DTU Compute
MSc in Business Analytics

Biosignal synchrony analysis in collaboration activities

Carlos Ramos González

Master Thesis

Biosignal synchrony analysis in collaboration activities
Carlos Ramos González

Master Thesis
February, 2023

By
Carlos Ramos González

Cover photo: Vibeke Hempler, 2012
Published by: DTU, MSc in Business Analytics, Building 101, 2800 Kgs. Lyngby Denmark
www.dtu.dk

Approval

This thesis has been prepared over six months at the Section for Department of DTU Computer, Section for Statistics and Data Analysis, at the Technical University of Denmark, DTU, in partial fulfillment for the degree Master of Science in Engineering, MSc Eng.

It is assumed that the reader has a basic knowledge in the areas of statistics and data science.

Carlos Ramos González - s202961

.....

Signature

.....

Date

Abstract

The study of biosignals is a field that has experienced a rapid growth and interest in the last years, thanks to its variety of potential applications and the new techniques in data science. The use of these signals can be applied to understand synchrony and correlation between individuals working in an activity together. Hence, there is a big potential to understand if certain biological processes translated into biosignals provide information on interaction between humans, analyzing the synchrony during the collaborative process. To evaluate this, multiple individuals have taken part in a controlled collaboration experiment, where a set of biosignals have been recorded using a wearable device, together with their emotional state answered in a questionnaire. These biosignals include Electro Dermal Activity (EDA), Heart Rate (HR) and Temperature. The data acquired during the experiment is reliable and useful, and shows level of consistency along the experiments, with certain deviations. The participants who worked together have shown higher level of synchrony than during in non-collaborative scenarios, both in terms of biosignals and feelings. The feelings are seen to be directly related to the extracted features from the biosignals. Also, the potential application of this acquired data for predictive purposes has been analyzed, bringing very positive results with the use of Machine Learning (ML).

Acronyms

ACF Auto Correlation Function

AI Artificial Intelligence

ANOVA Analysis of Variance

ANS Autonomic Nervous System

ASD Autism Spectrum Disorder

BVP Blood Volume Pulse

CC Cross Correlation

CNN Convolutional Neural Network

CSCL Computer-Supported Collaborative Learning

CV Coefficient of Variation

DL Deep Learning

DTW Dynamic Time Warping

EDA Electro Dermal Activity

EDR Electro Dermal Response

ECG Electrocardiogram

ECTS European Credit Transfer System

EEG Electroencephalogram

EGG Electrogastrogram

EMG Electromiogram

EOG Electrooculogram

ERG Electroretinogram

FFT Fast Fourier Transformation

GDPR General Data Protection Regulation

GSR Galvanic Skin Response

HR Heart Rate

Hz Hertz

IBI Inter Beating Interval

IID Independent and Identically Distributed

ILO intended learning objectives

IPR Intellectual Property Rights

ML Machine Learning

NCD Non Communicable Disease

PACF Partial Auto Correlation Function

PDR Project Definition Report

OCD Obsessive-Compulsive Disorder

PGR Psychogalvanic Reflex

PPG Photoplethysmogram

RFC Random Forest Classifier

SPCA Sparse Principal Component Analysis

SCR Skin Condutance Response

SCL Skin Conductance Level

TLCC Time Lagged Cross Correlation

Contents

1 Introduction

1.1 Background

Collaboration and interaction among people are complex phenomena, whose evaluation is hard to perform, as they include several subjective variables. These difficulties in the quantitative and qualitative measurements of collaboration have been present for decades, without adequate proposals to target the problem.

A reliable method to understand collaboration metrics would allow us to understand which variables and factors have influence in this process. This knowledge can be used to monitor the collaboration level in daily activities; enhance and optimize the factors which drive collaboration; or link people who share complementary profiles and may have high collaboration levels.

Hence, this theoretical problem addresses several real problems in everyday life, and society would benefit of a deeper understanding of what collaboration implies.

A potential solution to understand collaboration with objective variables is the study of biosignals of the individuals subjected to collaborative work. Biosignals are defined as space-time records of a biological event, which can be measured and analyzed for different applications.

