is one of the first reviewed studies to clearly consider the unbalance of classes that is typical of emotion related studies involving music, and the accuracy performances are compare against the “chance level” of each subject, which is defined as majority-voting accuracy. The performances of

the classification are evaluated as multi-modal contribution of several physiological sources, but the maximum accuracy obtained by EEG as uni-modal source is 80% for valence and 72% for arousal. Unfortunately, no detailed accuracy scores are provided, but instead the MCC scores are presented for each modality. The average MCC score for valence is 0.247 (0.17) with the highest score being 0.596 (0.3), while the average MCC score for arousal is 0.177 (0.04) and the highest score being 0.23 (0.22). Given the impact of personal perception on the distribution of emotional labels, MCC is an important evaluation metric as later explained, when discussing the optimization process of the current research (see Chapter 4.2.7).

Recently, Keelawal et al. [33] following up their previous studies [26], [34] compared the use of the deep learning algorithm Convolutional Neural Networks (CNN), with other methods for emotion recognition on EEG signals. CNN has shown very good results and potential in the generalization of unseen subjects, therefore they aimed to study how to tune the hyper-parameters to obtain beneficial optimizations. Their results show that the temporal information in distinct window sizes significantly affects the recognition performance, and CNN was more responsive to window changes than SVM. Subject-independent classification with (leave-one-subject-out (LOSO) strategy and 10-fold cross-validation of arousal achieved highest accuracy of 56.85% and MCC of 0.1369, window size of 10 seconds, while valence recognition performed a highest accuracy of 73.34% and MCC of 0.4669 with an 8 second window size. CNN has been recently applied to EEG-based emotion recognition with the advantage of circumventing feature engineering and improving classification accuracy thanks to its advantages at capturing adjacent spatial information. The fact that emotional responses can evolve creates the necessity of continuous annotation of emotions to allow capturing the temporal dynamic of emotion. Spatial information is also important using CNN: the placement of adjacent electrodes in the input matrix can be impactful, meaning the accuracy can be improved with an optimal arrangement of the order of EEG electrodes over the most contributing regions of the brain in emotional processing. The experiment was conducted with 12 male students from Osaka University, using a collection of 40 MIDI files, which were equally distributed over the four quadrants of the arousal-valence space. Subjects were instructed to select 8 familiar songs and 8 unfamiliar songs. Each song was played for 2 minutes with a 16 second silence interval in between. The electrodes were chosen near the frontal lobe. After listening to the songs, the subjects were detached from the EEG equipment and were asked to listen again in the same order and annotate the emotions by clicking on the arousal-valence space at the corresponding position (every 2-3 seconds). During the EEG pre-processing a band-pass filter was applied (0.5-60Hz) and ICA was computed using the Info-Max algorithm

and then the components were evaluated based on the power spectral density, scalp topography, and location of the equivalent current dipole. Four different architectures were tested with respectively 3, 4, 5 and 6 convolutional layers and the same model as represented in Figure 3.10.

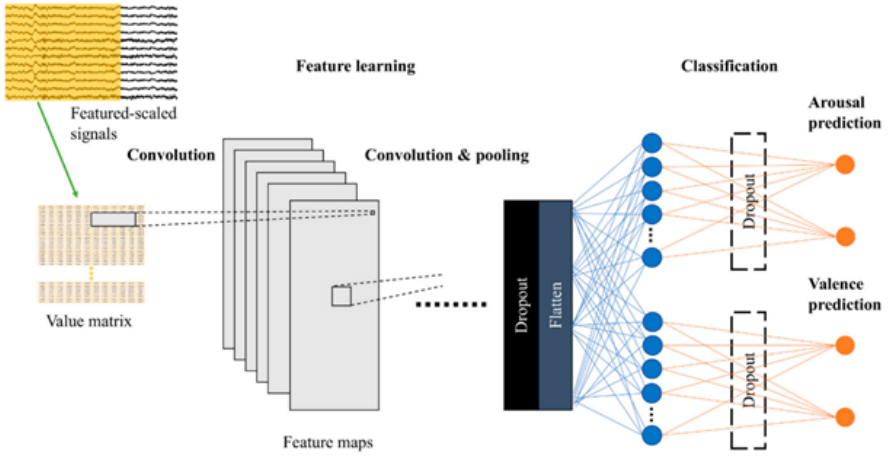

Figure 3.10: Optimized architecture of the model in [33].

Increasing the window frame led to higher performance in all CNN architectures, for both arousal and valence, with a higher improvement of the valence condition considering the MCC ranges of both conditions (0.1302 for valence and 0.0951 for arousal). Arousal classification scored the best results with window size of 10 seconds, obtaining 56.85% accuracy and MCC value of 0.1369, and scored the lowest with a 1-second window size obtaining 52.09% accuracy and MCC of 0.0418. On the other hand, valence classification obtained the best accuracy at 73.34% with a window size of 8 seconds and MCC value of 0.4669, while the lowest accuracy was 66.83% and MCC of 0.3367 with 1-second window size. According to these results, there could be enormous variations between the signal of each subject and expanding the window size could reduce fluctuations among distinct subjects. Compared to their previous work using SVM (linear, polynomial and RBF kernels) on the same EEG dataset, the window size was way more influential for CNN. Electrodes sorting instead had less marked effects on the classification in comparison with a random order, with 3D Physical order for valence classification obtaining significant differences with 72.94% accuracy and MCC of 0.4588. MinCBO also had some significant improvement overall. The results of subject-independent classification in this experiment seems to confirm that strategy is achievable with computationally demanding offline pre-processing pipelines and a fine-tuning process of the deep learning models. These considerations, together with the use of a standard EEG headset, suggest that subject-independent classification might be still too challenging for real-time Emotion-Recognition using a wearable device.

The general trend of investigating Emotion-Recognition using EEG is to support the task with discrete or continuous emotion annotations to reduce the biasing effects of pre-labelled emotional labels, as well as selecting the recording channels that are positionally more relevant in the analysis of emotional processing in the brain. Fractal Dimension algorithms and Power Spectral Density calculation seem to be the most promising pre-processing techniques for features extraction. The reviewed studies focus on the frontal asymmetrical hemispheric tendencies in the Alpha and Theta powers, which are usually computed as differential or rational indexes between symmetrical pair of electrodes. The absolute activation in frequency-band specific Theta, Alpha, Beta and Gamma powers in the frontal, central and parietal area are all relevant in the study of emotions, but there was not a prominent method for features computation. The limitation of the EEG equipment used for the current study, as well as the light preprocessing approach, constrained the analysis so that only frontal Theta, Alpha and Beta frequency-band specific powers were considered. Supervised learning algorithms, like Linear Regression, GBDT and SVM, but also deep learning algorithms like MLP and CNN, were possible valid choices for the task, depending on the amount of available data and the classification strategy. The most popular strategy among the reviewed studies is subject-dependent analysis, that generally yields better accuracy, compared to subject-independent strategy that requires fine tuning of features and models. Not all the reviewed studies reported if the distribution of self-reported emotional labels was balanced or unbalanced, which may cause major problems in the classification and diminishes the reliability of reported accuracy scores in absence of more informative coefficients like F1-score and MCC score. In the next chapters, the methods used in the classification task and the problems encountered further elaborate on these issues.

Chapter 4

Methods

This chapter describes the methods used to carry on the research, from the data collection to the preprocessing and the results of the analysis. It includes a brief description of the population that took part into the experiment, the experimental task and conditions, the procedure adopted, and the equipment used. Then it delves into the analysis process starting with the preparation and preprocessing of the data and concluding with the intermediate outcomes of the classification. As previouslyu mentioned, the experiment is the result of a collaboration between the University of Twente, that provided the space, the recruitment system and part of the recording equipment, and myBrainTechnologies that provided their EEG-capable Melomind headsets and the rest of the equipment. The experimental design was reviewed and adjusted for ethical approval by the Ethics Committee of the university with reference number RP 2021-43 and compliant with the safety measures to prevent the spread of the Covid-19 virus.

4.1 Experiment

The experimental phase of the research was crucial for the proper evaluation of Melomind. Previous studies conducted internally at myBrainTechnologies provided the theoretical backbone to work with neuromarkers and music, however the datasets of the employees were collected with a different experimental setup that did not account for the continuous measurement of emotional valence and arousal, and in any case there was always a chance data could be biased by the previous knowledge these ”expert” subjects had on the topic and the technology. A feasible alternative could have been the simulation of a wearable device, similarly to the approach taken by Wu et al. [25], but that was kept as emergency solution in case it was not possible to run an experiment under the safety restrictions enforced by the Covid-19 pandemic. However, the open datasets available online for emotion analysis

[27], [35], [36] are recorded with a very different equipment than the Melomind, that would have not properly reflected the challenges of using a wearable neural interface. Finally, the opportunity to conduct the study with the students of the University of Twente and the availability of myBrainTechnologies to ship all the technical and hygienic materials to the Netherlands made it possible to proceed with the experimental protocol in the desired conditions.

4.1.1 Experimental Annotation app for data collection

An app was developed to collect continuous annotations of perceived emotion, inspired by the design of the FEELTRACE tool [37] and the app developed by Thammasan et al. [26]. The Experimental Annotation (EA) app was developed in Python using the Psychopy[1] engine for experimental behavioral sciences. The app is a collection of timed routines that alternate guided instructions, annotation tasks on a simplified GUI representing the valence-arousal space and Likert scales to report familiarity/liking scores in the range [1-5]. Three training sessions have been included:

- T1: the participant is presented with some background information about the VA model and how to use the annotation tool.
- T2: the participant is asked to annotate on the VA space the perceived emotion while listening to 2 minutes of mixed music genres.
- T3: the participant is presented with the simulation of a trial of the experiment, including reporting of familiarity/liking and the two listening conditions (see Chapter 4.1.5).

The EA app was designed and developed at myBrainTechnologies and then tested with the other employees during a short pilot period to adjust the instructions, the clarity of the GUI and the input method. Two input methods were evaluated with A/B testing methodology, using mouse and joystick respectively. The results of the test (see Appendix A.1) confirmed that using mouse as input source required less training and effort, thus softening the cognitive load of the annotation task while music listening. Using the joystick would have enabled collecting annotation even in an eyes-closed listening condition thanks to the tactile feedback, but at the cost of requiring more training and concentration. To record experimental timed events, the EA app was connected to the Melomind through a TriggerBox with an USB cable, a customized Arduino Nano board that can send binary-encoded labels using the serial port.

[1] https://www.psychopy.org/

4.1.2 Participants

In respect with the Covid-19 safety measures enforced by University of Twente, 45 healthy participants (28 females) participated in the experiment, all students, or ex-students of the university. The mean age of the population is 23.8 ± 3.1, with the oldest student being 31 years old and the youngest student being 18 years old at the time of the experiment (see Appendix A.2.1). The lowest educational level was the enrollment as bachelor student and the highest educational level was having completed a master's degree. Almost half of the participants (20) were Dutch, while all the rest came from different countries, but all of them had at least a C1 or equivalent English proficiency as requirement to enroll in University of Twente. Prior to be confirmed as participants, they were invited through an invitation form informing them on the nature of the experiment and collecting personal information such as demographic information, health conditions, drugs consumption, musical literacy, and some behavioral information on their habits in listening and searching for music to later support the design of a prototype. Almost 5% of the participants reported to be left-handed, but none asked for an inverted setup of the equipment after it was offered to them (see Appendix A.2.1). The only strict criterion to participate in the experiment was the capability to hear music, eventually through a hearing support system. None of the applicants was discarded nor required additional support for their health conditions. Prior to their experimental session, they were asked to refrain from consuming recreational drugs and alcohol in the 12 hours before the experiment and caffeine and tobacco in the hour before the experiment to prevent unexpected biases in the cerebral activity.

4.1.3 Stimuli selection

The stimuli were selected to represent an even as possible distribution of emotions according to the 4 classes identifiable by the quadrants of the Valence-Arousal space. These 4 classes are the possible combinations of positive and negative valence with high and low arousal. To keep consistency with related studies [27], the classes have been named as follows going clockwise from the top-right quadrant:

- HAHV: High-Arousal and High-Valence.
- LAHV: Low-Arousal and High-Valence.
- LALV: Low-Arousal and Low-Valence.
- HALV: High-Arousal and Low-Valence.

Selecting the right stimuli is a non-trivial task especially in the case of music since many factors can bias the personal perception. For example, familiarity with a certain song might elicit stronger emotions or create an effect of anticipation [38], [39], [40], while cultural biases or genre preference might completely change how a song is perceived [5], [41]. Choosing to include lyrics or no lyrics may shift the attention of the listener from the meaning to the melody and vice versa. It is clearly hard to address all the possible issues but considering the scope of this study and the research on a realistic use-case scenario, most of these factors were either mitigated or considered with the due precautions.

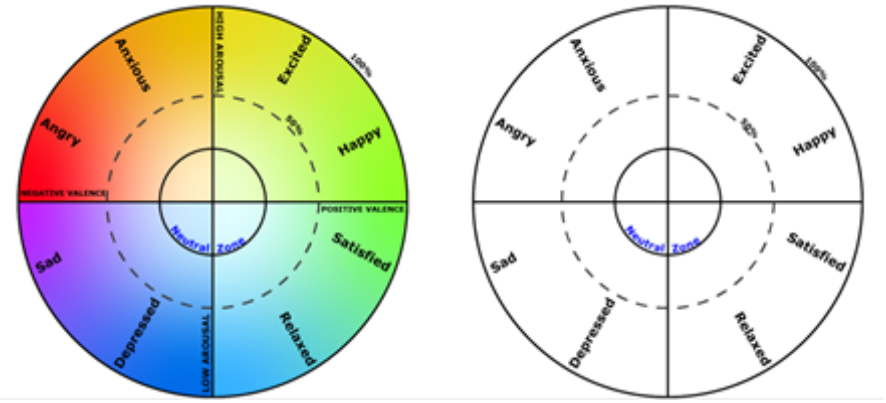

Figure 4.1: The Valence-Arousal space GUI used for the training with color cues on the left, and the uncolored Valence-Arousal space GUI used for the experiment session on the right.

The stimuli were finally selected as a subset of 8 songs (see Appendix A.2.2) from the music database created by Koelstra et al. [27], according to their emotional tagging. They used the popular online music database last.fm[2] to retrieve 120 songs with associated music videos through their APIs, emotionally labelled by thousands of users. They then screened them down 40 stimuli during a web assessment session with at least 14 volunteers for each stimulus. The 8 songs selected for this study are a randomly picked subset of those 40 stimuli whose emotional web assessment belonged to the same Valence-Arousal quadrant as the last.fm tagging. For each quadrant there are exactly 2 songs and in total 8 emotions are supposedly portrayed: excitement, happiness, satisfaction, relaxation, depression, sadness, anger, and anxiety. It is important to point out that the web assessment conducted by Koelstra et al. was done using the music videos of these songs, and that the placement of the emotions in the VA space used in this experiment (see Figure 4.1) is a functional simplification of Russel's model [12].

[2]https://www.last.fm/

4.1.4 Conditions

There is no common agreement in the academic world on which should be the best recording condition for an EEG experiment about emotions, but most researchers agree on the minimization of external stimuli. Only few studies tried to assess the impact of recording in eyes-open (EO) condition and eyes-closed (EC) condition on emotion analysis. Barry et al. reported [42] differences in topography and power levels, due to the processing of visual input, and recommend considering them when choosing baseline conditions. Chang et al. [43] analyzed recording conditions in relation to music listening and reported that frontal Theta power significantly increased in the EC condition, while asymmetries indices in the Alpha power on parietal and temporal sites reflected emotional valence for EC and EO states respectively. In addition, participants rated music as more pleasant and more positive while listening with their eyes closed. These differences in the listening conditions did not seem to significantly impact on the current study that only used frontal electrodes but were considered in the design of the experimental task and in the choice of the resting state baseline. Another problem is caused by ocular movements and blinks that generates large artifacts in the EEG signal [44]. As a consequence, the data collected is of lower quality and requires more computationally expensive preprocessing. In the worst cases, some data must be pruned or reconstructed, varying from a few channels to the entire dataset of a participant. Eye artifacts are typically found in the data recorded by the electrodes placed on the frontal area of the scalp and are usually filtered away by subtracting EOG, if recorded, from the EEG signal. However it is not the case of this study that could not take advantage of extra sensors to record EOG . In general, we can assume that an EC condition yields better quality data than an EO condition because the quantity of ocular artifacts will be reduced to the minimum and there is no underlying visual stimuli processing. The downsides of experimenting in the EC condition are the obvious limitations on the task that could be presented to the participants and a possible increase of power in the Alpha band of the spectrum, that is usually amplified during resting and focused states. The main advantage of the EO condition is the possibility to ask the participants for more complex tasks, at the cost of generating more ocular and muscular artifacts and eventually introducing multiple cognitive tasks at once, that can affect the analysis. Prior to the experiment, we run an internal pilot test at myBrainTechnologies to explore the best compromise options between having good quality data, the maximum amount of data and collecting the behavioral data we needed. Thammasan et al. [26] opted for a double listening protocol, in EC condition to record EEG and in EO condition to collect affective annotations. This translates into listening twice to the same song but recording only in the EC closed condition and then overlapping the annotations taken during the EO conditions. Our limitation of 2 frontal electrodes

already constrained the collectable amount of data, so we decided to extend this protocol with a double listening and recording approach, in both conditions. During the pilot we explored the feasibility of collecting annotations in both conditions using a joystick, but then opted for collecting annotations only during EO condition with a mouse and then reuse the same annotations on the EEG data collected in EC condition. As final consideration, the two conditions can be both present in realistic scenarios, with EO being the most common listening condition, for example in an office or free-time scenario in which a user listen to music while performing other cognitive tasks (work, homework, gaming...). Listening in EC condition resembles a more relaxed scenario, for example when listening to music at the opera or on a comfortable couch in the evening.

4.1.5 Task

During the experimental session, participants were presented a main task during which their physiological signal were recorded. The task is divided into 3 sub-tasks for both conditions during each trial, with a total of 8 trials. The average length of the recording was approximately 35 minutes of recording. The sub-tasks in each trial, repeated for both conditions, were the following:

- Listening to white noise: before presenting the stimulus, participants listened to 15 seconds of white noise to "reset" their emotional state.
- Listening to the stimulus in EO/EC conditions: participants listened to 60 seconds excerpts of each song. During the EO condition participants were requested to continuously annotate their emotions on the VA space, during the EC condition they focused solely on the music. The order of presentation of the conditions depended on the assigned group.
- Rating the stimulus: after listening to the song, participants were requested to give a rating in terms of familiarity and liking of the excerpts, using Likert scales ranging from 1 to 5. If for any reason they failed to give a score before the 20 seconds timer expired, the score would be automatically set to 3, that represented a neutral answer.

So, during each trial, the participants would listen to two different songs, rate them and only annotate during the EO condition (see Fig. 4.2).

It is important to underline that the order of the conditions might induce some bias in the annotation task and the rating of each stimulus. To minimize the statistical effects, participants were randomly assigned to two groups in equal distribution, ECEO and EOEC, according to an inversion criterion that would determine the order

Figure 4.2: Diagram of an experimental trial starting with EO listening condition.

in which the conditions were presented during the entire session. This solution seemed more elegant and less confusing than fully randomizing the order of the conditions for each trial, possibly creating confusion in the instructions. In addition, instead to presenting the same song consecutively in both conditions, we decided to split the session into two parts in which the stimuli are presented in pseudo-random class order in the first part, and then inverted in the second part as shown in Figure 4.3.

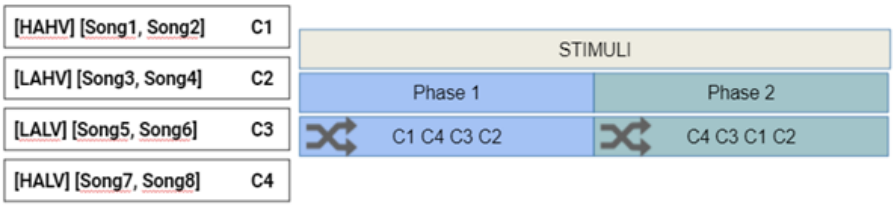

Figure 4.3: Pseudo randomization scheme. Participants listen to all songs twice, in both conditions, during the two parts of the experiment.

This approach also reduces the familiarity effect caused by listening twice to the same song in a short span of time and mitigates cases of extreme emotional fluctuation within each trial, for example if a very sad song would be followed by a very happy song and then again another sad song. This emotional fluctuation phenomenon cannot be fully eliminated, but it is statistically balanced by the pseudo-randomization of the order in which the classes are presented.

4.1.6 Equipment

To record EEG signals from subjects the standard Melomind (Fig. 2.2) was switched in frontal setup using electrodes placed over [AF3 AF4] positions of the 10-20 sys-

tem. The standard Melomind has two more textile electrodes placed behind the ear lobes to be used as ground and reference and can record signals with a sampling rate of 250Hz. The Melomind was connected via USB cable to a laptop through a TriggerBox and via Bluetooth to a smartphone to remotely control the start and end of the acquisition with the proprietary Acquisier app developed by myBrainTechnologies. The TriggerBox is a micro-controller that can be used by third apps to send binary strings to the Melomind to mark an event; during analysis these strings can be decoded using a dictionary to flag a specific segment of EEG, for example a trial of a specific VA class. In this way splitting the trials and group them by class is simple, even if they were presented in random order during the experimental session. An Empatica E4 wristband was used to collect bio signals from the non-dominant wrist of the participant, namely: blood-volume pressure (BVP), body temperature, heart rate (HR) and electro-dermal activity (EDA) (Fig. 4.4).

Figure 4.4: EmpaticaE4, a wearable device that can record physiological data in real-time.

The Empatica E4 was connected to an Android tablet running the Empatica application, so the researcher could monitor in real-time the data collection. The EA app was entirely developed as a set of automated routines in Psychopy, including trainings for the task, synchronization of the triggers with each experimental event and instructions for the users. The timers of each routine were calibrated during the pilot to allow even slower readers to follow up.

The participants could interact with the experimental application through an external monitor connected to the researcher's laptop and an agnostic mouse, although all participants decided to use the right hand. Finally, all the sessions were recorded with a GoPro Hero 7 to monitor accidental events and eventually support the emotion recognition task through facial expressions in a later study. An example of setup for the experiment can be see in Fig. 4.5.

Figure 4.5: Experimental setup with Melomind, Empatica E4 and GoProH7. The EA app is running on the monitor, while the participant is annotating emotions using the mouse (on the left) and then keeping his eyes closed as instructed in the following task (on the right).

4.1.7 Procedure

The experiment was conducted in a controlled environment made available by BMS lab at the University of Twente, with a strict protocol for sanitizing the equipment between each session, no direct skin-contact with the researcher during the setup, opening of the air flows every 10-15 minutes and at least 1.5 meters of distance with the researcher during the experimental task.

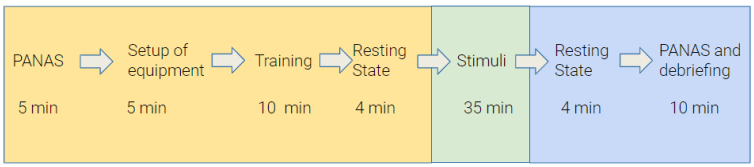

Figure 4.6: Scheme of the experimental procedure, estimated to last 75 minutes.

Upon their arrival, participants were invited to sanitize their hands, to read and sign the informed consent form and then to fill a PANAS questionnaire [14]. The PANAS is used to measure the change in positive and negative feeling and emotions in a specific span of time, from a few minutes up to a few weeks. In this case it was used to measure these changes when conducting the research task and analyse the impact of the protocol on the population that participated in the experiment (see Chapter 6.3. After the questionnaire, the participants could start the training session divided into three parts for a total of 10 minutes, without recording any EEG. The first part gave some introduction and background about the Valence-Arousal model and allowed them to get some confidence with the annotation GUI. The second part proposed a mix of 4 music excerpts, one for each Valence-Arousal class, and asked

them to annotate in real time with the emotions they were feeling. Finally, the third part was a complete simulation of an experimental trial, including instructions, white noises, both listening conditions and ratings. After the training, participants were asked to fit again the device on their head, then the researcher re-positioned the reference and recording electrodes to obtain the best possible quality signal using the Quality Checker (QC) tool of the Acquisier app. The Empatica E4 was then fit on their wrist to allow a precise calculation of heart rate, and finally the Melomind was connected to the laptop through the TriggerBox. Participants were also advised to avoid sudden head movements. Before starting the session, their resting state baseline was recorded, 2 minutes in EO condition and 2 minutes in EC condition. When they were ready, they could start the session and follow the automated instructions, with the order of conditions determined by their assigned group. Halfway through the session they could take a 5-minute break, look away from the screen and drink some water, but they were not allowed to remove the equipment. After completing the second part of the session, another resting state was recorded with the same settings of the previous one, and then they filled the second PANAS questionnaire. The resting state recordings are also part of the standard myBrainTechnologies protocol to compare the mental changes in the resting states after an experiment and to be used as baseline during EEG analysis. For this study only the resting state in the EC condition prior to the experimental task was eventually used as recording baseline. At the end of the session all participants were asked to fill a feedback form to briefly evaluate the comfort of the experience, the clarity of the instructions and GUI, any difficulty in the annotation task and to report some behavioral preferences during music listening. Finally, they were debriefed on the purpose of the experiment and dismissed. The total length of the session was of 75 minutes on average, with a maximum of 90 minutes in some cases where the calibration of the equipment was not satisfactory, and the participants were then compensated for their participation.

4.2 Data analysis

4.2.1 Data preparation

The first step in the analysis process was to reorganize each participant's dataset in a systematic collection that could be automatically parsed. Due to the pseudo-randomization of the classes and the two different conditions, all trials, white noises, and resting states were flagged using an encoded label through the TriggerBox during the experimental phase. The labels were sent using timed events by the EA app with a precision in the order of milliseconds. The resulting dictionary of events was used to split the EEG recording and extract trials, white noises, and resting

states. Each trial was associated with the appropriate label in the format ”condition/class_*_*” , where condition could be a value between “EO” and “EC” to represent the recording condition, the first * could be a number in the range [1-4] to represent the valence-arousal class, and the second * a letter between “A” or “B” to represent the order of presentation of a song within each trial. For example, ”EO/class_2_A” represents the part A of a trial using music from the LAHV class and recording in EO condition. In addition, for each EEG segment, all the QC labels were saved for later use in the preprocessing pipeline (see Chapter 4.2.2). Then another dictionary containing the order of presentation of the classes was used to associate the metadata saved by the EA app, namely the valence-arousal annotations and the familiarity/liking scores for each song, to the respective entry in the newly organized dataset. During the experiment sessions, some rare bugs in the recording application caused brief interruptions in the recording or the failure to register triggered events. For this reason, a total of 6 participants who had missing parts in their EEG recordings or did not have the dictionary of events were excluded for further analysis. Their data could still be utilized in future studies by synchronizing the splitting functions with the timestamps saved in the metadata, but because of the time cost, it was decided to exclude them for the current research.

4.2.2 Automated Pre-processing Pipeline

Given the goal of estimating the performances of a real-time oriented model, the preprocessing of the data had to consist of a lightweight and automated process that could be integrated in an application at some point. Consequently, more run-in tools like EEGLab and PREP [32], both based on the MATLAB programming language and very popular for offline analysis, were discarded in favor for more real-time oriented tools. Therefore, the Automated Pre-processing Pipeline (AuPP) was implemented as a combination functions of the open-source Python library MNE[3] and the SignalProcessingToolbox (SPT) from myBrainTechnologies, a closed-source library that is more suitable to handle the proprietary data format of Melomind. After loading each participant’s prepared dataset, the AuPP splits the signal in time windows of 5 seconds, removes the DC offset, applies a notch filter to remove power-line noise in the 50Hz and the 100Hz frequency bands, then applies a band-pass filter in the range 0.1Hz - 30Hz to remove slow and possibly large amplitude drifts and some movement artifacts outside of the frequency bands of interest (Fig. 22). Unfortunately, this light preprocessing is not suitable to remove most of the muscular artifacts, especially those generated by ocular movements that are frequently present in the frontal electrodes. In addition, having two electrodes hinders artifact detection

[3]https://mne.tools/stable/index.html

using more sophisticated signal processing algorithms like ICA and Principal Component Analysis (PCA), that require a higher number of electrodes to effectively separate the signal in components and identify artifacts. To deal with artifacts, the AuPP features two methods that can be used independently or in conjunction.

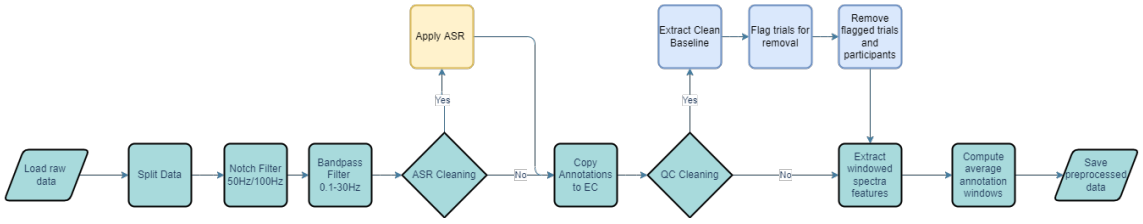

Figure 4.7: Flowchart representing the preprocessing steps of the AuPP.

The first one is Artifact Subspace Reconstruction (ASR), available in the open-source MEEGkit[4] library, that automatically tries to clean the signal by removing transient and large-amplitude artefacts. The second one is a custom method called Quality Index Removal (QIRem), implemented for this study, and based on the QC proprietary classification-based method developed by myBrainTechnologies. The QC algorithm has been developed to support researchers in real-time visual assessment of the quality of the signal [45], and for each second of recording it assigns a label representing the quality of the signal as follows:

- Low Quality: LOW-Q = -1 and 0.
- Medium Quality with muscular artefacts: MED-MUSC = 0.25.
- Medium Quality: MED-Q = 0.5.
- High Quality: HIGH-Q = 1.

The QIRem method takes the QC labels and redistribute them on a simpler scale from 0 to 1, where 0 corresponds to LOW-Q, 0.5 corresponds to MED-MUSC-Q and MED-Q and 1 corresponds to HIGH-Q, and then for each 5 seconds time window it calculates an average of the QC labels.

Figure 4.8: Example of how QIRem flags bad segments flagged for removal.

[4]https://nbara.github.io/python-meegkit/index.html

If the average is below a specified *threshold* parameter, the EEG split is removed from the dataset (Fig. 4.8). After removing all the contaminated splits if the dataset has lost more than the percentage of data amount specified by the *allowed loss* parameter, the entire participant's dataset is flagged for exclusion from the analysis. During the tuning of the AuPP, it was finally decided to avoid using ASR to clean the signal, because it requires to be trained on a sufficiently clean segment of signal, unfortunately not guaranteed to exist for all subjects, and most often resulted in very aggressive cleaning that would flatten the signal (Fig. 4.9).

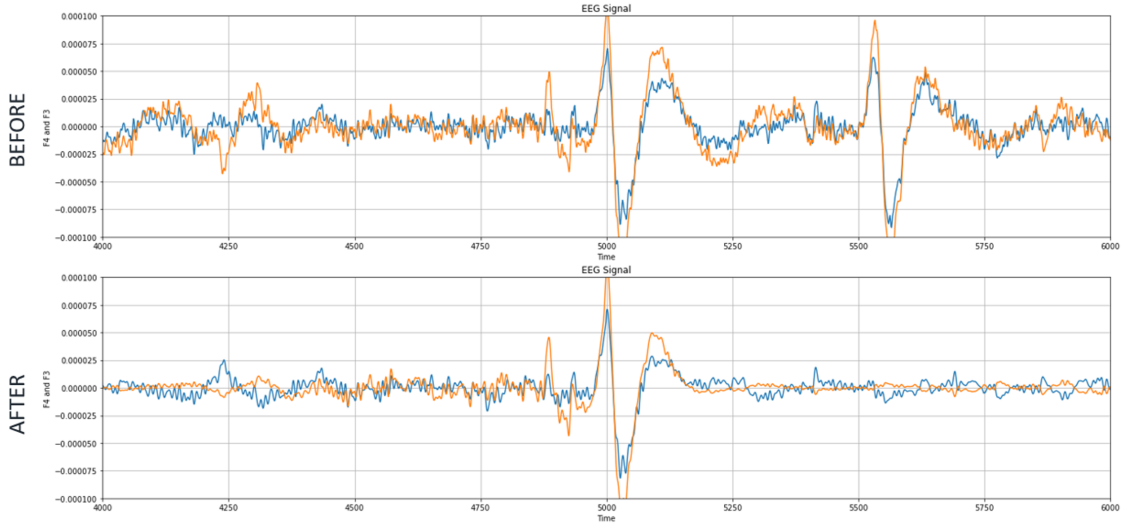

Figure 4.9: Examples of the effects of ASR on the signal.

The QIRem method was setup with *threshold* set to 0.5 and *allowed loss* set to 0.25, meaning that the average quality of each time window had to be equal or above 0.5 in the simplified scale and that at most 25% of data could be pruned before flagging the entire dataset for exclusion. With the current configuration, 10 participants were excluded from further analysis, hence the reason why the threshold was kept around medium quality (0.5), allowing some artefacts to persist in the data. This approach is a compromise choice that carries three main problems that must be addressed in future studies:

- Very aggressive: bad quality data are not cleaned, but removed instead, possibly losing meaningful information and control over the distribution of the class labels.
- Exclusive: 10 out of 39 participants were excluded from analysis, which summed up to those excluded for other reasons is more than 1/3 of the entire experimental dataset.
- Not Optimized: one limitation of the Quality Checker algorithm is that it was

trained for the consumer use on Melomind with electrodes placed on the parietal area of the scalp (P3, P4), so while able to discriminate good and bad quality segments of signal, it has no specific label for ocular artifacts. An upgraded version is under the work to provide classification of these artifacts.

4.2.3 Features Computation

To compute the spectral features, the PSD of each 5 seconds window of the EEG signal was extracted and filtered for theta, alpha and beta frequency band using the SPT wrapper for the FFT. Before computing features, the time-frequency power was normalized using the decibel conversion method as described by M. X. Cohen [46]. Time-frequency power follows a 1/f shape function, meaning that frequency spectrum tends to show decreasing power at increasing frequencies, EEG included. Consequently, there are 4 main limitations:

1. Difficulty in making power comparisons across such bands. Raw power values change in scale as a function of frequency, meaning that lower frequencies (Delta, Theta) will show larger effect than higher frequencies in terms of overall magnitude.

2. Aggregation of subject-independent effects will not yield good results because of differences influenced by skill thickness, sulcal anatomy, cortical surface, recording environment or other internal and external factors.

3. Task-related changes in power can be tainted by background activity, particularly for frequencies that tend to have higher power, especially during baseline periods (Alpha).

4. Raw powers do not follow a normal distribution because they cannot be negative, and they are strongly positively distorted.

Using decibel conversion, which is an expression of power as the ratio between strength of one signal (frequency-band-specific-power) and the strength of another signal (a baseline level of power in the same frequency band).

$$dB_{tf} = 10log10\left(\frac{activity_{tf}}{\overline{baseline_{tf}}}\right)$$

The scale and interpretation of frequency-band-specific power becomes the change in power relative to the baseline. Any frequency-band-specific activity constant over time will be removed, including background activity. As a baseline for normalization, the resting state in the EC condition was used, to prevent ocular artifacts from contaminating the trials. The baseline resting state, previously divided in 5 seconds

time window, and pre-processed together with the other trials, was pruned from low quality segments using the QIRem function and then averaged across all the time windows, for each channel. The main advantages obtainable by normalizing the date are the following:

- All power data are re-scaled to the same scale and thus can be compared visually and statistically
- Normalization computed in respect to a pre-trial baseline enables to disentangle time-frequency dynamics from background or task-unrelated dynamics
- All power results are in a common and easily numerically interpretable metrics
- Parametric statistical analysis is appropriated to use (for baseline-normalized power data normally distributed) and quantitative group-level analyses and integration with other data (behavioral performance, questionnaires) is facilitated.

The features extracted include the neuromarkers described in Chapter 2 and additional properties of the power spectrum that could strengthen the models' ability to discriminate emotional dimensions. It was decided in a later stage to use these properties of the raw signal because they could be conveniently extracted using the SPT and potentially make up for information lost during the computation of the neuromarkers. For each 5s time window the following measurements were computed and stored, for a total of 40 features among the two channels to be used in classification:

1. Normalized power in Theta, Alpha and Beta frequency bands
2. Approach-Withdrawal Index
3. Frontal-Midline Theta Index
4. Spectral Asymmetry Indexes
5. Skewness of the power in Theta, Alpha and Beta frequency bands
6. Kurtosis of the power in Theta, Alpha and Beta frequency bands
7. Standard deviation of the power in Theta, Alpha and Beta frequency bands
8. Ratio of the power in Theta, Alpha and Beta frequency bands
9. Relative spectral difference of the power in Theta, Alpha and Beta frequency bands

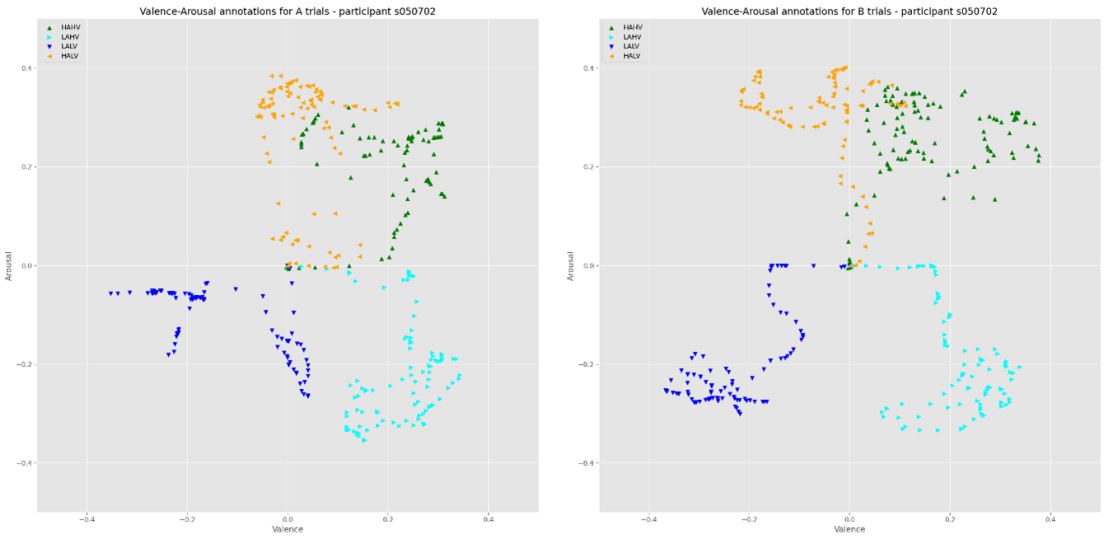

Figure 4.10: Example of annotations of a single participant for all trials. The labels are color coded according to the pre-labelled VA class of each song.

Finally, the raw VA annotations (Fig. 4.10) were averaged for each time-window and then converted into positive labels whether the average was positive, or negative labels otherwise. Consequentially, each time window was labelled twice: first as HA or LA (High/Low Arousal), and then as HV or LV (High/Low Valence). The union of the two labels generates one of the class labels of the Valence-Arousal quadrant that represent the emotion elicited in that specific time window, according to the notation proposed by Koelstra et al. [27] (see Chapter 4.1.3). In some studies, the notation for valence is defined as PV and NV (Positive/Negative Valence), which is better aligned with the etymology of positive and negative emotions. The labels were also copied for each song from the EO listening condition to the respective EC listening condition.

4.2.4 Classification

The classification pipeline was implemented using the open-source Python library Scikit-Learn[5]. Multiple experiments were conducted with two supervised learning models, SVM and MLP. These models are a popular choice for the Emotion-Recognition task thanks to their relative simplicity yet their superior capacity to handle not linearly separable data compared to statistical linear models (see Chapter 3.1). The SVM architecture was defined using RBF kernel, that usually grants better accuracy, and it is relatively easy to calibrate, and decision function one-vs-one for binary classification and one-vs-rest for multi-class classification. The architec-

[5]https://scikit-learn.org/stable/index.html

ture for MLP was based on the LBFGS optimizer, a quasi-Newtonian method, that is more suitable for small datasets and can converge faster with better perforances and *ReLU* was chosen over *TanH* as activation function because it reduces the impact of vanishing gradients, even if no substantial difference was observed while testing both. The problem has been set as a separate binary classification of Arousal and Valence using a subject-dependent strategy, similarly to most related studies. As explained in the Section 4.2.5, during the intermediate experiments the listening condition did not reveal significant differences in the classification performances, therefore all the trials of each subject were later unified under a third condition named "EO&EC" to take advantage of the greater amount of data-points. Then, from each subject dataset, a total of 40 previously computed features were loaded in the classification pipeline. To begin with, the data were split into a training dataset and a test

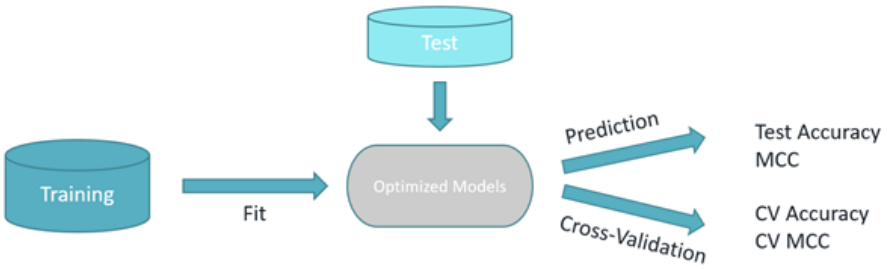

Figure 4.11: Approach used to compute Test Accuracy, MCC, CV Accuracy and CV MCC with separate splits of data.

dataset in 80:20 proportion and scaled using MaxAbsScaler, that re-scales the data to its maximum absolute value without shifting or centering it, thus preserving any sparsity.PCA was used to identify the features contributing for the 95% of the variability of the dataset and projecting them on a lower dimensional space, reducing the dimensionality to 12 components and greatly reducing the overall computation time, often referred to as *curse of dimensionality*. After applying PCA, the training dataset was used to tune the best hyper-parameters each classifier, respectively *C* and *Gamma* for SVM and *Alpha* and *Hidden Layer Sizes* for MLP, using GridSearch with a K-Fold Cross-Validation strategy, k=5 (see Fig, 4.11). Finally, the tuned classifiers were trained with K-Fold Cross Validation leave-one-block-out (LOBO) for testing, and the relevant score metrics were collected.

4.2.5 Intermediate experiments

The classification problem was initially addressed as a binary classification problem for Arousal and Valence and then as a multi-class classification problem for

Valence-Arousal. During the initial experimentation phase, these models were manually tuned and studied keeping the listening condition separated and then adding a third condition defined as EO&EC, composed by data from both listening conditions joint together. To explore each condition and each type of model, a binary classifier for Arousal, a binary classifier for Valence and a multi-class classifier for Valence-Arousal were trained, following both subject-dependent and subject-independent strategy, for a total of 36 possible combinations. All the other hyper-parameters were initially manually tuned and kept identical for each classifier (see Appendix A.3.1). The first full-scale experiment was launched with subject-dependent strategy and to reduce the dimensionality of the data, the 5 most relevant features were selected using forward sequential features selection (SFS) with cross-validation to select the features most contributing to the variability of the data and using them to train each subject's classifiers with cross-validated scores. Consequently, the selected features were ranked by adoption rate among all participants (see Appendix A.3.1) and appointed as TOP5 features. It should be emphasized that this operation was computationally intensive and took several hours to complete, underlying the necessity for a better approach to envision a real-time application. Another subject-dependent experiment was run, this time using the same pre-selected TOP5 features for each participant, followed by a subject-independent experiment with the same configuration with cross-validation and LOSO. The results of all experiments have been collected and averaged (see Appendix A.3.1, A.3.2, A.3.4). A few observations were made that led to further investigation and the search for a better approach:

- EC and EO conditions did not show significant differences in performance, suggesting that the study could continue using just the EO&EC condition, greatly reducing computation time and simplifying the analysis.
- The multi-class classifiers for VA reported the worst performances, thus was discarded to focus the efforts on optimizing the binary classifiers.
- Many datasets were highly unbalanced in the distribution of the labels among the four VA classes and both MLP and SVM struggled to discriminate positive and negative class, with SVM always default guessing the majority class.
- Subject-Independent performance always resulted very close and sometimes worse than default guessing the majority class.
- Selecting the average TOP5 features among all participants is idealistically a good choice for a subject-independent strategy, but given the more promising subject-dependent results, this approach is at best losing data variability for some of the subjects and at worse using the least contributing features for some others.