There exists a big set of these signals, which usually are associated with electric current in our body, such as: Electroencephalogram (EEG), Electrocardiogram (ECG), Electromiogram (EMG), Electrooculogram (EOG), Electroretinogram (ERG), Electrogastrogram (EGG), Galvanic Skin Response (GSR) or EDA.

The biosignals are not only restricted to an electrical potential on a given organ, like the ones enumerated before. They could have mechanical nature (accelerometer or mechanomyogram), chemical nature (level of oxygen and pH in a given tissue) and acoustic nature.

Some of the aforesaid mentioned signals require complex systems/procedures to be measured. The list of biosignals used for this project is reduced to a set of easier-to-acquire data, which include EEG, ECG, EDA, HR, skin temperature and body movement.

1.1.1 History of Biosignals

Biosignals have been used by humans since the very beginning of our history. The use of inspection, palpation, percussion and auscultation techniques have been used for thousands of years to diagnose any kind of disease or condition. Erasistratus, prominent Greek physiologist, already used in the III rd Century BC the HR for diagnosis. However, the use of this biosignals was mainly qualitative and not quantitative [1].

The electrical nature present in the bodies of animals was described in the late XVIII Century by English scientist John Walsh and Italian anatomist Luigi Galvani. However, no quantitative analysis was performed more than the qualitative idea that muscle activation is due to electrical force [2].

The study and use of more refined biosignals did not happen until 1856, when Kolliker and Mueller discovered the electrical activity associated to heart rate. In 1870, Alexander Muirhead registered the first electrocardiogram during its studies in St Bartholomew's

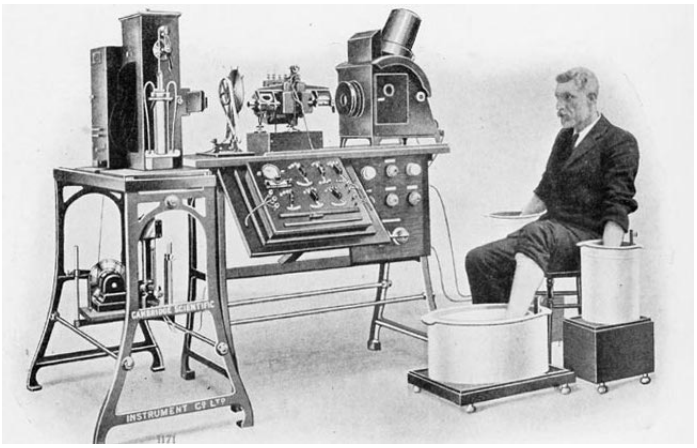


Figure 1.1: History of Biosignals: ECG machine of the 1910s, with electrodes attached to the patients measuring the signal [2]

Hospital, London [3]. The importance and advances of biosignals, like the ECG, is notorious as it is regarded as a key diagnostic and investigative tool. For example, Willem Einthoven, creator of the mercury capillary electrometer applied to ECG, received the Nobel Prize in physiology and medicine in 1924 [4].

ECG is just an example of the expansion of techniques that have converted the biosignals analysis to a more quantitative nature.

1.1.2 WristAngel Project

This project is part of a bigger, ambitious project called WristAngel [1], whose aim is to improve psychotherapeutic treatment of patients with Obsessive-Compulsive Disorder (OCD). According to the project responsible, acOCD, is a chronic, costly and disabling brain disorder for which existing treatments usually produce disappointing outcomes. According to the Diagnostic and Statistical Manual of Mental Disorders, the OCD is defined by the presence of obsessions and/or compulsions that are time-consuming, distressing or disabling. Obsessions are repetitive thoughts, urges or images that are intrusive and unwanted, and that in most individuals cause anxiety or distress, for example recurrent thoughts about accidental death [5].

The WristAngel project is supported by the EmpaticaE4, a wearable device that monitors several biosignals and can help to predict and identify OCD-episodes with the help of Artificial Intelligence (AI). More information about this device can be found in section 2.1.1.