Further investigation in the subjects labelling behavior and in the models' predictions (see Appendix A.3.4) led to the decision to optimize the strategy in function of subject-dependent classification.

4.2.6 Unbalanced labelling

The stimuli were selected to elicit a wide as possible range of emotions to cover evenly the Valence-Arousal quadrants. However, when selecting music that has been pre-labelled with the average opinion of hundreds of annotators, there is always a possibility that part of the population disagrees. These subjects are not per se "bad annotators", but their personal taste and perception lies outside of median. The distribution of labels across all subjects in Figure 4.12 also reveals that positive classes are in general the most reported, with HAHV and LAHV taking absolute lead, and corresponding to emotions very common during the music listening experience: excitement, happiness, satisfaction, calm etc.

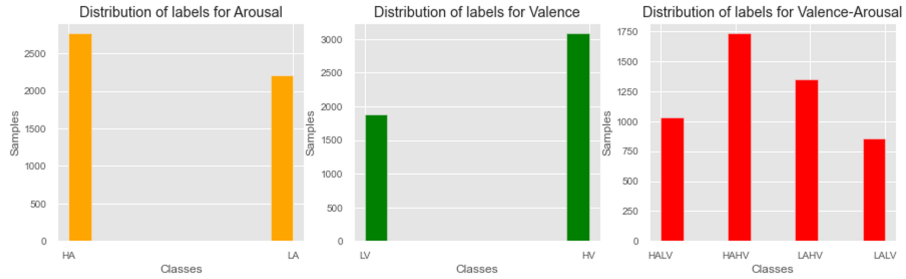

Figure 4.12: Distribution of arousal, valence, and Valence-Arousal labels across all subjects.

Furthermore, the personal perception of a subject can be heavily biased by external factors such as memories, genres preferences or unexpected events occurred over the day that are outside of the experiment control capabilities. In some cases, this led to very extreme distributions of data (Fig. 4.13).

Although just a few data points of this subject lie in the negative spectrum of Arousal and Valence, the continuous annotations allow to capture them while discrete annotations at the end of each song might have ended up all in the HAHV class, thus hindering any classification. However, a classifier trained to obtain the maximum accuracy would always overfit and opt to classify the majority class, misleading the interpretation of its performances.

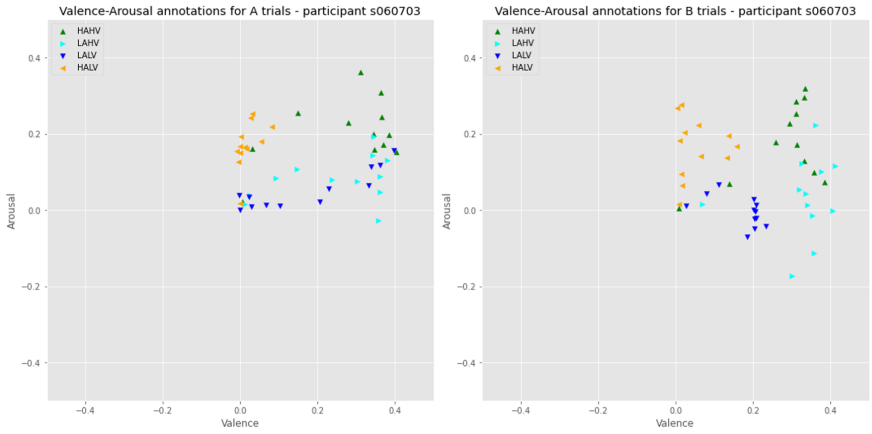

Figure 4.13: Annotations averaged over a 5 second window for a subject with unbalanced dataset. The labels are color coded according to the prelabelled VA class of each song.

4.2.7 Optimizations

All the problematics were addressed before proceeding with the final experiment. First, the dimensionality of the features was reduced using PCA to select the 95% most contributing components, greatly reducing the computational time and the risk of excluding meaningful features. Then MCC [47] was introduced as scoring parameter to provide better interpretability of the accuracy scores. The MCC is a correlation coefficient between the observed and predicted binary classification and is defined by the following formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC value ranges between -1 and +1, where +1 represents a perfect prediction, 0 a random prediction and -1 and inverse prediction. Unlike the F1 score, that is the harmonic mean of precision and recall, the MCC considers all the four quadrants of the confusion matrix of a prediction, making it a more reliable measure for the learning performances of a model for binary classification, even when the dataset is unbalanced [48]. The large number of factors that may hinder the Emotion-Recognition task are reflected in the large variability of the classification results across subjects and thus using solely the accuracy to measure the performances is not optimal and more studies [31], [33] now relying on MCC score's reliability to explain the learning capabilities of their models. To optimize the models for the subject-dependent strategy, GridSearch was used with K-Fold Cross-Validation on the training split of each subject dataset to find the configuration that would yield the highest MCC score (see Fig. 4.14).

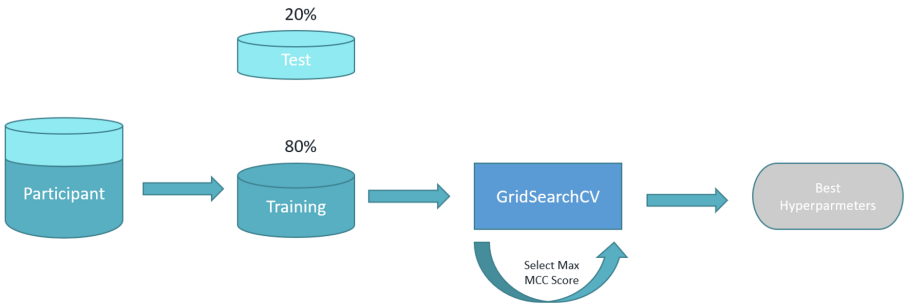

Figure 4.14: Scheme of data splitting strategy for cross-validated GridSearch optimization.

In addition, SVM models were also initialized with the weights of each class based on the distribution of the labels. Currently there is a known limitation in the Scikit-learn library, as it does not seem to support custom loss functions to optimize scoring parameters other than accuracy, a missing feature reported by users of the library [49], [50]. Consequently, when feeding MCC as scoring parameter, it only means that after optimizing accuracy for all possible configurations, the configuration with highest MCC is selected. Two scoring strategies for GridSearch were tested in the intermediate experiments:

- Maximization of MCC score, defined as “Max MCC” strategy
- Maximization of Accuracy score, defined as “Max Accuracy” strategy

Maximising accuracy scores is often desirable in machine learning because it gives a clear idea of a model’s ability to predict a class. A ”good” accuracy in most cases should be above 90%, while for an optimal score should be above 99%. However, given the presence of many unbalanced datasets in the current research, the ”Max Accuracy” strategy for GridSearch generated more often over-fitted models that would have higher accuracy, at the cost of never predicting the minority class, especially with SVM classifiers (see Appendix A.3.5). With MLP classifiers this phenomenon was less evident, also because they already produced a larger number of over-fitted and under-fitted models than their SVM counterparts, however in spite of using the same strategy and MCC as principal metric to show evidences of learning capabilities, the ”Max MCC” strategy was chosen to proceed with the final experiment.

Chapter 5

Results

In this chapter the results of the classification using “Max MCC” optimization strategy are presented. Some premises are necessary before proceeding with the interpretation of the data:

1. The average scores at the bottom of each table are between models with the same architecture, but different hyper-parameters. Furthermore, as part of the ”Max MCC” strategy, having an accuracy lower than the majority class guessing (defined as ”chance level” in the tables) but with a positive MCC has been preferred compared to having the highest possible accuracy with a zero or negative MCC.

2. All models are underfitting or overfitting to some degree on their dataset but, considering the challenges of classifying emotions, models with positive MCC and cross-validated MCC (CV MCC) are reported as good models regardless their scoring. As a consequence, only extreme cases of overfitting, i.e. models with 0 or negative MCC and positive CV MCC, and extreme case of underfitting, i.e. models with 0 or negative CV MCC are labelled as such in tables 5.1 and 5.2.

3. CV Accuracy, CV Acc Std, CV MCC and CV MCC Std stand for Cross-Validated Accuracy and MCC and their respective standard deviation.

4. Test accuracy and MCC are computed using an unseen test split of the data. CV Accuracy and CV MCC have been computed on the training split of the data; thus, their purpose is to provide a measure of consistency of the test scores (see Fig 4.11).

5. The confidence interval has been calculated on the test split, Z=1.96.

6. Chance level represents the default guessing of the majority class.

7. Standard deviation of each measure, if present, is between round parenthesis in the tables.

First, the classification performances of SVM and MLP are compared for both arousal and valence classification. Then, the results are compared with related work.

5.1 Support-Vector Machines vs Multi-Layer Perceptron

The results of the subject-dependent arousal classification experiment are reported in Table 5.1, while the results for valence classification are reported in Table 5.2. In these tables, learning models can be identified by a positive MCC score supported by a positive CV MCC and are highlighted in blue. The models that over-fitted, highlighted in yellow, are characterized by negative MCC and positive CV MCC, meaning that they learned well on the training data but could not discriminate classes on the test data. Under-fitted models instead are characterized by zero or negative CV MCC and positive or negative MCC, because they did not have the adequate capabilities to capture the underlying structure of the training data, but possibly obtained a good test accuracy by random guessing and they are highlighted in orange. In arousal classification the average majority class guessing (defined as chance level in the table) is $58 \pm 8\%$, the highest and consistent test accuracy score is 84% with MCC score of 0.20 using SVM and 88% with MCC score of 0.78 using MLP. For SVM classifiers, the average test accuracy is $61 \pm 9\%$ with average MCC of 0.16 ± 0.20, $61 \pm 6\%$ cross-validated (CV) accuracy and CV MCC of 0.24 ± 0.12. For MLP classifiers, the average test accuracy is $58 \pm 12\%$, with average MCC of 0.13 ± 0.20, $57 \pm 8\%$ CV accuracy and CV MCC of 0.15 ± 0.16.

In valence classification (see Table 5.2) the average majority class guessing (defined as chance level in the table) is $65 \pm 12\%$, the highest and consistent test accuracy score is 89% with MCC score of 0.27 using SVM and 77% with MCC score of 0.48 using MLP. For SVM classifiers, the average test accuracy is $67 \pm 12\%$ with average MCC of 0.13 ± 0.18, $61 \pm 6\%$ CV accuracy and CV MCC of 0.26 ± 0.13. For MLP classifiers, the average test accuracy is $65 \pm 12\%$, with average MCC of 0.11 ± 0.18, $56 \pm 9\%$ CV accuracy and CV MCC of 0.13 ± 0.18. The large variances are affected by the unbalanced distribution of positive and negative classes (see Chapter 4.2.6), but it are also a clear symptom of diffused over-fitting, especially for MLP classifiers.

A final remark can be obtained by comparing the number of learning, over-fitting and under-fitting models between the two type of classifiers (see Fig. 5.1. In arousal classification it was possible to train 25 SVM learning classifiers, with 4 over-fitting

Table 5.1: Arousal classification results using MCC as scoring parameter for GridSearch. Learning models are highlighted in blue, over-fitted and under-fitted models are highlighted in yellow and orange, respectively.

Arousal	SVM							MLP							EXTRA		
Participant	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Chance Level	Avg Familiarity	Avg Liking
s010701	0.50	-0.07	0.66	0.08	0.37	0.20	0.16	0.66	0.32	0.65	0.09	0.30	0.19	0.16	0.57	2.49	2.51
s010702	0.53	0.12	0.59	0.04	0.26	0.07	0.16	0.61	0.23	0.61	0.12	0.23	0.23	0.16	0.54	2.87	3.97
s010703	0.57	0.13	0.56	0.07	0.12	0.13	0.16	0.57	0.09	0.55	0.09	0.09	0.17	0.16	0.65	2.04	3.55
s010704	0.56	0.13	0.59	0.06	0.18	0.12	0.16	0.49	-0.02	0.48	0.05	-0.04	0.10	0.15	0.51	1.00	2.56
s020702	0.63	0.30	0.64	0.03	0.30	0.06	0.17	0.63	0.30	0.60	0.08	0.21	0.17	0.17	0.52	2.71	3.71
s020703	0.71	0.41	0.61	0.10	0.22	0.20	0.15	0.65	0.28	0.49	0.11	-0.04	0.21	0.16	0.57	1.92	3.52
s020704	0.72	0.41	0.64	0.10	0.32	0.23	0.18	0.88	0.78	0.63	0.11	0.28	0.24	0.18	0.65	2.09	2.47
s050702	0.63	0.26	0.54	0.09	0.09	0.20	0.16	0.57	0.16	0.43	0.09	-0.14	0.19	0.16	0.52	1.96	2.70
s050704	0.53	0.11	0.65	0.07	0.31	0.15	0.17	0.50	0.07	0.65	0.06	0.30	0.13	0.17	0.54	1.88	2.97
s060703	0.84	0.20	0.53	0.14	0.08	0.32	0.13	0.87	0.00	0.58	0.14	0.20	0.29	0.15	0.86	2.35	3.68
s070702	0.44	-0.14	0.59	0.08	0.18	0.16	0.17	0.59	0.15	0.49	0.10	-0.03	0.21	0.17	0.58	2.08	3.30
s170601	0.65	0.22	0.52	0.03	0.05	0.07	0.15	0.43	-0.05	0.47	0.09	-0.06	0.20	0.16	0.58	2.39	2.91
s210602	0.60	0.05	0.61	0.07	0.24	0.16	0.18	0.60	0.16	0.61	0.09	0.24	0.20	0.18	0.59	2.55	2.60
s220602	0.63	-0.22	0.58	0.05	0.20	0.11	0.15	0.47	-0.06	0.55	0.08	0.11	0.16	0.15	0.72	2.17	3.17
s230602	0.53	0.08	0.63	0.03	0.35	0.11	0.16	0.53	0.06	0.64	0.07	0.30	0.16	0.16	0.59	3.12	3.37
s230603	0.77	0.56	0.79	0.09	0.57	0.18	0.15	0.71	0.42	0.79	0.03	0.60	0.05	0.16	0.52	1.67	2.66
s230604	0.56	0.12	0.64	0.10	0.28	0.21	0.17	0.35	-0.29	0.64	0.04	0.29	0.08	0.17	0.5	2.53	3.15
s250601	0.61	0.18	0.60	0.04	0.20	0.08	0.16	0.58	0.12	0.52	0.12	0.05	0.24	0.16	0.53	1.62	3.24
s250602	0.69	0.52	0.65	0.10	0.34	0.21	0.15	0.69	0.42	0.66	0.07	0.33	0.14	0.13	0.56	2.78	3.69
s250604	0.56	0.06	0.60	0.05	0.27	0.09	0.16	0.56	0.00	0.50	0.00	0.00	0.00	0.16	0.58	1.99	3.48
s280601	0.65	0.30	0.56	0.06	0.12	0.13	0.15	0.59	0.13	0.54	0.08	0.07	0.17	0.15	0.55	1.79	2.86
s280603	0.50	0.04	0.58	0.05	0.18	0.11	0.16	0.33	0.00	0.50	0.00	0.00	0.00	0.16	0.53	3.34	3.69
s280604	0.61	0.02	0.60	0.07	0.22	0.15	0.16	0.53	0.00	0.55	0.05	0.11	0.09	0.16	0.63	1.95	3.88
s290601	0.62	0.14	0.58	0.09	0.17	0.19	0.16	0.49	-0.10	0.62	0.07	0.24	0.14	0.16	0.62	3.10	4.27
s290602	0.68	0.36	0.68	0.05	0.37	0.10	0.15	0.59	0.19	0.63	0.14	0.26	0.28	0.15	0.51	2.39	3.64
s290603	0.72	0.43	0.73	0.08	0.46	0.15	0.16	0.56	0.13	0.64	0.07	0.30	0.14	0.17	0.51	2.51	3.15
s290604	0.58	0.19	0.54	0.06	0.12	0.15	0.17	0.48	-0.01	0.52	0.04	0.08	0.15	0.17	0.53	1.77	4.24
s290605	0.67	0.02	0.57	0.09	0.16	0.25	0.16	0.67	0.19	0.53	0.04	0.08	0.09	0.17	0.65	1.68	3.50
s300602	0.43	-0.19	0.53	0.02	0.11	0.08	0.16	0.54	0.13	0.47	0.12	-0.06	0.25	0.16	0.55	2.87	3.88
Average (Std)	**0.61 (0.09)**	**0.16 (0.20)**	**0.61 (0.06)**		**0.24 (0.12)**		**0.16 (0.01)**	**0.58 (0.12)**	**0.13 (0.20)**	**0.57 (0.08)**		**0.15 (0.16)**		**0.16 (0.01)**	**0.58 (0.08)**	**2.26 (0.53)**	**3.32 (0.52)**

Table 5.2: Valence classification results using MCC as scoring parameter for GridSearch. Learning models are highlighted in blue, over-fitted and under-fitted models are highlighted in yellow and orange, respectively.

Valence	SVM							MLP							EXTRA		
Participant	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Chance Level	Avg Familiarity	Avg Liking
s010701	0.50	0.06	0.66	0.07	0.34	0.13	0.16	0.53	0.04	0.60	0.07	0.21	0.14	0.15	0.51	2.49	2.51
s010702	0.78	0.07	0.58	0.10	0.14	0.18	0.14	0.86	0.00	0.45	0.06	-0.09	0.10	0.11	0.78	2.87	3.97
s010703	0.49	-0.11	0.60	0.07	0.20	0.14	0.16	0.59	0.18	0.52	0.06	0.04	0.14	0.16	0.65	2.04	3.55
s010704	0.36	-0.28	0.60	0.09	0.22	0.19	0.15	0.46	-0.06	0.61	0.05	0.23	0.09	0.16	0.52	1.00	2.56
s020702	0.73	-0.10	0.60	0.07	0.31	0.17	0.16	0.73	-0.10	0.59	0.10	0.22	0.26	0.16	0.69	2.71	3.71
s020703	0.76	-0.09	0.59	0.07	0.26	0.20	0.14	0.71	0.10	0.56	0.08	0.14	0.18	0.14	0.8	1.92	3.52
s020704	0.68	0.36	0.59	0.07	0.27	0.17	0.18	0.60	0.14	0.43	0.14	-0.14	0.30	0.18	0.58	2.09	2.47
s050702	0.66	0.34	0.54	0.10	0.08	0.19	0.16	0.71	0.42	0.47	0.11	-0.06	0.21	0.15	0.63	1.96	2.70
s050704	0.88	-0.05	0.65	0.19	0.33	0.40	0.11	0.91	0.00	0.50	0.00	0.00	0.00	0.10	0.91	1.88	2.97
s060703	0.87	-0.06	0.62	0.13	0.28	0.29	0.12	0.87	-0.06	0.48	0.02	-0.03	0.04	0.14	0.92	2.35	3.68
s070702	0.59	0.19	0.64	0.13	0.28	0.27	0.17	0.72	0.44	0.55	0.10	0.11	0.20	0.17	0.53	2.08	3.30
s170601	0.68	0.26	0.74	0.04	0.50	0.09	0.15	0.68	0.45	0.71	0.06	0.44	0.13	0.15	0.59	2.39	2.91
s210602	0.70	0.41	0.57	0.05	0.23	0.13	0.16	0.47	-0.14	0.55	0.08	0.11	0.17	0.18	0.57	2.55	2.60
s220602	0.66	-0.07	0.56	0.06	0.17	0.13	0.15	0.61	0.09	0.53	0.02	0.07	0.05	0.16	0.64	2.17	3.17
s230602	0.58	0.19	0.69	0.06	0.39	0.12	0.16	0.55	0.17	0.67	0.07	0.34	0.14	0.16	0.51	3.12	3.37
s230603	0.61	0.23	0.74	0.04	0.49	0.10	0.17	0.61	0.23	0.73	0.06	0.47	0.12	0.17	0.51	1.67	2.66
s230604	0.65	0.25	0.62	0.06	0.26	0.12	0.16	0.65	0.25	0.62	0.04	0.25	0.09	0.17	0.6	2.53	3.15
s250601	0.66	0.30	0.64	0.05	0.35	0.11	0.15	0.58	0.14	0.62	0.06	0.24	0.13	0.15	0.58	1.62	3.24
s250602	0.66	0.24	0.75	0.06	0.51	0.12	0.16	0.77	0.48	0.67	0.07	0.35	0.14	0.14	0.53	2.78	3.69
s250604	0.75	0.26	0.62	0.11	0.31	0.25	0.14	0.64	0.01	0.62	0.11	0.26	0.22	0.15	0.67	1.99	3.48
s280601	0.59	0.31	0.61	0.08	0.28	0.16	0.16	0.62	0.35	0.64	0.08	0.31	0.17	0.16	0.64	1.79	2.86
s280603	0.58	0.21	0.59	0.08	0.18	0.16	0.16	0.58	0.16	0.47	0.06	-0.07	0.13	0.16	0.51	3.34	3.69
s280604	0.55	0.00	0.51	0.02	0.06	0.08	0.16	0.50	-0.04	0.38	0.08	-0.24	0.17	0.16	0.56	1.95	3.88
s290601	0.89	0.27	0.50	0.08	0.02	0.20	0.10	0.84	0.17	0.47	0.06	-0.06	0.12	0.13	0.83	3.10	4.27
s290602	0.84	0.00	0.51	0.04	0.03	0.17	0.12	0.62	-0.23	0.48	0.10	-0.02	0.19	0.14	0.79	2.39	3.64
s290603	0.63	0.29	0.59	0.08	0.28	0.25	0.17	0.53	0.01	0.52	0.09	0.08	0.22	0.17	0.65	2.51	3.15
s290604	0.65	0.20	0.64	0.06	0.29	0.10	0.17	0.61	0.16	0.61	0.07	0.25	0.14	0.16	0.61	1.77	4.24
s290605	0.67	0.14	0.59	0.05	0.19	0.10	0.16	0.67	0.14	0.56	0.02	0.13	0.04	0.16	0.66	1.68	3.50
s300602	0.74	0.04	0.66	0.06	0.34	0.11	0.14	0.60	-0.23	0.65	0.09	0.28	0.17	0.14	0.75	2.87	3.88
Average (Std)	**0.67 (0.12)**	**0.13 (0.18)**	**0.61 (0.06)**		**0.26 (0.13)**		**0.15 (0.02)**	**0.65 (0.12)**	**0.11 (0.18)**	**0.56 (0.09)**		**0.13 (0.18)**		**0.15 (0.02)**	**0.65 (0.12)**	**2.26 (0.53)**	**3.32 (0.52)**

models and 0 under-fitting models against 15 MLP learning classifiers, 6 over-fitting models and 8 under-fitting models. In valence classification it was possible to train 20 SVM learning classifiers, with 9 over-fitting models and 0 under-fitting models against 17 MLP learning classifiers, 4 over-fitting models and 8 under-fitting models. Overall, the classification performances between SVM and MLP are not significantly different, but SVM proved to be more reliable and obtained average classification accuracy above the majority default guessing threshold and generated more models able to discriminate unseen data. The are other model-specific advantages and

disadvantages a to be taken into account in the implementation of a real-time application, for example most MLP implementations support partial fitting of the models that is very useful with continuous streams of data, however these technical considerations lie outside the scope of this research.

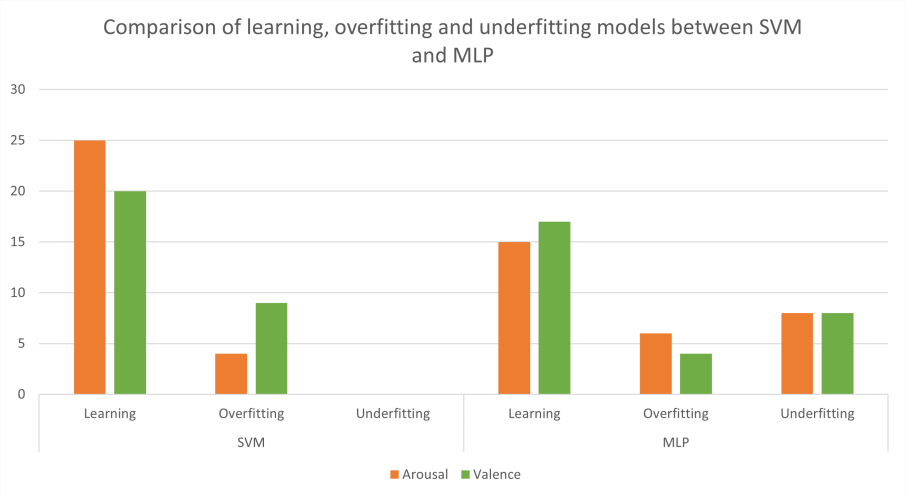

Figure 5.1: Comparison of SVM and MLP in terms of number of learning, overfitting and underfitting models.

5.2 Comparison with related work

Comparing the result of the current research with related work is non-trivial because of the methodological differences in data collection, processing, and evaluation. These differences have been considered to sort the comparisons from most comparable to least comparable.

The first and most comparable study [31] used self-reported continuous annotation as main labelling method for classification, the data were collected from 9 subjects with a wearable EEG headset equipped with 8 dry electrodes and processed using the automated PREP pipeline in MATLAB. The accuracy scores of SVM classifiers using EEG and LOBO cross-validation are only reported through plots, valence classification scored average accuracy of $68 \pm 10\%$ and arousal classification scored average accuracy $64 \pm 10\%$. The comparison only takes into account results from SVM classifiers, which in the current study reports average test accuracy of $67 \pm 12\%$ and average LOBO cross-validated accuracy on the training set of $61 \pm 6\%$ for valence classification . Arousal classification scored an average test accuracy of $61 \pm 9\%$ and average LOBO cross-validated accuracy of $61 \pm 6\%$. They provided a table reporting the MCC scores for each classification modality, the average CV MCC reported for valence is 0.247 ± 0.17 and the average CV MCC for arousal is 0.177 ± 0.04. The

current study reports average test MCC score of 0.13 ± 0.18 and average CV MCC score of 0.26 ± 0.13 for valence, while for arousal the average test MCC score is 0.16 ± 0.20 and average CV MCC score is 0.24 ± 0.12. Finally, the highest CV MCC score reported for a single subject is 0.596 ± 0.30 for valence and 0.23 ± 0.22 for arousal, while in the current study the highest CV MCC score reported is 0.51 ± 0.12 for valence and 0.57 ± 0.18 for arousal. These results are aligned with the current study and provide individual insights for each subject that can be easily compared, so in conclusion if we consider only the single EEG modality from [31], the learning capabilities of the models trained in the current study are similar despite a lighter pre-processing and a lower number of electrodes. This comparison suggests that there definitely is room for improving the software pipeline on the current dataset, especially the AuPP, to generate more stable models with lower variance and risk of over-fitting and under-fitting even before considering an improvement of the hardware equipment and another data collection phase.
The second comparable study [26] is the one that inspired the self-reporting of emotions using continuous annotation. The focus of this study, however, was to compare traditional discrete annotations to continuous annotation and to evaluate two approaches for features extraction. For comparison purposes, the scores reported usingPSD features will be considered instead of the scores obtained with FD features. The accuracy scores, and relative standard deviations are mostly reported in plots and partially during the discussion, so an estimate is provided. Using SVM, valence classification score is $81.2 \pm 8\%$ average CV accuracy and arousal classification score is $75 \pm 8\%$ average CV accuracy. With MLP, valence classification score is $80.2 \pm 10\%$ average CV accuracy and arousal classification score is $75 \pm 10\%$ average CV accuracy. These scores are significantly higher than default guessing of the majority class and clearly outperform the results obtained in the current study, that are not on average significantly different than default guessing. No further insights are provided on subject-dependent performances, nor on MCC scores, but their models are consistently reliable in performing better than default guessing. This comparison suggest that the choice of stimuli, i.e. music with or without lyrics and the distribution of emotional classes, can highly impact on the classification performances. Furthermore, improving the features engineering part to take full advantage of the continuous annotations, for example applying a sliding window segmentation technique on the EEG data, can help better capturing the underlying emotional fluctuation. Finally, they also dealt more aggressively on unbalanced datasets, for example by removing those subjects where one of the two binary classes (either in arousal or valence classification) was missing.
The third study [25] aimed at artificially simulating a wearable device by selecting, 2, 4 and then 8 frontal electrodes from the subjects of the DEAP [27] dataset, col-

lected with and EEG system with 32 wet electrodes and discrete self-reported labels. The asymmetry indexes used in this study are sim Only the results for the 2 electrodes configuration using SVM for subject-dependent classification are reported and no specific description of the preprocessing pipeline is provided, so the assumption is that they used the data already preprocessed by the original authors of the dataset. The average CV accuracy scored for valence classification is $68.4 \pm 2.0\%$. The authors also present better accuracy scores using GBDT and Random Forests classifiers, achieving respectively $75.10 \pm 2\%$ and $72.15 \pm 2\%$ accuracy for subject-dependent classification, and subsequently reported 61.82%accuracy for subject-independent classification using GBDT. A balancing strategy for the labels is explained, but no specific distribution of positive and negative class is provided, nor individual insights for subject-dependent classification so it is not possible to evaluate the reliability of the accuracy scores compared to default guessing the majority class. In conclusion, even assuming an equal distribution of positive and negative labels, this study results with SVM are in line with the $67 \pm 12\%$ average test accuracy in valence classification of the current research, but with a definitely lower average variance that suggests that models trained were less prone to over-fitting. In addition, this comparison suggest that other supervised learning algorithms that are constituted by ensembles of weaker classifiers, for example Random Forests and GBDT (that is conceptually similar), can obtain much better performances than simpler models such as SVM and MLP.

The fourth and last comparable study [23] collected self-reported continuous annotations from 26 subjects using a standard EEG system with 32 wet electrodes. The data were preprocessed with visual inspection on EEGLab and then features were extracted from 12 pairs of symmetrical electrodes. The authors provide 3 classifications schemes, but only the “one-against-one scheme is comparable in terms of binary classification of valence and arousal and therefore reported. The reported average CV accuracy score for valence classification is $94.86 \pm 17.6\%$ and for arousal classification is $94.43 \pm 21.2\%$. The authors did not report the distribution of the classes for each subject so it is not possible to evaluate the reliability of the accuracy scores for each subject. The high average variance suggests that they also generated a number of over-fitted models, similarly to the current study, nevertheless their follow-up study [24] applied a better features engineering process that lowered the overall variance that enabled them to obtain a lower but more stable average classification accuracy of $82.29 \pm 3.06\%$ of four emotional states (joy, anger, sadness and pleasure) using SVM and even to identify 30 subject-independent features relevant to emotional processing across subjects. These studies by Lin et al. report interesting strategies for selecting and ranking EEG features for the transition from subject-dependent classification to subject-independent classification, however af-

ter experimenting with electrodes reduction their models performances dramatically dropped when the electrodes number was lower than 18 (9 pairs), suggesting that subject-independent classification is still out of reach of wearable devices like Melomind.

Chapter 6

Discussion

In this chapter, the research results are discussed and contextualized with the research questions (see Chapter 1) and the idea that inspired this work: evaluating the feasibility of performing real-time Emotion-Recognition with a wearable device for a future affective BCI system.

6.1 Challenges towards real-time Emotion-Recognition

When imagining users adopting a technology in their life, many functional and design aspects are rightfully expected. For example, the design of the device needs to be sleek and intuitive, and the flow of the application must be seamlessly integrated and well performing. This is clearly not the case for current Brain-Computer Interfaces, that are still in their infancy and far from mass adoption, especially for non-clinical applications. Lin, Jung and Onton [51] reviewed a collection of methods that could greatly improve the quality of the user experience and finally open the way for affective BCI to reach the consumer market. The core challenge of affective BCI is to create a plug-and-play BCI system with limited electrodes that can consistently perform accurate Emotion-Recognition regardless of the person that is using it. The other challenges consequently follow:

- **Reduction and selection of electrodes:** the number of electrodes must be relatively small, between 2 and 8, and strategically placed over the cortical areas delegated to the processing of emotions. They also should be soft dry electrodes and the system should be able to automatically recognize bad channels and excluding them from the processing.

- **Automatic artifacts cleaning**: artefacts are one of the most impacting problems, and currently the most popular approaches for offline artifact cleaning are ICA, that is not applicable to small sets of electrodes and too computa-

tionally expensive for online cleaning, and visual inspection. Alternative approaches might be achieved using regression to train algorithms in recognizing specific types of artifacts or features of the signal that indicate whether the quality is good or not and adjust it accordingly. Cleaning artifacts is also vital to the efficient use of as much data as possible. In the current study the artifacts removal strategy led to the exclusion of 10 participants and the removal of up to 25% of the data of the remaining participants, a considerable cost considering the time and efforts needed to collect the data with an experimental protocol.

- **Features selection and reduction**: selecting the most suitable features allows to reduce the computational cost by removing redundant and less meaningful features, and this is particularly correlated to the selection of electrodes as well. Lin et al. [51] through their extensive research over the years discovered that differential asymmetries are the most consistent type of features among subjects and sessions, reliable also for subject-independent classification. In more recent studies, Thammasan et al. [26], Keelawat et al [33] and Avramidis et al. [52] , extracted EEG features using fractal dimension algorithms instead of PSD, obtaining significantly better performances in both subject-dependent and subject-independent classification. Another possibility is to use more complex models like CNN [33] to automatically extract features from the EEG signal, sacrificing the ease of interpretability of the model and the direct connection with the theoretical neuroscience.

- **Users training and calibration**: a very time consuming and frustrating process is the calibration of BCI systems for new users, especially if they have no experience of brain-controlled inputs. A system that can only classify emotions with a subject-dependent strategy is bound to train the users in reporting their emotions and then train a classifier with the collected data, all under the assumption that the resulting dataset is not too unbalanced. This of course is not feasible, and ultimately subject-independent classification should be the aim for real-time systems. However, this might not suffice, and even subject-independent classifiers might have to be tuned with short calibration sessions to adjust for each specific case, for example by selecting more susceptible features for a certain user that matches similar brain “signatures” from a group of users that the model has already been trained on. Zero-training strategies [53], [54] have been object of study over the last decade and using spatial filters and transfer learning makes it possible to train a sub-optimal decoder and then use unsupervised learning to transform it into a user specific decoder. Of course, these solutions are still very experimental and use case specific, and further investigation is required.

The rest of the chapter will contextualize these challenges in the current study. Using a wearable EEG is a solution to the first problem, but also a constraint around which every other problem had to be worked around.

6.2 Self-reported emotional labels

Reporting emotions is a non-trivial task and could be subjectively difficult. From the feedback forms collected after the experiment, 7 subjects reported to have found difficulties in choosing the quadrant of the valence-arousal space, 8 subjects reported difficulties in assessing the emotional intensity (in both dimensions) and 9 subjects reported difficulties in assessing the specific emotions. In addition, participants reported an average fatigue score of 2.47 ± 0.97 after the first half of the experiment and a fatigue score of 3.27 ± 0.86, on a scale from 1 to 5. Some emotions were reported to be missing from the simplified valence-arousal space, for example ”annoyance”. Another missing emotion was “bittersweet”, that is commonly experienced in music listening yet difficult to report because of its composition of not-adjacent emotions (sadness, happiness) on the valence-arousal space. Continuous annotation of emotions has several advantages over discrete annotation, allowing a more granular reporting that consider emotional “oscillations” over the duration of a stimulus and more distributed labels across all classes that can favor the classification task. It is a powerful tool for researchers to build affective datasets but heading towards a real product it will eventually be an obstacle as it requires an extra layer of training before the calibration. Discrete annotations are a simpler task that has virtually no impact on the recording phase and can be more easily integrated in a calibration tool, but seemingly yielding poorer performances compared with the continuous approach [26]. An example to compromise the benefits of both approaches could be a subject-independent classifier trained using an offline dataset collected with continuous labelling, then discrete annotations collected by an online system would be used to integrate subject-specific differences during a calibration phase.

6.3 Familiarity, liking and PANAS

During the experiment, participants were asked to rate every song they listened to in terms of how much they liked the song referred to as “liking” and how much the song was familiar to them referred to as “familiarity, both on a scale from 1 to 5 where “3” represents neutrality. These scores could be used in a music recommending system to enhance the quality and relevancy of recommendations, however for the current research they were only used to assess the impact of the selected stimuli. The aver-

age “liking” score across all trials and all participants 3.32 ± 0.52, indicating that the selection has been in general positively perceived. Average “familiarity” score was 2.26 ± 0.53, quite below neutrality despite most of the songs being from internationally known artists. Having general low familiarity is positive for classification as emotional biases caused by memories can occur. Looking at Tables 5.1 and Table 5.2 in the previous chapter it is also possible to observe that some of the participants with highest average “familiarity” are also among the worst classification scores, but the analysis was out of the scope of this research. Further investigation is needed regarding the correlations of “liking” and “familiarity” with the emotional dimensions of valence and arousal. Before each session, participants were also asked to fill a PANAS questionnaire.

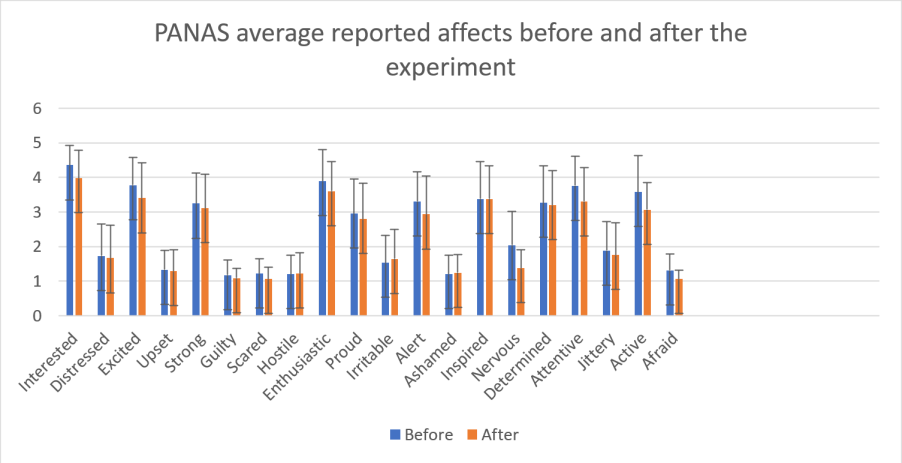

Figure 6.1: Average reported affects before and after the experimental session.

Comparing the assessed affects before and after the experiment (Fig. 6.1), it is possible to observe that general interest, enthusiasm, and excitement of the participants decreased, which was expected considering that for the majority of the participants it was the first time participating in an EEG experiment. Attentiveness and activeness also significantly decreased, suggesting that the task proposed, and the length of the session were to some degree causing fatigue. Follow-up experiments can account for these effects by shortening the length of musical excerpts, reducing the number of conditions, or distributing the data collection from each participant over multiple sessions.

6.4 Features selection and performances evaluation

The proposed principal features based on previous findings on differential and rational asymmetries were the neuromarkers: AWI, FMTI and SASI. The frequency-band specific features of the EEG signal were extracted using the SPT proprietary

library from myBrainTechnologies. Intermediary experiments using forward SFS revealed a general trend in selecting the neuromarkers as most significant features, with some exceptions. PCA was then used instead of SFS as better performing subject-dependent features selection method to reduce the dimensionality of the features vector in lower computational time. It was not possible to directly compare the impact of using neuromarkers instead of frequency-band specific features, but the results were very promising for a subset of participants. The evaluation of model performances using MCC score led to better understanding and correcting models who were over-fitting toward the majority classes because of uneven distributions of labels. It made possible to easily identify under-fitting models too, however no solution was found to prevent it. Overall, MCC score seems to be a more reliable score to describe the learning capabilities of the models, and it is better for comparison with the most recent related studies that make use of it, instead of classification accuracy only.

6.5 From subject-dependent to subject-independent classification

Clearly, subject-dependent classifiers are not an optimal solution for a real-time system and pose a threat to the usability by requiring full training sessions for new users. Besides, training a separated model for each user is not a long-term scalable solution. However, due to the subjective differences, subject-dependent remains the current preferred choice for Emotion-Recognition. Subject-independent strategy is only applicable to set of features that can represent the same emotional patterns for everyone, for example differential asymmetry was reported to be very promising [51], but also assumes that all the datasets used for training are as clean as possible to prevent the contamination of artefacts. In the current study it was not possible to obtain better than default-guessing during intermediate experiments on subject-independent classification. Some conditions are likely to hinder classification:

- Artifacts: Data might be too affected by artifacts, hindering any classification.
- Features: Selected features might not be suitable for all subjects, or they might be suitable only for groups of subjects with similar “brain signature”.
- Labels: Distribution of the classes can be highly uneven, and self-reported labels can be unreliable.
- Electrodes: Two electrodes might not contain sufficient significant data; they can be misplaced or bad conducting.

The enormous performance difference among subjects in subject-dependent classifications suggests that improvements in the first three conditions are not only possible, but also necessary towards the creation of wearable affective BCIs. If the low number of electrodes were the main impediment to classification, it would not have been possible for some participants to score more than 80% classification accuracy. After reducing the variability created by these conditions and once subject-dependent classification will yield consistent performances among subjects, it will be possible to further develop a hybrid paradigm where a subject-independent classifier serves as sub-optimal basis for fast calibration of subject-dependent classifiers. Finally, if these goals are met, full subject-independent classification will be the next step towards affective BCIs.

6.6 Reflections on future research

The analytical study of emotions is heavily impacted by subjective differences, and the current state of art technology and signal processing techniques still struggle to provide the desired performances to build seamless affective BCI systems. Many critical factors starting from the data collection phase to the classification task can hinder the expected outcome. Considering the interest in a follow-up study, the factors that mainly affected the current study are addressed in the paragraphs below.

Selection of the stimuli. The selection of the stimuli is critical for the success for affective experiments, as there is nothing worse than selecting ineffective stimuli. Music is usually a favorable stimulus, since only a few subjects do not react at all to it. Even so, subjective preferences can highly impact the perception of the same song in a population of participants. This study conveniently “recycled” a selection of songs previously used in other experiments and proposed the same playlist to all participants for the sake of inter-subject comparison. This choice led the participants to a constrained experience, that for some resulted in listening to music genres they usually would not listen to. In addition, the annotation experience suffered of great variability, leading some subjects to never report some one or more valence-arousal classes. There are alternative solutions for the stimuli selection, as some related studies experimented [26]. For example, the researcher can ask the subjects to pick music from a selection that covers the emotional spectrum during a pre-listening phase. A better compromise could be a hybrid selection: part of the stimuli selected by the researchers and the other part selected by the participants. Adapting the selection to account for the subjective preferences is very time consuming in the design of an experiment, but it is also what real users of an affective recommending system would expect.

Self-assessment of emotions. As several participants reported, it was not always simple to assess in real-time perceived emotions, with the consequence of unreliable annotations. On one hand, knowing exactly what feelings are being experienced requires considerable self-perception. On the other hand, models for assessing emotions have limitations. As mentioned earlier (see Chapter 6.2), the VA space used in the experiment was simplified to make it more understandable, sacrificing the representation of some emotional states. Using a more complete representation does not automatically improve the participants ability to report emotions, on the contrary it is likely to introduce more confusion. Other studies [24] adopted an even more simplified version that only associates one emotion to each quadrant of the VA space, that maybe partially explains the significantly better performances in classification. Continuous emotion annotations require more effort but are very valuable to capture emotion oscillations and provide classifiers with more realistic emotional distribution. Once the models for the classification of emotion will have reached maturity, continuous annotations will be discarded in favor of discrete reports, that better suit a real product. The solutions for Emotion-Recognition cannot possibly start from a complex representation of emotional states, thus building up from more simplified models might be a better strategy now to support the ongoing development of the field and then later scale up the complexity with more sophisticated tools.

Experimental sessions. Two main observations were made based on the feedback received and the analysis of data. First one, a single session of 75-90 minutes (30-35 of EEG recording) can be fatiguing for most people. Decreased attention and discomfort are not favorable conditions for recording brain activity and are likely to be reflected also on the emotional assessment. Second one, a single session does not ensure that classification performances are consistent for the same subject over time. External factors experienced prior to the session might bias the emotional perception of a subject, for example if there as a breakup with a loved one or if very good news brightened the day. In addition, the oscillatory nature of brain waves is known to generate differences in the EEG signature of the same person over time. Multiple shorter session can prevent fatigue by reducing the overall mental load, and at the same time mitigate the effect of variations over time in the emotional assessment and in the EEG.

Automated lightweight preprocessing. The requirements for a preprocessing pipeline in an online system are not easy to meet. Computational time needs to be in the order of seconds or even better milliseconds; at the same time the quality of data must be ensured to prevent poor classification performances. Ocular artifacts are

the greatest threat in the analysis of emotions because of the topological position near the frontal lobe, where emotions are processed. Using EOG to subtract eye movements from the EEG signal is effective for offline studies but is clearly not suitable for a product. Methods like ICA and PCA are very effective for EEG recordings using standard equipment with many electrodes, but nevertheless computationally expensive. For a final product using wearable devices with limited capabilities, inferring the presence of artifacts by classifying EEG based on the signal qualities [45] can be achieved with low computational effort and enable surgical precision in the cleaning process. ASR has been proven effective in artifact removal for both offline data analysis and online applications [55], but for optimal outcome, i.e. removing artifacts and retaining the significance of the signal, it requires to be tuned using good quality EEG samples. Investigating the best combination of approaches for lightweight processing can be a suitable research question also outside the specific scope of affective BCI.