1.2 Thesis scope and goals

1.2.1 Scope

The scope of this M.Sc. Thesis, even though it is part of the WristAngel project for OCD patients, it is exclusively restricted to non-patients. The access of pediatric patients data is very difficult and confidential due to its nature and will not be used.

Another point to remark is that no other biosignals than the ones measured by the Empatica E4 Device are measured nor used in the project. More recent devices from Empatica are available in the market at the time of this project release, but are not used.

[1] https://www.cachet.dk/research/research_projects/wrist-angel

This project is part of the study program of the author, and corresponds to a workload of 30 European Credit Transfer System (ECTS) credits, which is equivalent to 840 hours of work. The project is developed during a time span of approximately 6 months, which correspond to around 30 hours of work per week. The first month has been used on the PDR. This document includes a broad project description, the intended learning objectives and a project plan stating activities, milestones, deliverables and risks of the project.

1.2.2 Goals

The main research question of this thesis is the following:

Can we identify synchrony in biosignals of people performing a collaborative task with a wearable device?

This question addresses a real word problem for real world users, which in the mid-long term could benefit from a solution. The answer to this question can be done breaking the main research question in smaller ones:

- **Question 1**: Can we detect collaboration levels from the synchronization biosignals?
- **Question 2**: Can we generate reliable and useful biosignal data for the synchrony analysis? Can we set-up and carry out a new experiment to extend the dataset? Is data consistent along the different experiments?
- **Question 3**: Which biosignals show the biggest levels of synchrony among participants?
- **Question 4**: Can we generate predictive models with this data?

1.3 Literature review

1.3.1 Literature review: Biosignal features extraction

As seen in section 1.1.1, the biosignals have been widely used for years mainly for diagnosis purposes. However, the recent advances in computational methods for signal preprocessing, the advances in biosensors and applications, and the rapid growth and evolution of data analysis have produced a drastic acceleration of biosignals use. There are several representative studies of the combination of all these advances, for example we can find papers which address topics like: analysis of blink pattern activity using Convolutional Neural Network (CNN) [6]; stress detection techniques design via supervised ML[7] or hidden features acquisition of patients using Deep Learning (DL) [8]. The book *"Trends in biomedical signal feature extraction"* [9] sets forth the evolution of biosignals analysis over the last decades and presents the most recent techniques applied in the area.

As from Figure 1.2, we can find 4 generations of feature extraction techniques in biosignals:

- **Time domain feature extraction**: How values change over time. Some applications are: biometric modeling, time evolution of a given phenomenon.
- **Frequency domain feature extaction**: rate of change in signal values (usually transformed via Fourier transform). Some applications are: Signal separation and compression or ECG analysis.
- **Joint time frequency domain feature extraction**: decomposition of signals to acquire characteristics that in real world (non-linear and non-stationary signals) can be revealed.

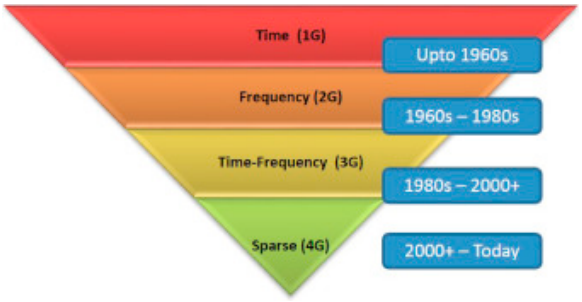


Figure 1.2: Biosignal features extraction: Evolution of biomedical signal feature extraction along different dimensions and decades. [9]

- **Sparse domain feature extraction:** further evolution of decomposing non-linear, non-stationary signals into sparse representations (minimizing signal representation). Like the previous, the use of Sparse Principal Component Analysis (SPCA) it has multiple applications to obtain insightful features on the analyzed signals.

The extraction of these features is highly dependent on the quality and reliability of the data generated and stored by the Empatica E4 device. Empatica has improved the access of monitoring for patients through wearable sensors. There are multiple examples of projects where biosignal features have been successfully extracted from an Empatica device and used. For example, we can find: Influenza and common cold detection before symptoms [10]; digital phenotyping measure of negative symptoms in schizophrenia via acceleration signals [11]; or arousal classification using EDA [12].