Features extraction and selection. Selecting the right features has proved to be non-trivial and affected by subject-dependent differences. The extraction process also carries a computation cost, thus just extracting any possible feature and then apply subjective selection criteria is not feasible. The current study used PCA to collapse the features in the minimum number of components that could account for subjective differences and retain at least 95% of the variability of the data. The neuromarkers are a step towards the delineation of an optimal subset of features for emotional assessment but require more investigation both from neuroscience and data science perspectives to determine which combination of differential/rational measurements and EEG frequency bands are more relevant for the study of emotions. The identification of more subject-independent features, like the asymmetry indexes [51], is also essential towards the development of affective BCI systems and can be the main topic of a dedicated research. The possibility to use other physiological signals to build a multi-modal classification system has also been explored, and physiological signals were collected during this research for eventual follow up studies. A related study already assessed the increased performances that can be obtained by decision-level multi-modal fusion [31], an adaptive approach to select by majority voting the optimal uni-modal sources among several physiological measurements (EEG, ECG, GSR) for classification. Any system designed on other physiological data than EEG is, however, outside the strict scope of affective BCI.

Unbalanced datasets. One of the main obstacles in training and evaluating classification models was the uneven distribution of labels. Multiple experimental sessions and subjective stimuli selection can minimize the possibility of having very

unbalanced datasets, as explained in the previous paragraph. However, it would still be possible to hit the same obstacle, regardless the minimization. In the field of machine learning there are standard methods to deal with unbalanced methods. Assigning weights to the classes to penalize the prediction of the majority class is one of such methods, used in the current to improve the discriminative power of SVM classifiers. Up-sampling of the minority class is another method that creates copies of labels and data in the training dataset to reduce the bias of the classifier. However, copying affective data does not seem optimal as it would simply repeat the same emotional pattern and will not likely cover the entire emotional spectrum of the minority class. Some studies investigated the possibility of simulating realistic EEG data [56] from biologically plausible signals. Good affective EEG data from multiple subjects could be collected for the realization of a plausible EEG simulator that can account for individual variability. The data generator could then be calibrated over a small sample of real EEG data from a real subject and be used to counterbalance the distribution of emotional classes, thus improving the classification performances.

Chapter 7

Conclusions and recommendations

The goal of the experiment setup for this research was to answer the main research question: *"What are the accuracy and MCC scores of subject-dependent classification of music-elicited emotional valence and arousal in the EEG signal using SVM and MLP algorithms with Melomind?"* and the two sub-research questions that extended it. Each question is answered separately in the following section and then some recommendations for future work are proposed in the last section.

7.1 Conclusions

"What are the accuracy and MCC scores of subject-dependent classification of music-elicited emotional valence and arousal in the EEG signal using SVM and MLP algorithms with Melomind?".

For arousal classification, SVM scored higher average test accuracy of $61 \pm 9\%$ and higher average MCC score of 0.16 ± 0.20 compared to MLP that scored average test accuracy of $58 \pm 12\%$ and average MCC score of 0.13 ± 0.20. Cross-validated scores are consistent, with SVM scoring higher average CV accuracy of $61 \pm 6\%$ and higher CV MCC of 0.24 ± 0.12, while MLP scored average CV accuracy of $58 \pm 8\%$ and average CV MCC of 0.15 ± 0.16. The highest consistent test accuracy was 84% with 0.20 MCC and 0.08 ± 0.32 CV MCC score for SVM; the highest consistent test accuracy was 88% with 0.78 MCC and 0.28 ± 0.24 CV MCC score for MLP. Similarly, for valence classification, SVM scored higher average test accuracy of $67 \pm 12\%$ and higher average MCC score of 0.13 ± 0.18 compared to MLP that scored average test accuracy of $65 \pm 12\%$ and average MCC score of 0.11 ± 0.18. Cross-validated scores are again consistent, with SVM scoring higher average CV accuracy of $61 \pm 6\%$ and higher CV MCC of 0.26 ± 0.13, while MLP scored average CV accuracy of $56 \pm 9\%$ and average CV MCC of 0.13 ± 0.18. The highest consistent test accuracy was 89% with 0.27 MCC and 0.02 ± 0.20 CV MCC score for SVM; the highest consistent test

accuracy was 77% with 0.48 MCC and 0.35 ± 0.14 CV MCC score for MLP. Overall, SVM models yielded more consistency between test and cross-validated scores: only 4 and 9 SVM models over-fitted for arousal and valence classification respectively, against 6 overfitting and 8 underfitting models in arousal classification and 4 overfitting and 8 underfitting models for valence classification using MLP.

"What are the most relevant selected Power Spectral Density features to perform the Emotion-Recognition using SVM and MLP algorithms with Melomind?".

Intermediate experiments selecting features with SFS suggested that the neuromarkers were more relevant than other raw features of the EEG signal for a number of participants for subject-dependent classification, but not for the entire population, thus the choice of using PCA instead. The final results were obtained after compressing the features using PCA and the contribution of the individual features to the components was not measured. This aggregation of neuromarkers and frequency band-specific spectral features showed encouraging results, but also great variability among subjects that will require more study on the causes and possible solutions to mitigate this effect. No neuromarker or subset of features could be proved to be relevant towards subject-independent classification.

"What is the best classification strategy applicable to the current software and hardware capabilities of Melomind using SVM and MLP algorithms?"

For this research it was possible to obtain subjective appreciable results using a subject-dependent strategy. Subject-independent classification was discarded during intermediate experiments due to inability of the model to perform better than default guessing the majority class. Thus, subject-dependent classification strategy is for the moment the most suitable using Melomind. Further investigation using more sophisticated signal processing techniques and approaches to deal with unbalanced datasets might lead to more consistent results and enable the development of better strategies that can be applied in an online system with minimal training and calibration time.

7.2 Recommendations

Over the course of this research, I faced many design and technical challenges, but the one that took most time to be handled was preparing the data for classification. The preprocessing of the data required continuous reiteration and visualization to understand what was happening when applying certain tools for improving the quality of the EEG signal. Ultimately, tools for artifact cleaning like ASR were discarded because it was not possible to train them in removing the artifacts without affecting

also the ”good” segments of signal. The choice of excluding or slicing data was necessary to proceed with the real classification experiment, but took an heavy toll on the dataset and left an unsatisfactory feeling of incompleteness. The AuPP is the part of the project that required most time to be developed and tuned and yet it is also the most fragile and the first one that should be rewritten and improved almost entirely. The dataset collected for this research has still a lot of potential information to give and the first step in that direction can be obtained by cleaning the artifacts and including as many datasets as possible. Several types of artifacts are often present in the EEG signal [57]:

- External artifacts: noise caused by interference of other electronic devices, like power-line noise, smartphone frequencies, bad electrodes positioning, electrodes movement.
- Muscle artifacts: caused by the movement of facial muscles (tongue movement, swallowing) or neck muscles and often appear in frontal and temporal lobes recording.
- Cardiac artifacts: the heart activity can also be detected by EEG electrodes, especially in the left temporal region.
- Physiological artifacts: these type of artifacts include eyes and eyelids movements and eye blinks and are prominent in the fronto-parietal areas.

There are standard signal processing methods to deal with these artifacts, such as filtering (notch and band-pass) to remove interfering electric frequencies, while PCA and ICA can be used to decompose the signal into components to identify almost any type of artifact, but are only suitable for offline analysis and with a large number of electrodes. Recent researches focusing on automated removal necessarily require to use classification or regression to continuously identify when a segment of signal contains which type of artifact and then apply the right correction. Yeh Sai et al. [58] performed artifacts identification and removal with wavelet-ICA without visual inspection using a pre-trained SVM classifier trained on data contaminated by eye blink artifacts. Their approach allowed the successful removal of target artifacts while retaining most of the EEG source signals of interest. A very recent deep learning approach by Rajabioun et al. [59] was able to successfully classify up to 7 different types of artifacts with 78.12% accuracy score using CNN on a dataset collected to include: blinks, eyes movements, eyebrows movements, head movements, jaw clinches and jaw movements. The effort of cleaning artifacts from EEG signal is a common struggle for researchers in academia or neurotech companies and the development of novel methods that can automatically handle artifacts is becoming essential for the design of interactive BCI applications that fulfill modern

user experience requirements. Considering the aim of this research to evaluate Melomind, a wearable technology with limited hardware, for real-time applications it is clear that this is not only the biggest challenge, but also the one affecting all the other directions for future research and developments. In In conclusion, my primary recommendation for a follow-up study is the collection of contaminated data for the development of cleaning tools for automatic preprocessing of the EEG signal that could be then applied to the dataset collected for this research and all future datasets that will be collected using Melomind or similar wearabled devices with dry electrodes. Besides artifacts cleaning, this research on affective BCI offers various insights and ideas for further exploring the field of Affective Computing as discussed in Chapter 6.6. In the time assigned to the project it was only possible to scratch the surface, but hopefully an interesting and actual overview of the state-of-art methods, devices, and strategies was provided to the readers. Researchers and designers interested in building the affective technologies of tomorrow are welcomed to take part in this challenging and exciting world, and share their ideas, insights and solutions with the passionate community that is growing around Brain-Computer Interfaces, that can only exist thanks to the shared efforts, ideas and cooperation that I had the pleasure of discovering during my last two years of experience as scholar of Human-Computer Interaction and Design.

Bibliography

[1] (). “Gartner’s 2016 hype cycle for emerging technologies identifies three key trends that organizations must track to gain competitive advantage,” Gartner, [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2016-08-16-gartners-2016-hype-cycle-for-emerging-technologies-identifies-three-key-trends-that-organizations-must-track-to-gain-competitive-advantage (visited on 11/29/2021).

[2] A. Parfenov. (). “BrainFlow 4.6.0,” [Online]. Available: https://brainflow.org/2021-08-17-enophone/ (visited on 11/29/2021).

[3] ——, (). “OpenBCI galea with BrainFlow,” [Online]. Available: https://brainflow.org/2021-01-26-galea-brainflow/ (visited on 11/29/2021).

[4] N. Statt. (Sep. 23, 2019). “Facebook acquires neural interface startup CTRL-labs for its mind-reading wristband,” The Verge, [Online]. Available: https://www.theverge.com/2019/9/23/20881032/facebook-ctrl-labs-acquisition-neural-interface-armband-ar-vr-deal (visited on 11/29/2021).

[5] H.-Y. Chang, S.-C. Huang, and J.-H. Wu, “A personalized music recommendation system based on electroencephalography feedback,” *Multimedia Tools and Applications*, vol. 76, no. 19, pp. 19 523–19 542, Oct. 1, 2017, ISSN: 1573-7721. DOI: 10.1007/s11042-015-3202-4. [Online]. Available: https://doi.org/10.1007/s11042-015-3202-4 (visited on 07/12/2021).

[6] A. Abdul, J. Chen, H.-Y. Liao, and S.-H. Chang, “An emotion-aware personalized music recommendation system using a convolutional neural networks approach,” *Applied Sciences*, vol. 8, p. 1103, Jul. 8, 2018. DOI: 10.3390/app8071103.

[7] R. W. Picard, “MIT media laboratory; perceptual computing; 20 ames st., cambridge, MA 02139 picard@media.mit.edu, http://www.media.mit.edu/˜picard/,” p. 16,

[8] N. Gertz, "Autonomy online: Jacques ellul and the facebook emotional manipulation study," *Research Ethics*, vol. 12, no. 1, pp. 55–61, Jan. 1, 2016, ISSN: 1747-0161. DOI: 10.1177/1747016115579534. [Online]. Available: https://doi.org/10.1177/1747016115579534 (visited on 07/14/2021).

[9] N. Tromp, P. Hekkert, and P.-P. Verbeek, "Design for socially responsible behavior: A classification of influence based on intended user experience," *Design Issues*, vol. 27, no. 3, pp. 3–19, Jul. 1, 2011, ISSN: 0747-9360. DOI: 10.1162/DESI_a_00087. [Online]. Available: https://doi.org/10.1162/DESI_a_00087 (visited on 07/14/2021).

[10] R. W. Picard, "Affective computing: Challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 55–64, Jul. 2003, ISSN: 10715819. DOI: 10.1016/S1071-5819(03)00052-1. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1071581903000521 (visited on 10/14/2021).

[11] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors (Basel, Switzerland)*, vol. 12, no. 2, pp. 1211–1279, Jan. 31, 2012, ISSN: 1424-8220. DOI: 10.3390/s120201211. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3304110/ (visited on 09/17/2021).

[12] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, Dec. 1, 1980. DOI: 10.1037/h0077714.

[13] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1, 1994, ISSN: 0005-7916. DOI: 10.1016/0005-7916(94)90063-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005791694900639 (visited on 08/08/2021).

[14] D. Watson, L. Anna, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," p. 8,

[15] L. A. Schmidt and L. J. Trainor, "Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions," *Cognition and Emotion*, vol. 15, no. 4, pp. 487–500, Jul. 1, 2001, Publisher: Routledge _eprint: https://doi.org/10.1080/02699930126048, ISSN: 0269-9931. DOI: 10.1080/02699930126048. [Online]. Available: https://doi.org/10.1080/02699930126048 (visited on 03/07/2021).

[16] R. J. Davidson, "Anterior cerebral asymmetry and the nature of emotion," *Brain and Cognition*, vol. 20, no. 1, pp. 125–151, Sep. 1992, ISSN: 0278-2626. DOI: 10.1016/0278-2626(92)90065-t.

[17] G. Dawson, "Frontal electroencephalographic correlates of individual differences in emotion expression in infants: A brain systems perspective on emotion," *Monographs of the Society for Research in Child Development*, vol. 59, no. 2, pp. 135–151, 1994, ISSN: 0037-976X.

[18] W. Heller, "Neuropsychological mechanisms of individual differences in emotion, personality, and arousal," *Neuropsychology*, vol. 7, no. 4, pp. 476–489, Oct. 1993, ISSN: 0894-4105. DOI: 10.1037/0894-4105.7.4.476. [Online]. Available: http://www.scopus.com/inward/record.url?scp=0001052432&partnerID=8YFLogxK (visited on 09/17/2021).

[19] L. Orgo, M. Bachmann, J. Lass, and H. Hinrikus, "Effect of negative and positive emotions on EEG spectral asymmetry," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2015, pp. 8107–8110, Aug. 2015, ISSN: 2694-0604. DOI: 10.1109/EMBC.2015.7320275.

[20] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch, "Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music," p. 12,

[21] B. Reuderink, C. Mühl, and M. Poel, "Valence, arousal and dominance in the EEG during game play," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 6, pp. 45–62, Dec. 1, 2013. DOI: 10.1504/IJAACS.2013.050691.

[22] G. Zhao, Y. Zhang, and Y. Ge, "Frontal EEG asymmetry and middle line power difference in discrete emotions," *Frontiers in Behavioral Neuroscience*, vol. 12, p. 225, 2018, ISSN: 1662-5153. DOI: 10.3389/fnbeh.2018.00225. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbeh.2018.00225 (visited on 11/23/2021).

[23] Y. Lin, C. Wang, T. Wu, S. Jeng, and J. Chen, "EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ISSN: 2379-190X, Apr. 2009, pp. 489–492. DOI: 10.1109/ICASSP.2009.4959627.

[24] Y. Lin, C. Wang, T. Jung, T. Wu, S. Jeng, J. Duann, and J. Chen, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010, Conference Name: IEEE Transactions on Biomedical Engineering, ISSN: 1558-2531. DOI: 10.1109/TBME.2010.2048568.

[25] S. Wu, X. Xu, L. Shu, and B. Hu, "Estimation of valence of emotion using two frontal EEG channels," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, pp. 1127–1130. DOI: 10.1109/BIBM.2017.8217815.

[26] N. Thammasan, K. Moriyama, K.-i. Fukui, and M. Numao, "Continuous music-emotion recognition based on electroencephalogram," *IEICE Transactions on Information and Systems*, vol. E99.D, pp. 1234–1241, Apr. 1, 2016. DOI: 10.1587/transinf.2015EDP7251.

[27] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012, Conference Name: IEEE Transactions on Affective Computing, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.15.

[28] T. Eerola and J. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models, and stimuli," *Music Perception*, Feb. 1, 2013. DOI: 10.1525/mp.2012.30.3.307.

[29] M. Soleymani, J. J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Amsterdam, Netherlands: IEEE, Sep. 2009, pp. 1–7, ISBN: 978-1-4244-4800-5. DOI: 10.1109/ACII.2009.5349563. [Online]. Available: http://ieeexplore.ieee.org/document/5349563/ (visited on 12/02/2021).

[30] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, Jun. 1, 1988, ISSN: 0167-2789. DOI: 10.1016/0167-2789(88)90081-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167278988900814 (visited on 03/17/2021).

[31] N. Thammasan, J. L. Hagad, K.-i. Fukui, and M. Numao, "Multimodal stability-sensitive emotion recognition based on brainwave and physiological signals," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Oct. 2017, pp. 44–49. DOI: 10.1109/ACIIW.2017.8272584.

[32] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The PREP pipeline: Standardized preprocessing for large-scale EEG analysis," *Frontiers in Neuroinformatics*, vol. 9, p. 16, 2015, ISSN: 1662-5196. DOI: 10.3389/fninf.2015.00016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2015.00016 (visited on 11/12/2021).

[33] P. Keelawat, N. Thammasan, M. Numao, and B. Kijsirikul, "A comparative study of window size and channel arrangement on EEG-emotion recognition using deep CNN," *Sensors*, vol. 21, no. 5, p. 1678, Jan. 2021, Number: 5 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/s21051678. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1678 (visited on 03/11/2021).

[34] ——, "Spatiotemporal emotion recognition using deep CNN based on EEG during music listening," *arXiv:1910.09719 [cs, eess, stat]*, Oct. 21, 2019. arXiv: 1910.09719. [Online]. Available: http://arxiv.org/abs/1910.09719 (visited on 07/12/2021).

[35] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan. 2012, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.25. [Online]. Available: http://ieeexplore.ieee.org/document/5975141/ (visited on 01/27/2022).

[36] M. Soleymani, A. Aljanaki, and Y.-H. Yang, "DEAM: MediaEval database for emotional analysis in music," p. 3,

[37] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schr{\textbackslash}"oder, *'FEELTRACE': An instrument for recording perceived emotion in real time*. Jan. 1, 2000, Journal Abbreviation: Proceedings of the ISCA Workshop on Speech and Emotion Publication Title: Proceedings of the ISCA Workshop on Speech and Emotion.

[38] S. Sangnark, P. Autthasan, P. Ponglertnapakorn, P. Chalekarn, T. Sudhawiyangkul, M. Trakulruangroj, S. Songsermsawad, R. Assabumrungrat, S. Amplod, K. Ounjai, and T. Wilaiprasitporn, "Revealing preference in popular music through familiarity and brain response," *IEEE Sensors Journal*, pp. 1–1, 2021, ISSN: 1530-437X, 1558-1748, 2379-9153. DOI: 10.1109/JSEN.2021.3073040. arXiv: 2102.00159. [Online]. Available: http://arxiv.org/abs/2102.00159 (visited on 05/17/2021).

[39] M. Ward, J. Goodman, and J. Irwin, "The same old song: The power of familiarity in music choice," *Marketing Letters*, vol. 25, May 1, 2013. DOI: 10.1007/s11002-013-9238-1.

[40] V. Salimpoor, M. Benovoy, K. Larcher, A. Dagher, and R. Zatorre, "Anatomically distinct dopamine release during anticipation and experience of peak emotion to music," *Nature neuroscience*, vol. 14, pp. 257–62, Feb. 1, 2011. DOI: 10.1038/nn.2726.

[41] L. Fang, J. Shang, and N. Chen, "Perception of western musical modes: A chinese study," *Frontiers in Psychology*, vol. 8, 2017, Publisher: Frontiers, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01905. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01905/full (visited on 03/01/2021).

[42] R. J. Barry, A. R. Clarke, S. J. Johnstone, C. A. Magee, and J. A. Rushby, "EEG differences between eyes-closed and eyes-open resting conditions," *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, vol. 118, no. 12, pp. 2765–2773, Dec. 2007, ISSN: 1388-2457. DOI: 10.1016/j.clinph.2007.07.028.

[43] Y.-H. Chang, Y.-Y. Lee, K.-C. Liang, I.-P. Chen, C.-G. Tsai, and S. Hsieh, "Experiencing affective music in eyes-closed and eyes-open states: An electroencephalography study," *Frontiers in Psychology*, vol. 6, p. 1160, 2015, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.01160. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2015.01160 (visited on 10/06/2021).

[44] D. Hagemann and E. Naumann, "The effects of ocular artifacts on (lateralized) broadband power in the EEG," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 112, pp. 215–31, Mar. 1, 2001. DOI: 10.1016/S1388-2457(00)00541-1.

[45] F. Grosselin, X. Navarro-Sune, A. Vozzi, K. Pandremmenou, F. De Vico Fallani, Y. Attal, and M. Chavez, "Quality assessment of single-channel EEG for wearable devices," *Sensors (Basel, Switzerland)*, vol. 19, no. 3, E601, Jan. 31, 2019, ISSN: 1424-8220. DOI: 10.3390/s19030601.

[46] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*. Cambridge, MA, USA: MIT Press, Jan. 17, 2014, 600 pp., ISBN: 978-0-262-01987-3.

[47] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica Et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, Oct. 20, 1975, ISSN: 0006-3002. DOI: 10.1016/0005-2795(75)90109-9.

[48] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, p. 6, Jan. 2, 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7.

[49] D. Martin. (). "Stacked turtles," Stacked Turtles, [Online]. Available: https://kiwidamien.github.io (visited on 12/02/2021).

[50] (). "Please support customize loss function in SGDClassifier/regressor · issue #1701 · scikit-learn/scikit-learn," GitHub, [Online]. Available: https://github.com/scikit-learn/scikit-learn/issues/1701 (visited on 12/02/2021).

[51] Y.-P. Lin, T.-P. Jung, and J. Onton, "Toward affective brain-computer interface: Fundamentals and analysis of EEG-based emotion classification," in, Jan. 2, 2015, pp. 315–341, ISBN: 978-1-118-13066-7. DOI: 10.1002/9781118910566.ch13.

[52] K. Avramidis, A. Zlatintsi, C. Garoufis, and P. Maragos, "Multiscale fractal analysis on EEG signals for music-induced emotion recognition," *arXiv:2010.16310 [cs]*, Mar. 2, 2021. arXiv: 2010.16310. [Online]. Available: http://arxiv.org/abs/2010.16310 (visited on 03/17/2021).

[53] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards zero training for brain-computer interfacing," *PloS One*, vol. 3, no. 8, e2967, Aug. 13, 2008, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0002967.

[54] J. H. Jeong, D.-J. Kim, and H. Kim, "Hybrid zero-training BCI based on convolutional neural network for lower-limb motor-imagery," in *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*, ISSN: 2572-7672, Feb. 2021, pp. 1–4. DOI: 10.1109/BCI51272.2021.9385316.

[55] C.-Y. Chang, S.-H. Hsu, L. Pion-Tonachini, and T.-P. Jung, "Evaluation of artifact subspace reconstruction for automatic EEG artifact removal," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2018, pp. 1242–1245, Jul. 2018, ISSN: 2694-0604. DOI: 10.1109/EMBC.2018.8512547.

[56] E. Barzegaran, S. Bosse, P. J. Kohler, and A. M. Norcia, "EEGSourceSim: A framework for realistic simulation of EEG scalp data using MRI-based forward models and biologically plausible signals and noise," *Journal of Neuroscience Methods*, vol. 328, p. 108 377, Dec. 1, 2019, ISSN: 1872-678X. DOI: 10.1016/j.jneumeth.2019.108377.

[57] A. Tandle, "Classification of artefacts in EEG signal recordings and overview of removing techniques," *International Journal of Computer Applications*, p. 5,

[58] C. Y. Sai, N. Mokhtar, H. Arof, P. Cumming, and M. Iwahashi, "Automated classification and removal of EEG artifacts with SVM and wavelet-ICA," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 664–670, May 2018, Conference Name: IEEE Journal of Biomedical and Health Informatics, ISSN: 2168-2208. DOI: 10.1109/JBHI.2017.2723420.

[59] R. Rajabioun, A. Ö. Akyürek, and E. A. Sezer, “Deep learning approach for EEG artifact identification and classification,” in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, ISSN: 2521-1641, Sep. 2021, pp. 320–325. DOI: 10.1109/UBMK52708.2021.9558979.

Appendix A

Appendix

The appendix roughly follows the structure of the main thesis. Some of the tables and plots from the intermediate experiments section (Appendix A.3) were obtained during an exploratory phase and thus are only approximated and do not contain subject-wise information. However they should give an overview of the logical reasoning that led to the final experiment, for which more individual information is reported in Appendix A.4. Individual ROC curves, labels distribution and confusion matrices are only reported for specific examples regarding the unbalance of some datasets (see Appendix A.3.3)

A.1 Pilot study

A pilot study was run internally with myBrainTechnologies employee to design the Experimental Annotator App. The choice of using mouse as input method was evaluated through A|B testing with two training sessions using a demo of the app. The table reports the average usability score for each training session.

A.1.1 Usability scores for ExperimentalAnnotator app demo

The usability scores were collected with forms asking to rate how simple it was on a scale from 0 to 5 to perform the following annotation tasks on the GUI representing VA space using either the mouse or a joystick:

- Selecting the desired quadrant
- Selecting the perceived emotion
- Selecting the correct intensity of the perceived emotion

The scores were collected using the training sessions developed for the experiment, so the first training gave the participants some background on the experiment

Figure A.1: Usability scores for training in condition A (joystick)

Figure A.2: Usability scores for training in condition A (joystick)

and the task while the second training asked the participants to annotate their emotions in real-time while listening to 4 music excerpts of 30 seconds each. As it can be seen from table A.1, on average joystick users improved their ability to select what they wanted while mouse users were consistent. This was not completely unexpected since the mouse is very likely to be used by most people working with computers, while joystick is more of a niche input for people who play games or use simulators.

The lower average ratings obtained by the joystick and the higher likelihood of skill gaps between participants was enough to drop the joystick as input method, which was only really interesting for the possibility of annotating emotions in eyes-closed condition.

Table A.1: Average usability scores for the two groups.

Input Type	Session	Selecting Quadrant (AVG)	Selecting Emotion (AVG)	Selecting Intensity (AVG)
A Joystick	1	3,75	3,5	3,5
	2	4,25	3,75	3,75
B Mouse	1	4,4	4,2	4
	2	4,4	4	4

A.2 Methods

A.2.1 Participants infographic

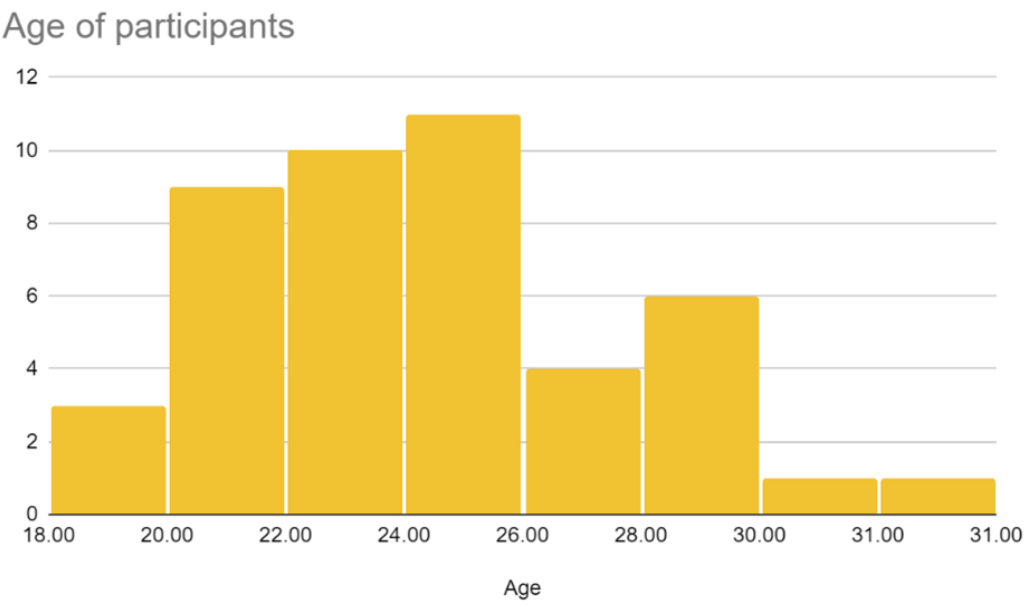

Figure A.3: Age distribution of the population that participated in the experiment

A.2.2 Emotion-Music playlist

The playlist is a subset of the stimuli proposed by Koelstra et al. [27] and selected using the emotional tagging available on last.fm. The arousal/valence intensity (Low,

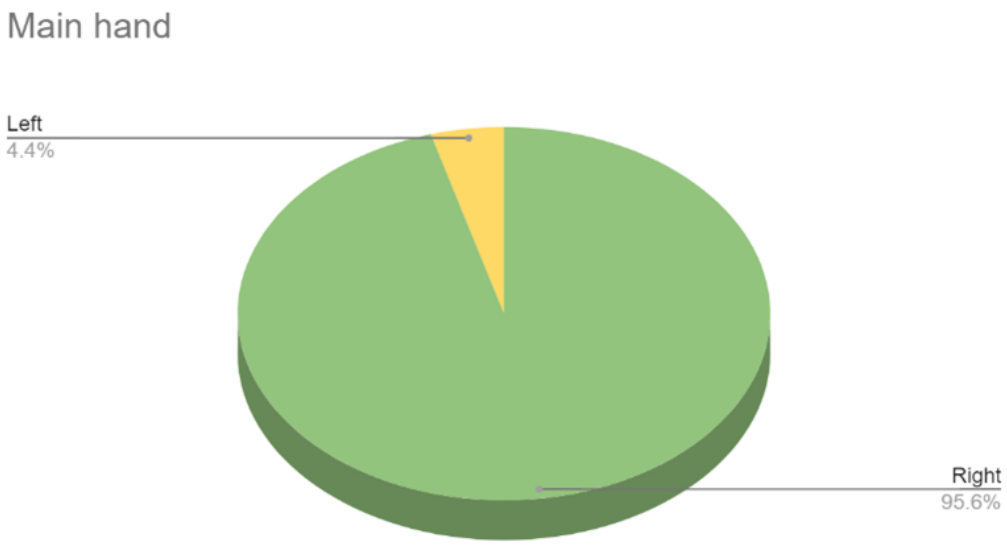

Figure A.4: Percentages of right-handed and left-handed over the population that participated in the experiment

Medium, High) are a transposition of the scores assigned by users who participated in the web assessment of the aforementioned study, but ultimately have been simplified as a binary representation, low or high.
High Arousal / High Valence (HAHV)

- (High arousal, medium positive valence) Excitement- Weapon of choice by Fatboy Slim
- (Medium positive arousal, high valence) Happiness - Love Today by Mika

Low Arousal / High Valence (LAHV)

- (Medium negative Arousal, high valence) Satisfaction - Nasty Naughty Boy, by Christina Aguilera
- (Low arousal, medium positive valence) Relaxation – Amber by 311

Low Arousal / Low Valence (LALV)

- (Medium negative arousal, High negative valence) Depression - Last Flowers by Radiohead
- (Low arousal, Medium negative valence) Sadness – Hurt by Johnny Cash

High Arousal / Low Valence (HALV)

- (Medium positive arousal, High negative valence) Anger – Threshold by Slayer
- (High positive arousal, Medium negative valence) Anxiety - Trapped Under Ice by Believe

Other songs were used in the training sessions, but no labelling is provided as their only purpose was to teach the users how to use the EA app GUI.

- The white stripes - Seven Nation Army
- Gary Jules - Mad World
- Oren Lavie - Her morning elegance
- The Cranberries - Zombie
- Maneskin - I wanna be your slave
- Rage against the machine - Killing in the name of

A.3 Intermediate experiments

During the development of the classification pipeline, several intermediate experiments were run to explore the data, the methodologies, and the different classifiers. The dataset of each subject was explored singularly, but only some examples are reported below

A.3.1 Subject-dependent experiment with Sequential Features Selection

Sequential Features Selection is a greedy features selection method used to reduce the amount of features by preferring those that improve the performances of a classifier while reducing the variance. Before computing SFS, GridSearch was used on the entire dataset including all participants to select the best subject-independent hyperparameters (tables A.2 and A.3).

Table A.2: Manually tuned hyper-paramters for MLP.

Classifier	Activation Function	Alpha	Hidden Layer Sizes	Learning Rate	Solver
Arousal	Tanh	0.00001	4, 4	Adaptive	lbfgs
Valence	Tanh	0.01	10, 2, 5	Adaptive	lbfgs
Valence-Arousal	Tanh	0.00001	10, 2, 5	Adaptive	adam

Table A.3: Manually tuned hyper-paramters for SVM.

Classifier	Kernel	C	Class Weights	Gamma
Arousal	RBF	0.0001	No	Scale : 1 / (n_features * X.var())
Valence	RBF	0.0001	No	Scale : 1 / (n_features * X.var())
Valence-Arousal	RBF	0.0001	No	Scale : 1 / (n_features * X.var())

Then, forward SFS was used to select 5 features for each participant and perform the first classification experiment with subject-dependent strategy for each type of classifier (SVM or MLP) and each condition (EO, EC or EO&EC). Cross-validated accuracy scores were averaged and reproted in A.4.

Table A.4: Average cross-validated accuracy for each classifier and listening condition using SFS.

Classifier	Cond.	Avg CV Accuracy	Max/Min CV Accuracy	Std	Avg CV Accuracy	Max/Min CV Accuracy	Std
Arousal	EO	0.57	0.85 / 0.44	0.05	0.69	0.84 / 0.51	0.18
Arousal	EC	0.59	0.89 / 0.51	0.04	0.70	0.94 / 0.46	0.17
Arousal	EO&EC	0.58	0.86 / 0.51	0.03	0.66	0.79 / 0.50	0.16
Valence	EO	0.65	0.93 / 0.51	0.05	0.67	0.82 / 0.47	0.20
Valence	EC	0.65	0.91 / 0.51	0.04	0.66	0.81 / 0.51	0.19
Valence	EO&EC	0.65	0.92 / 0.51	0.02	0.63	0.86 / 0.41	0.16
VA	EO	0.41	0.78 / 0.27	0.04	0.47	0.76 / 0.28	0.16
VA	EC	0.42	0.81 / 0.31	0.05	0.47	0.80 / 0.27	0.16
VA	EO&EC	0.41	0.80 / 0.28	0.03	0.42	0.80 / 0.25	0.12
		SVM			MLP		

In addition, the most selected features for each classifier and each condition were grouped (see table A.5) for later use in two more experiments under the name of TOP5 features.

Table A.5: Most frequently selected features using SFS, renamed TOP5 features.

Cond.	T.	Classifier	Top 5 selected features
EO&EC	5s	Arousal	SASI_IDX_F4, RSD_THETA_F4, ALPHA_NORM_F3, SKEWNESS_BETA_F3, RSD_ALPHA_F4
EO&EC	5s	Valence	AW_IDX, FMT_IDX, SASI_IDX_F4, SKEWNESS_BETA_F4, SASI_IDX_F3
EO&EC	5s	VA	SASI_IDX_F4, FMT_IDX, STD_ALPHA_F3, SASI_IDX_F3, THETA_NORM_F4
EO	5s	Arousal	KURTOSIS_BETA_F4, AW_IDX, FMT_IDX, RSD_THETA_F3, STD_ALPHA_F4
EO	5s	Valence	AW_IDX, BETA_NORM_F4, SASI_IDX_F3, FMT_IDX, SASI_IDX_F4
EO	5s	VA	SASI_IDX_F4, FMT_IDX, AW_IDX, RSD_THETA_F3, KURTOSIS_BETA_F4
EC	5s	Arousal	AW_IDX, RATIO_BETA_F4, SASI_IDX_F3, THETA_NORM_F3, ALPHA_NORM_F4
EC	5s	Valence	AW_IDX, SASI_IDX_F4, FMT_IDX, SKEWNESS_THETA_F4, SASI_IDX_F3
EC	5s	VA	FMT_IDX, SASI_IDX_F4, AW_IDX, RSD_THETA_F4, THETA_NORM_F4

A.3.2 Subject-dependent experiment with TOP5 features

For this experiment, the average TOP5 features selected using SFS were applied to all the generated models, for each classification task and in each condition similarly to the previous experiment. Average CV accuracy scores were lower, suggesting that applying the same features using this generalization criterion was not optimal for subject-dependent strategy. For this reason, and to optimize computational time, the dimensionality of the features array was later reduced using PCA instead in the final experiments.

Table A.6: Average cross-validated accuracy for each classifier and listening condition using TOP5 features.

Classifier	Cond.	Avg CV Accuracy	Max/Min CV Accuracy	Std	Avg CV Accuracy	Max/Min CV Accuracy	Std
Arousal	EO	0.57	0.85 / 0.44	0.05	0.68	0.80 / 0.54	0.17
Arousal	EC	0.60	0.89 / 0.51	0.04	0.60	0.78 / 0.49	0.16
Arousal	EO&EC	0.58	0.86 / 0.51	0.03	0.61	0.81 / 0.38	0.15
Valence	EO	0.65	0.93 / 0.51	0.05	0.62	0.84 / 0.45	0.15
Valence	EC	0.65	0.91 / 0.51	0.04	0.62	0.89 / 0.43	0.15
Valence	EO&EC	0.65	0.92 / 0.51	0.02	0.61	0.89 / 0.51	0.12
VA	EO	0.41	0.78 / 0.28	0.04	0.41	0.78 / 0.22	0.13
VA	EC	0.42	0.81 / 0.26	0.05	0.43	0.78 / 0.29	0.14
VA	EO&EC	0.41	0.80 / 0.28	0.03	0.41	0.78 / 0.23	0.06
		SVM			MLP		

A.3.3 Example of unbalanced dataset for arousal classification

A more insightful study was conducted on participants with suspiciously high Test accuracy and low CV accuracy to identify the causes. Here is reported an example from a participant with unbalanced arousal labels. First, the labelling behaviour during experimental trials is reported in figure A.5, then the total distribution of VA classes is reported in figure A.6.

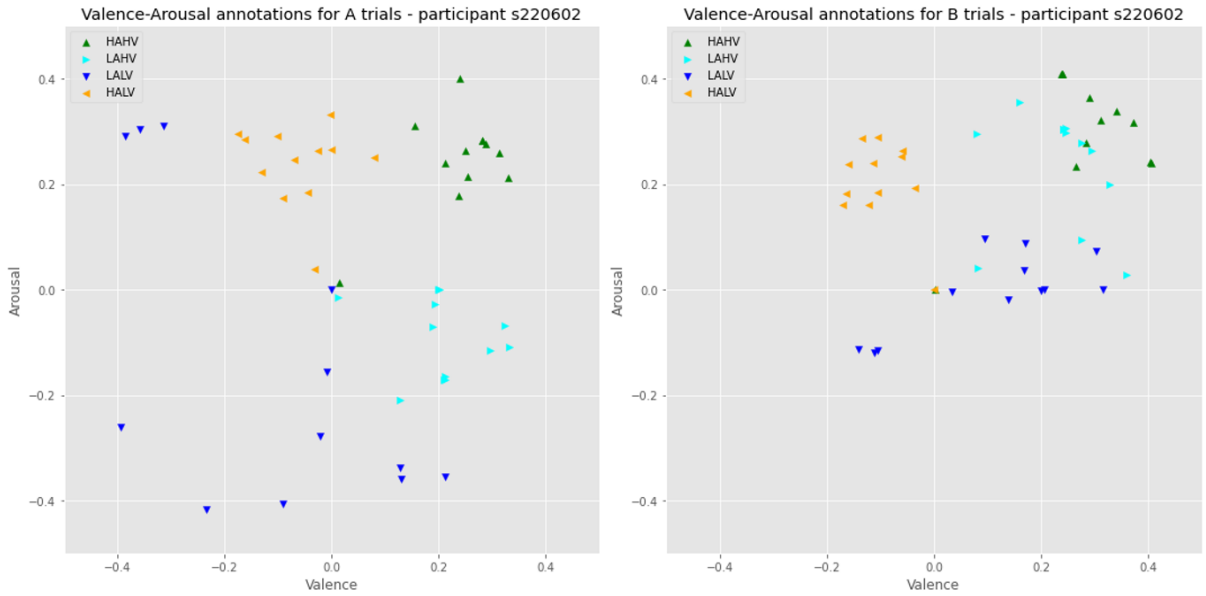

Figure A.5: Summary of annotation session for participant s220602. The labels are color coded according to the pre-labelled VA class of each song.

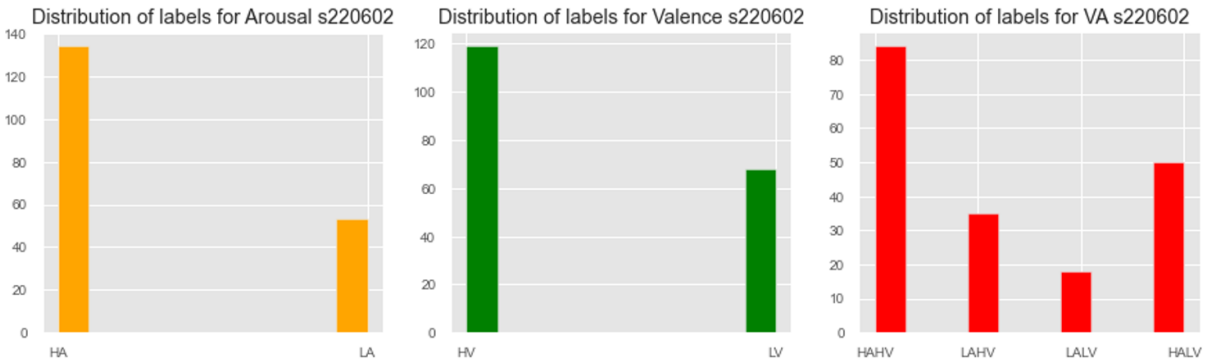

Figure A.6: Labels distribution of participant s220602. The LA* classes summed up are 1/4 of the entire dataset.

For both SVM and MLP classifiers cross-validated ROC curves were computed (see fig. A.7 and fig. A.8), showing the high variance between each split (in grey) against the mean accuracy (in blue), that is usually a symptom of overfitting.

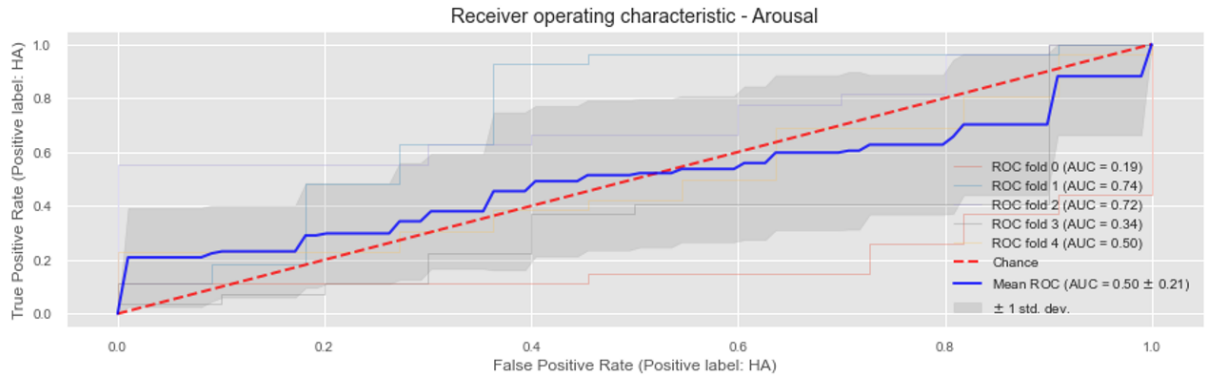

Figure A.7: Cross-validated ROC curve for participant s220602 for arousal classification using SVM.

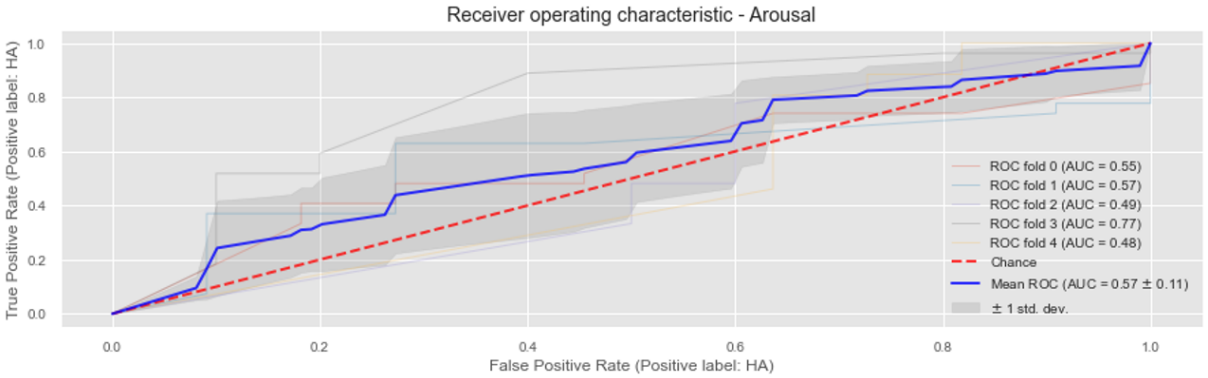

Figure A.8: Cross-validated ROC curve for participant s220602 for arousal classification using MLP.

Finally, PCA was computed to observe the distribution of labels across the first and the second principal components obtained by the features, and confusion matrices for both SVM and MLP revealed the difficulty in predicting the minority class due to the data not being linearly separable (see fig. A.9)

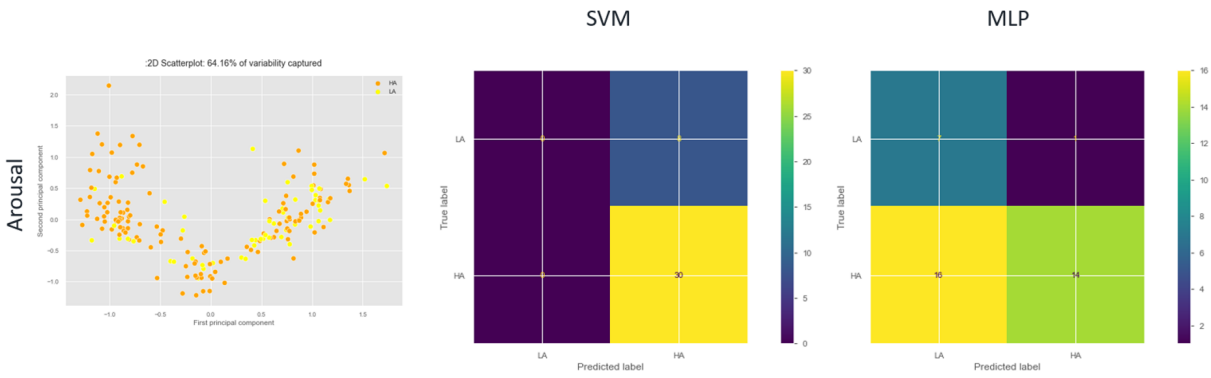

Figure A.9: PCA (left) and confusion matrices (right) of s220602 participant with unbalanced labels for emotional arousal classification.

A.3.4 Example of unbalanced dataset for valence classification

A more insightful study was conducted on participants with suspiciously high Test accuracy and low CV accuracy to identify the causes. Here is reported an example from a participant with unbalanced valence labels. First, the labelling behaviour during experimental trials is reported in figure A.10, then the total distribution of VA classes is reported in figure A.11.

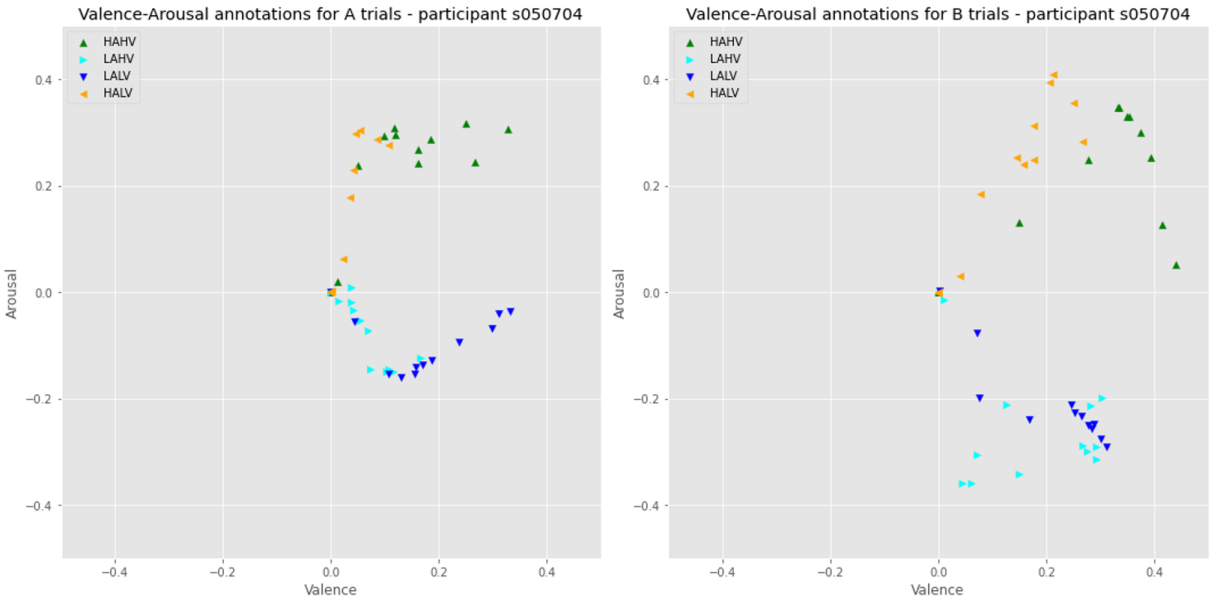

Figure A.10: Summary of annotation session for participant s050704. The labels are color coded according to the pre-labelled VA class of each song.

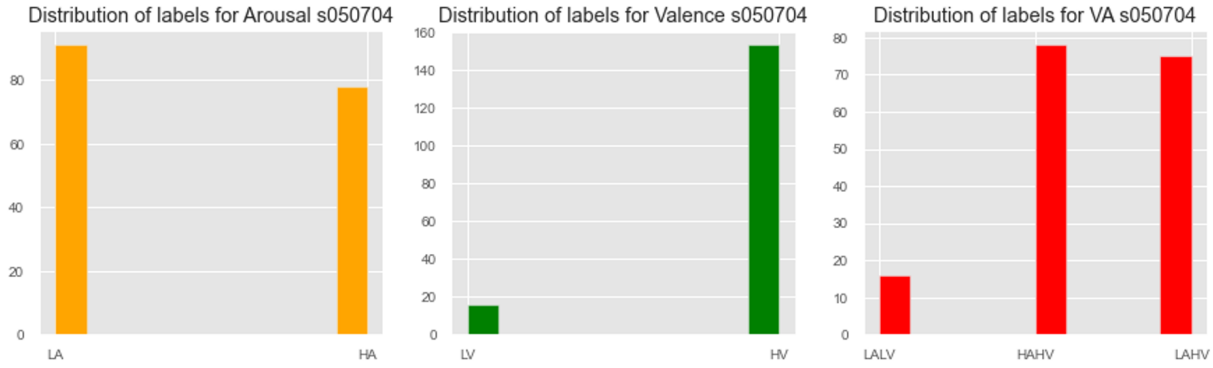

Figure A.11: Labels distribution of participant s050704. The class HALV is completely missing.

For both SVM and MLP classifiers cross-validated ROC curves were computed (see fig. A.12 and fig. A.13), showing the high variance between each split (in grey) against the mean accuracy (in blue), that is usually a symptom of overfitting.

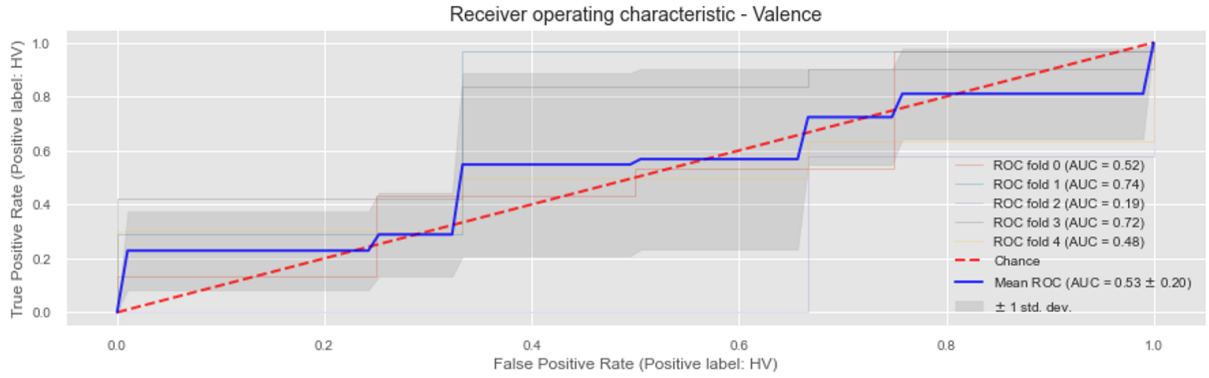

Figure A.12: Cross-validated ROC curve for participant s050704 for valence classification using SVM.

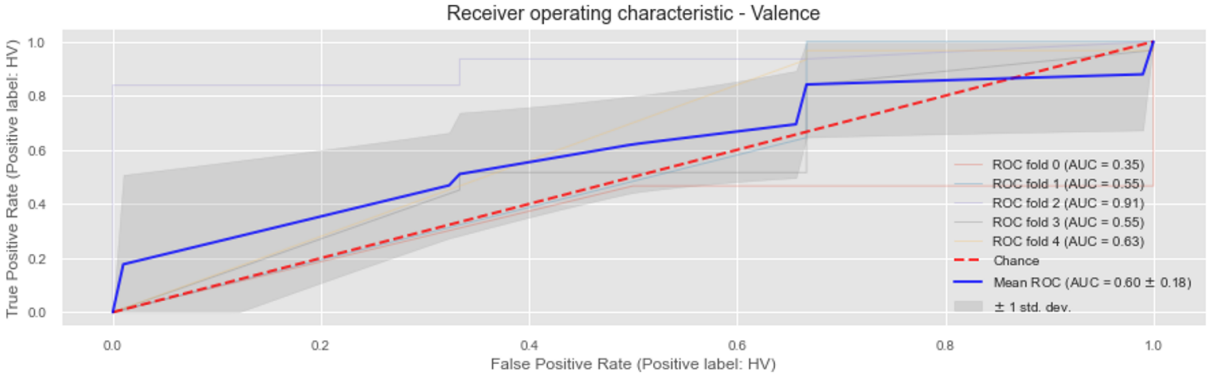

Figure A.13: Cross-validated ROC curve for participant s050704 for valence classification using MLP.

Finally, PCA was computed to observe the distribution of labels across the first and the second principal components obtained by the features, and confusion matrices for both SVM and MLP revealed the difficulty in predicting the minority class due to the data not being linearly separable (see fig. A.14)

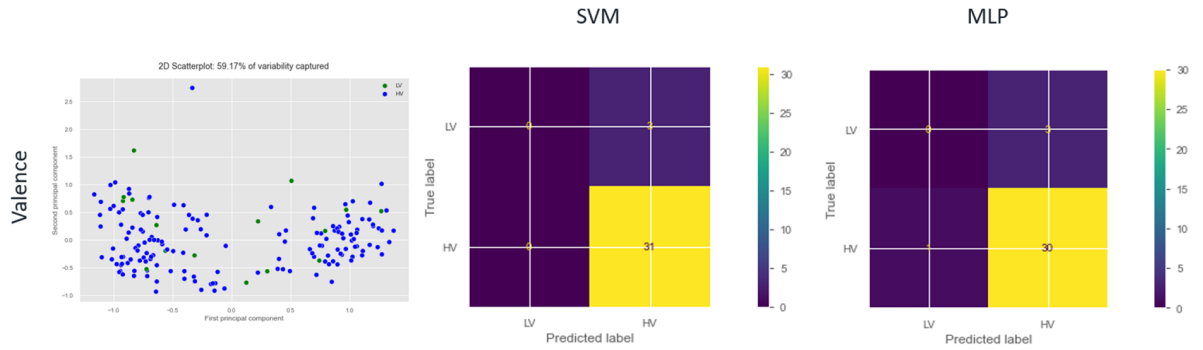

Figure A.14: PCA (left) and confusion matrices (right) of s050704 participant with unbalanced labels for emotional valence classification.

A.3.5 Subject-independent experiment with Top5 features

Since it was impossible to apply SFS to the entire dataset for subject-independent classification without days of computational time, for this experiment TOP5 features previously identified were used. The poor average results (see table A.7) obtained pushed to focus this research on subject-dependent strategy. Another subject-independent experiment was conducted using PCA on a sub-group of participants with balanced datasets, but the results were far below expectations and have not been saved for reporting.

Table A.7: Average cross-validated accuracy for each classifier and listening condition using TOP5 features.

Classifier	Cond.	Avg CV Acc	Max/Min Avg CV Acc	Avg Std	Avg Test Acc.	Avg F1	Avg CV ACC	Max/Min Avg CV Acc	Avg Std	Avg Test Acc.	Avg F1
Arousal	EO	0.56	0.57 / 0.55	0.005	0.56	0.40	0.55	0.56 / 0.53	0.03	0.55	0.48
Arousal	EC	0.58	0.59 / 0.57	0.004	0.59	0.44	0.57	0.58/0.55	0.03	0.55	0.48
Arousal	EO&EC	0.57	0.58 / 0.56	0.002	0.57	0.42	0.57	0.58 / 0.56	0.01	0.50	0.43
Valence	EO	0.66	0.67 / 0.64	0.005	0.66	0.53	0.66	0.66 / 0.63	0.02	0.60	0.57
Valence	EC	0.64	0.65 / 0.63	0.004	0.65	0.52	0.62	0.64 / 0.60	0.03	0.57	0.54
Valence	EO&EC	0.65	0.66 / 0.64	0.002	0.65	0.52	0.64	0.66 / 0.62	0.02	0.57	0.53
VA	EO	0.37	0.38 / 0.35	0.005	0.37	0.22	0.37	0.38 / 0.35	0.008	0.35	0.24
VA	EC	0.37	0.38 / 0.36	0.005	0.38	0.23	0.37	0.38 / 0.36	0.007	0.34	0.23
VA	EO&EC	0.37	0.38 / 0.35	0.003	0.37	0.22	0.37	0.38 / 0.35	0.001	0.37	0.25
		SVM					MLP				

A.3.6 Subject-dependent experiment with ”Max Accuracy” scoring strategy

Tables A.8 and A.9 report the results for arousal and valence subject-dependent classification respectively with a scoring strategy that optimizes hyperparameters to maximize ”accuracy” using GridSearch. While average accuracy and CV accuracy scores are higher, the number of overfitting and underfitting models that were generated, especially for SVM classifier is also higher. This approach for GridSearch optimization was discarded in the final experiment in favor of maximising MCC scores instead.

Table A.8: Arousal classification results using ”accuracy” as scoring parameter for GridSearch. Learning models are highlighted in blue, over-fitted and under-fitted models are highlighted in yellow and orange, respectively.

Arousal	SVM							MLP							EXTRA		
Participant	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Chance Level	Avg Famili	Avg Liking
s010701	0.50	-0.07	0.66	0.08	0.37	0.09	0.16	0.61	0.20	0.63	0.05	0.27	0.11	0.16	0.57	2.49	2.51
s010702	0.69	0.40	0.63	0.11	0.26	0.10	0.16	0.53	0.09	0.57	0.06	0.16	0.12	0.16	0.54	2.87	3.97
s010703	0.68	0.00	0.50	0.00	0.00	0.02	0.16	0.68	0.00	0.50	0.00	0.00	0.00	0.16	0.65	2.04	3.55
s010704	0.56	0.19	0.58	0.08	0.18	0.09	0.16	0.54	0.08	0.48	0.11	-0.03	0.22	0.15	0.51	1.00	2.56
s020702	0.63	0.30	0.64	0.03	0.30	0.04	0.17	0.63	0.30	0.60	0.08	0.21	0.17	0.17	0.52	2.71	3.71
s020703	0.71	0.41	0.61	0.10	0.22	0.09	0.15	0.76	0.53	0.54	0.07	0.09	0.14	0.16	0.57	1.92	3.52
s020704	0.72	0.41	0.64	0.10	0.32	0.09	0.18	0.76	0.51	0.63	0.08	0.27	0.16	0.18	0.65	2.09	2.47
s050702	0.57	0.14	0.54	0.09	0.09	0.09	0.16	0.54	0.08	0.57	0.07	0.14	0.15	0.16	0.52	1.96	2.70
s050704	0.53	0.11	0.65	0.07	0.31	0.08	0.17	0.50	0.07	0.65	0.06	0.30	0.13	0.17	0.54	1.88	2.97
s060703	0.87	0.00	0.50	0.00	0.00	0.03	0.13	0.87	0.00	0.50	0.00	0.00	0.00	0.15	0.86	2.35	3.68
s070702	0.44	-0.14	0.59	0.08	0.18	0.08	0.17	0.56	0.12	0.57	0.04	0.15	0.07	0.17	0.58	2.08	3.30
s170601	0.68	0.00	0.50	0.00	0.00	0.01	0.15	0.49	-0.06	0.55	0.05	0.10	0.11	0.16	0.58	2.39	2.91
s210602	0.50	-0.15	0.59	0.09	0.21	0.10	0.18	0.60	0.16	0.61	0.09	0.24	0.20	0.18	0.59	2.55	2.60
s220602	0.76	0.00	0.50	0.00	0.00	0.01	0.15	0.55	0.09	0.57	0.08	0.14	0.17	0.15	0.72	2.17	3.17
s230602	0.53	0.08	0.63	0.03	0.35	0.03	0.16	0.50	0.00	0.50	0.00	0.00	0.00	0.16	0.59	3.12	3.37
s230603	0.77	0.56	0.79	0.09	0.57	0.09	0.15	0.68	0.37	0.82	0.03	0.65	0.06	0.16	0.52	1.67	2.66
s230604	0.56	0.12	0.64	0.10	0.28	0.10	0.17	0.47	-0.06	0.58	0.08	0.16	0.16	0.17	0.5	2.53	3.15
s250601	0.61	0.18	0.60	0.04	0.20	0.04	0.16	0.47	0.01	0.44	0.05	-0.12	0.11	0.16	0.53	1.62	3.24
s250602	0.69	0.52	0.65	0.10	0.34	0.10	0.15	0.77	0.55	0.54	0.10	0.08	0.19	0.13	0.56	2.78	3.69
s250604	0.56	0.06	0.60	0.05	0.27	0.06	0.16	0.53	0.03	0.55	0.05	0.11	0.10	0.16	0.58	1.99	3.48
s280601	0.65	0.30	0.56	0.06	0.12	0.06	0.15	0.43	-0.02	0.51	0.06	0.02	0.13	0.15	0.55	1.79	2.86
s280603	0.50	0.04	0.58	0.05	0.18	0.06	0.16	0.61	0.00	0.52	0.10	0.05	0.20	0.16	0.53	3.34	3.69
s280604	0.61	0.02	0.60	0.07	0.22	0.07	0.16	0.58	-0.01	0.60	0.09	0.19	0.19	0.16	0.63	1.95	3.88
s290601	0.68	0.00	0.54	0.04	0.17	0.05	0.16	0.49	-0.10	0.62	0.07	0.24	0.14	0.16	0.62	3.10	4.27
s290602	0.68	0.36	0.68	0.05	0.37	0.05	0.15	0.65	0.30	0.67	0.06	0.35	0.13	0.15	0.51	2.39	3.64
s290603	0.72	0.43	0.73	0.08	0.46	0.08	0.16	0.53	0.06	0.66	0.02	0.32	0.05	0.17	0.51	2.51	3.15
s290604	0.58	0.19	0.54	0.06	0.12	0.05	0.17	0.58	0.17	0.53	0.07	0.05	0.15	0.17	0.53	1.77	4.24
s290605	0.67	0.02	0.57	0.09	0.16	0.10	0.16	0.64	-0.04	0.47	0.06	-0.09	0.14	0.17	0.65	1.68	3.50
s300602	0.43	-0.19	0.53	0.02	0.11	0.02	0.16	0.57	0.16	0.52	0.08	0.04	0.16	0.16	0.55	2.87	3.88
Average (Std)	**0.62 (0.10)**	**0.15 (0.21)**	**0.60 (0.07)**		**0.22 (0.14)**		**0.16 (0.01)**	**0.59 (0.10)**	**0.12 (0.18)**	**0.57 (0.07)**		**0.14 (0.15)**		**0.16 (0.01)**	**0.58 (0.08)**	**2.26 (0.53)**	**3.32 (0.52)**

Table A.9: Valence classification results using ”accuracy” as scoring parameter for GridSearch. Learning models are highlighted in blue, over-fitted and under-fitted models are highlighted in yellow and orange, respectively.

Valence	SVM							MLP							EXTRA		
Participant	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Test Accuracy	MCC	CV Accuracy	CV Acc Std	CV MCC	CV MCC Std	Confidence	Chance Level	Avg Famili	Avg Liking
s010701	0.50	0.06	0.66	0.07	0.34	0.13	0.16	0.58	0.15	0.65	0.10	0.30	0.20	0.16	0.51	2.49	2.51
s010702	0.86	0.00	0.52	0.04	0.09	0.17	0.11	0.86	0.00	0.50	0.00	0.00	0.00	0.11	0.78	2.87	3.97
s010703	0.62	0.21	0.58	0.05	0.19	0.13	0.16	0.54	0.04	0.59	0.09	0.18	0.18	0.16	0.65	2.04	3.55
s010704	0.44	-0.12	0.61	0.10	0.22	0.20	0.16	0.54	0.08	0.59	0.10	0.19	0.19	0.16	0.52	1.00	2.56
s020702	0.73	-0.10	0.60	0.07	0.31	0.17	0.16	0.73	-0.10	0.60	0.09	0.24	0.25	0.16	0.69	2.71	3.71
s020703	0.76	-0.09	0.59	0.07	0.26	0.20	0.14	0.74	0.04	0.51	0.04	0.03	0.12	0.15	0.8	1.92	3.52
s020704	0.68	0.36	0.59	0.07	0.27	0.17	0.18	0.72	0.40	0.56	0.11	0.11	0.24	0.18	0.58	2.09	2.47
s050702	0.74	0.00	0.50	0.00	0.00	0.00	0.14	0.43	0.16	0.50	0.04	0.00	0.08	0.16	0.63	1.96	2.70
s050704	0.91	0.00	0.59	0.12	0.20	0.28	0.10	0.91	0.00	0.50	0.00	0.00	0.00	0.10	0.91	1.88	2.97
s060703	0.90	0.00	0.50	0.00	0.00	0.00	0.10	0.87	-0.06	0.50	0.00	0.00	0.00	0.12	0.92	2.35	3.68
s070702	0.59	0.19	0.64	0.13	0.28	0.27	0.17	0.63	0.24	0.64	0.06	0.28	0.13	0.17	0.53	2.08	3.30
s170601	0.68	0.26	0.74	0.04	0.50	0.09	0.15	0.70	0.39	0.68	0.03	0.38	0.06	0.15	0.59	2.39	2.91
s210602	0.70	0.41	0.57	0.05	0.23	0.13	0.16	0.50	-0.18	0.55	0.07	0.10	0.16	0.18	0.57	2.55	2.60
s220602	0.66	-0.07	0.56	0.06	0.17	0.13	0.15	0.63	0.25	0.56	0.08	0.12	0.16	0.15	0.64	2.17	3.17
s230602	0.58	0.19	0.69	0.06	0.39	0.12	0.16	0.53	0.10	0.63	0.08	0.27	0.15	0.16	0.51	3.12	3.37
s230603	0.61	0.23	0.74	0.04	0.49	0.10	0.17	0.58	0.16	0.81	0.06	0.63	0.11	0.17	0.51	1.67	2.66
s230604	0.65	0.25	0.62	0.06	0.26	0.12	0.16	0.62	0.31	0.56	0.05	0.13	0.10	0.16	0.6	2.53	3.15
s250601	0.66	0.30	0.64	0.05	0.35	0.11	0.15	0.58	0.14	0.60	0.05	0.20	0.12	0.16	0.58	1.62	3.24
s250602	0.66	0.24	0.75	0.06	0.51	0.12	0.16	0.71	0.37	0.70	0.06	0.40	0.13	0.15	0.53	2.78	3.69
s250604	0.75	0.26	0.62	0.11	0.31	0.25	0.14	0.75	0.29	0.62	0.10	0.29	0.23	0.14	0.67	1.99	3.48
s280601	0.59	0.31	0.61	0.08	0.28	0.16	0.16	0.57	0.26	0.63	0.09	0.28	0.19	0.16	0.64	1.79	2.86
s280603	0.58	0.21	0.59	0.08	0.18	0.16	0.16	0.56	0.09	0.55	0.10	0.09	0.20	0.16	0.51	3.34	3.69
s280604	0.55	0.00	0.51	0.02	0.06	0.08	0.16	0.45	-0.16	0.42	0.05	-0.16	0.09	0.16	0.56	1.95	3.88
s290601	0.92	0.00	0.50	0.00	0.00	0.00	0.09	0.86	0.22	0.49	0.14	0.01	0.26	0.11	0.83	3.10	4.27
s290602	0.84	0.00	0.50	0.00	0.00	0.00	0.12	0.84	0.00	0.50	0.00	0.00	0.00	0.12	0.79	2.39	3.64
s290603	0.63	0.29	0.59	0.08	0.28	0.25	0.17	0.56	0.00	0.50	0.00	0.00	0.00	0.17	0.65	2.51	3.15
s290604	0.65	0.20	0.64	0.06	0.29	0.10	0.17	0.68	0.29	0.59	0.12	0.18	0.25	0.16	0.61	1.77	4.24
s290605	0.64	-0.04	0.54	0.05	0.13	0.19	0.16	0.55	-0.05	0.59	0.05	0.24	0.14	0.17	0.66	1.68	3.50
s300602	0.80	0.11	0.63	0.08	0.30	0.16	0.13	0.74	0.20	0.62	0.09	0.25	0.18	0.14	0.75	2.87	3.88
Average (Std)	**0.69 (0.12)**	**0.13 (0.15)**	**0.60 (0.07)**		**0.24 (0.14)**		**0.15 (0.02)**	**0.65 (0.13)**	**0.12 (0.16)**	**0.58 (0.08)**		**0.14 (0.12)**		**0.15 (0.02)**	**0.65 (0.12)**	**2.26 (0.53)**	**3.32 (0.52)**

A.4 Final experiment

The Test accuracy and MCC scores of the final experiment with subject-dependent strategy have been averaged and compared with default guessing of the majority class in fig A.15, while the cross-validated scores have been averaged and compared in fig. A.16. As it can be observed, the average accuracy scores are not significantly different than default guessing, but it should be considered that the maximization of MCC scores often sacrifices accuracy when datasets have unbalanced

labelling.

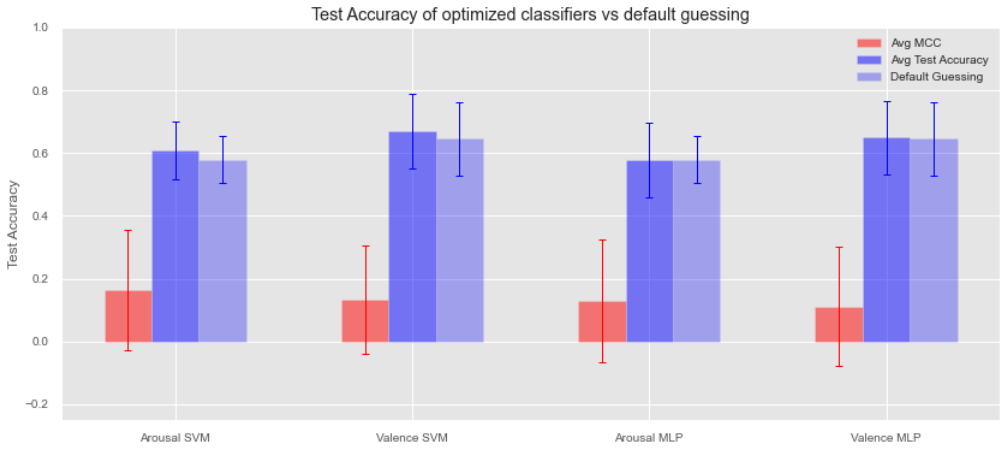

Figure A.15: Average Test Accuracy and MCC compare against majority class guessing.

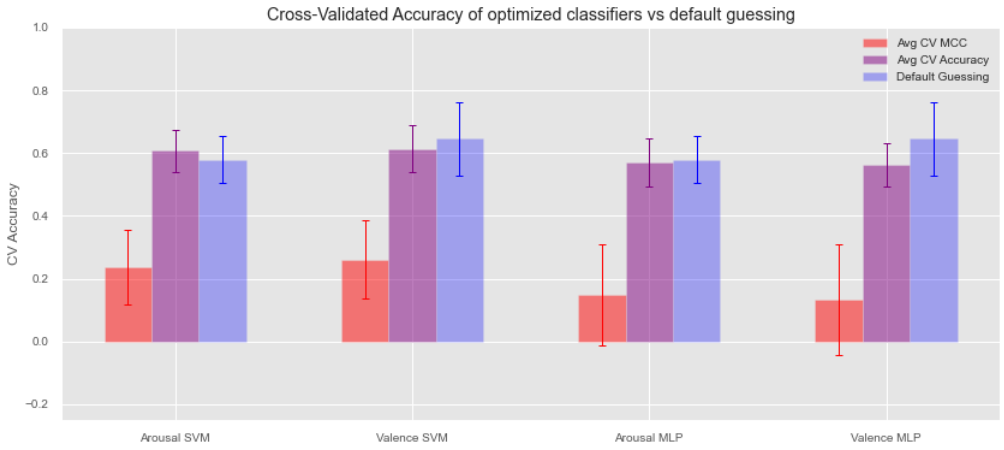

Figure A.16: Average CV Accuracy and CV MCC compare against majority class guessing.

A.4.1 Ranking of SVM classification performances for arousal classification

Test accuracy and MCC scores for arousal classification with SVM have been ranked by descending MCC (see fig. A.17), CV accuracy and CV MCC have been ranked by descending CV MCC (see fig. A.18), and compared with default guessing for

each participant. These ranking shows that the highest accuracy scores are often caused by unbalanced datasets.

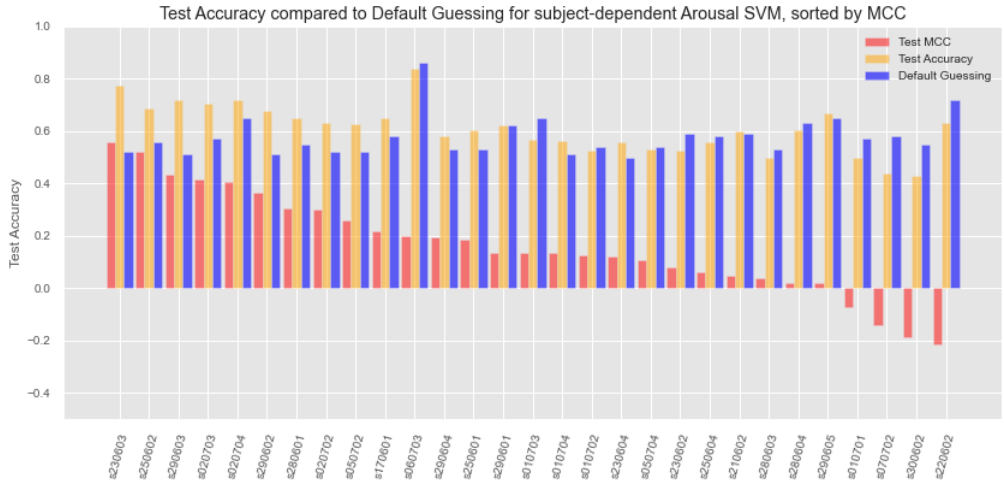

Figure A.17: Ranking of subject-dependent SVM models performances for arousal classification by Test MCC, descending.

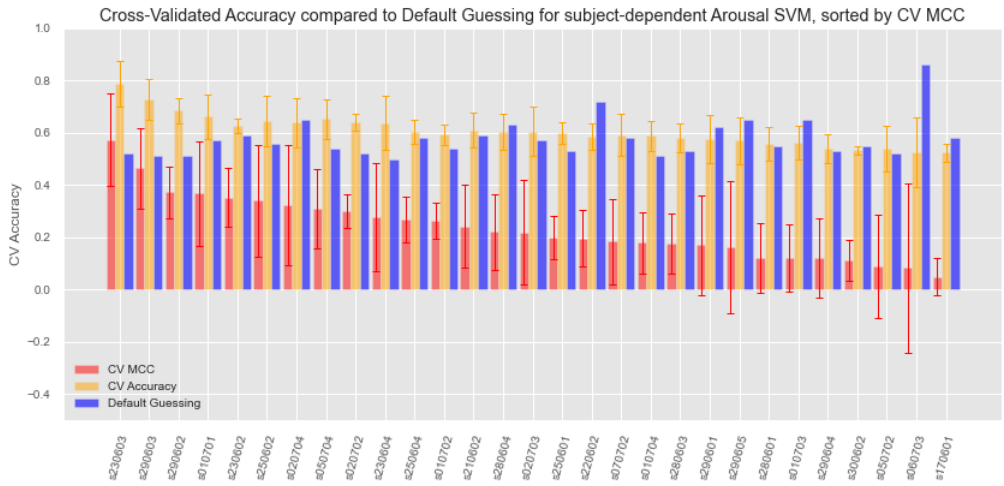

Figure A.18: Ranking of subject-dependent SVM models performances for arousal classification by CV MCC, descending.

A.4.2 Ranking of MLP classification performances for arousal classification

Test accuracy and MCC scores for arousal classification with MLP have been ranked by descending MCC (see fig. A.19), CV accuracy and CV MCC have been ranked

by descending CV MCC, and compared with default guessing for each participant (see fig. A.20). These ranking shows that the highest accuracy scores are often caused by unbalanced datasets.

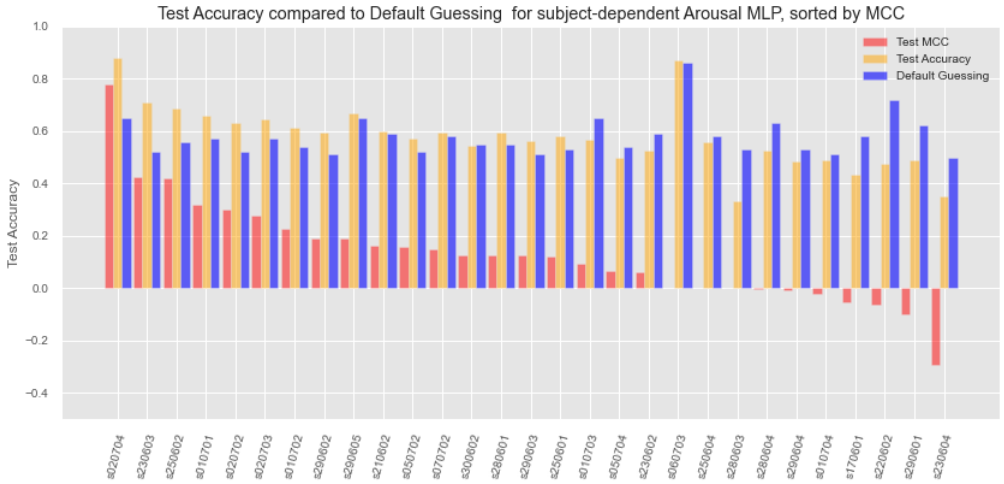

Figure A.19: Ranking of subject-dependent MLP models performances for arousal classification by Test MCC, descending.

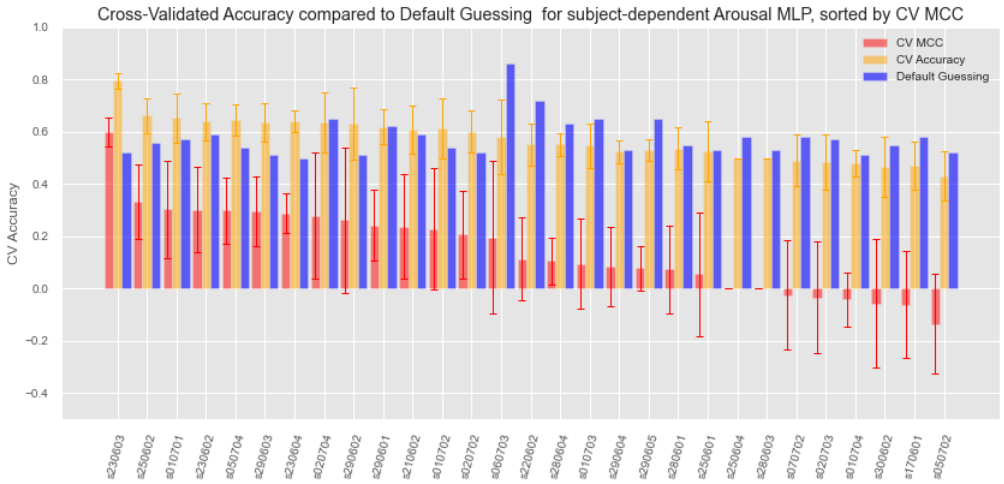

Figure A.20: Ranking of subject-dependent MLP models performances for arousal classification by CV MCC, descending.

A.4.3 Ranking of SVM classification performances for valence classification

Test accuracy and MCC scores for valence classification with SVM have been ranked by descending MCC (see fig. A.21), CV accuracy and CV MCC have been ranked

by descending CV MCC (see fig. A.22), and compared with default guessing for each participant. These ranking shows that the highest accuracy scores are often caused by unbalanced datasets.

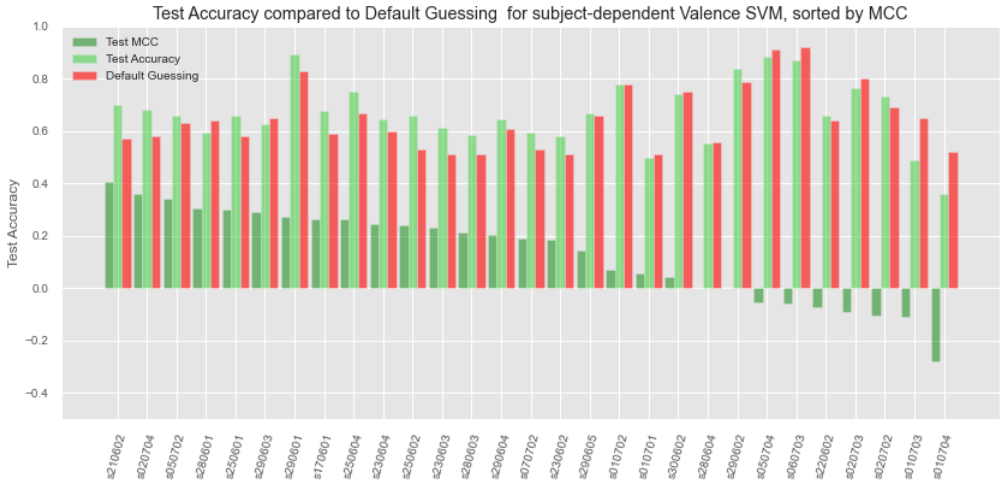

Figure A.21: Ranking of subject-dependent SVM models performances for valence classification by Test MCC, descending.

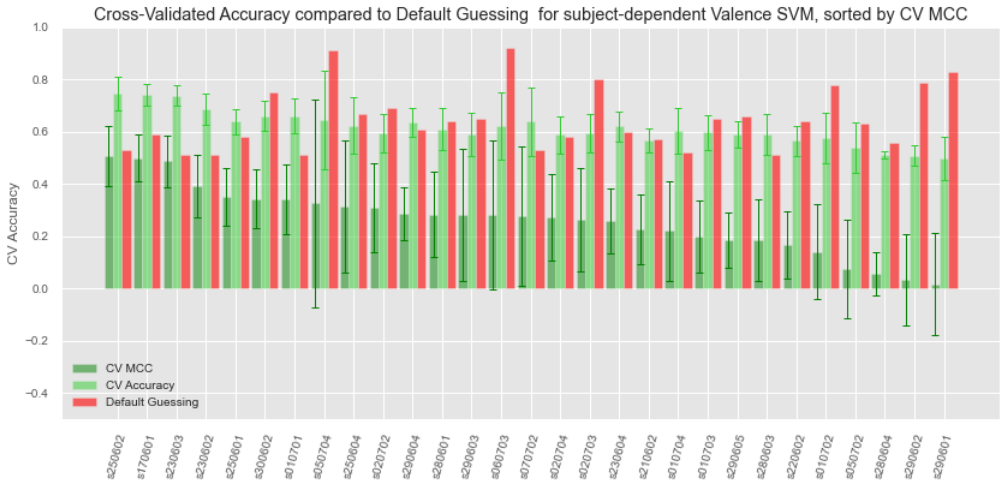

Figure A.22: Ranking of subject-dependent SVM models performances for valence classification by CV MCC, descending.

A.4.4 Ranking of MLP classification performances for valence classification

Test accuracy and MCC scores for valence classification with SVM have been ranked by descending MCC (see fig. A.23), CV accuracy and CV MCC have been ranked

by descending CV MCC (see fig. A.24), and compared with default guessing for each participant. These ranking shows that the highest accuracy scores are often caused by unbalanced datasets.

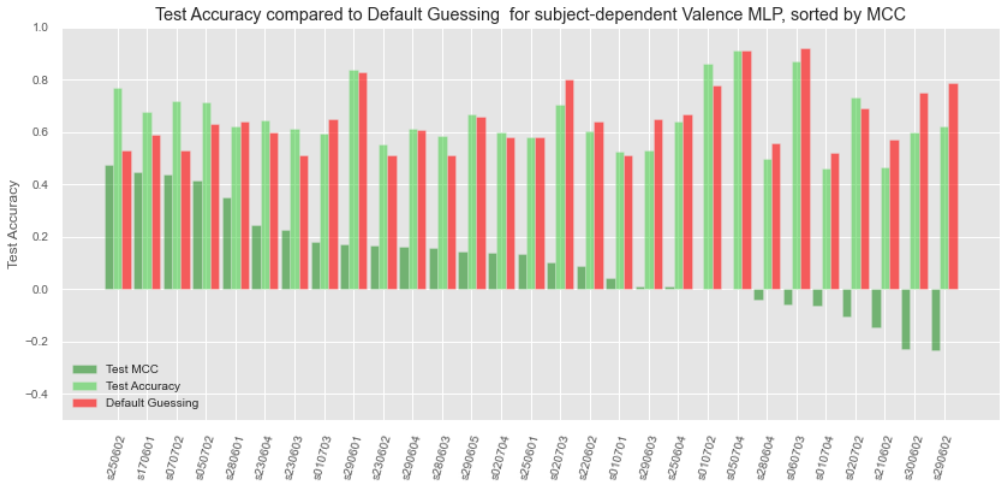

Figure A.23: Ranking of subject-dependent MLP models performances for valence classification by Test MCC, descending.

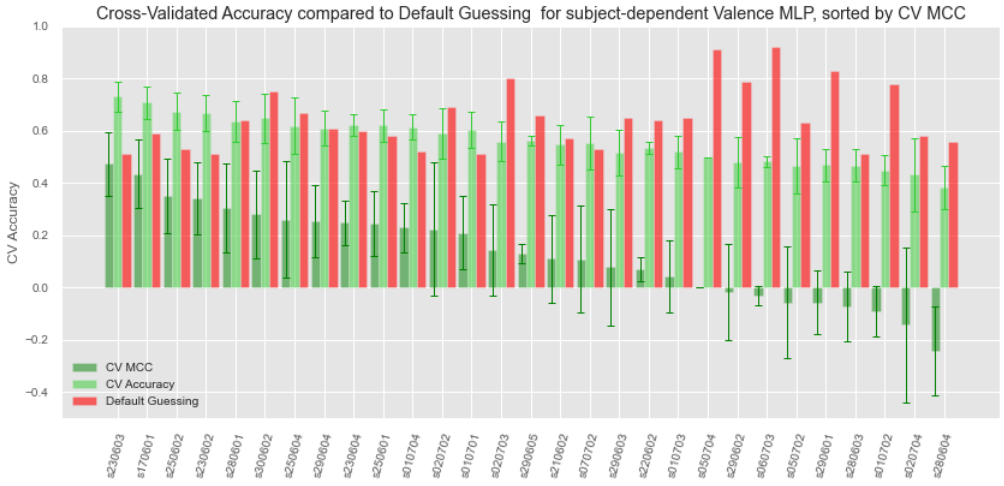

Figure A.24: Ranking of subject-dependent MLP models performances for valence classification by CV MCC, descending.