1.3.2 Literature review: Biosignals in collaborative processes

So far, only works related with general biosignal processing and applications have been discussed. The next set of works and ideas are more strongly related with the M.Sc. project, where we can find several works which link biosignals analysis of EDA, HR and temperature with social collaborative processes.

The study of 2018 by the Finnish Institute of Occupational Health [13] shows that physiologic signals, such as HR can offer an objective reliable base to measure collaboration. These measurements can become a proper quantitative and qualitative foundation to agree on such a biased measurement as it is interaction among humans.

The physiological signals can also indicate how people acting in collaborative environments with a common goal show similar behaviors. An interesting case of this pattern is the work published on 2022 by Fang et al. [14], where biosignals followed similar trends on individuals playing board games with same roles. In these processes, there is a learning process associated. When these learning participants collaborate as a group, their affective states are likely to converge, revealing synchrony [15].

Also, the work from Palumbo and colleagues [16], reinforces the idea that social processes are linked with body signals, as stating that "*social processes operate at the physiological level [...] synchrony due to external variables may be informative of shared levels of involvement*". However, this same work states that this research area is yet to be explored and we are in very early stages of the process. This is a motivation point for the thesis, as it touches areas not fully discovered. A similar work [17] shows the importance of psychophysiologic measures and their link with the exploration of conscious and unconscious processes.

One of the most important references for this project is the work from Mamberg [18],

where physiological synchrony and physiological arousal in groups are studied to evaluate Computer-Supported Collaborative Learning (CSCL). This study concluded that: the groups which experienced difficulties showed physiological synchrony, while those who did not experience difficulties did not show synchrony. It leads to the idea that currently there are no efficient methods and techniques to recognize socially shared regulation of learning, but new techniques may lead to better understanding on collaboration .

One of the most common signals that we can think of is the HR, as it is easily measurable with just a clock. This simple metric is a good indicator of regulated emotional responding, and can be used for better understanding of emotion processes [19]. As an example, the work from Wu et al. showed implied that certain states, such as amused condition, has a significantly lower heart rate than angry, fearful and neutral condition [20]. This idea is corroborated by the clinical trial performed by Kreibig in 2006 [21], where fear- and sadness-inducing films were used to analyze physiological patterns. The result of this work showed clear differential physiological response patterns for the different emotions induced by the films.

As stated before, this project is linked to the WristAngel project for OCD episodes prediction. The idea of studying synchrony in biosignals with interaction is not new. The team of Calabro published in 2021 [22] a dataset of biosignals studying the synchrony between patients with Autism Spectrum Disorder (ASD) and their therapist. Therefore, even though the application of these techniques is new, it seems that there is an increasing interest of AI applications within this medical field.

1.3.3 Literature review: Importance and growth

To understand the importance that the biosignals analysis in collaborative tasks, it is recommendable to understand the conclusions extracted from *Behrens et al*. [23]. This work showed that *"the strong relationship between our bodily responses and social behavior, and emphasize the importance of studying social processes"*, opening the door to further studies relating both fields.

There are multiple ongoing initiatives, works projects related to the expansion and applications of physiological biosignals in our daily life. A good example is the BITalino toolkit [24], which integrates multiple sensors with a software. It allows users to apply the kit in a wide range of daily activities.

Figure 1.3 shows some examples developed by BITalino users. The subfigures represent: (a) bicycle handlebars with heart rate monitor; (b)a muscle-activated door lock; (c) a keyboard for continuous ECG acquisition; (d) a game controller fitted with an ECG sensor; (e) heart monitoring on a mobile phone; and (f) an Android OS interface for heart monitoring.

The application of AI and DL on the biosignals features is analyzed by *Supratak et al*. [25], where the authors expose the current problems extracting worthy and reliable features for biosignal analysis. The usual way of extracting biosignal features is via hand-engineering (expert knowledge and established knowledge), which proves to be a fairly good solution. However, this hand-engineering feature approach needs to be combined with new Deep Learning techniques to learn and extract new features which are not extracted by previous knowledge. Nevertheless, the addition of these techniques, although promising, are computationally expensive and features are difficult to understand and analyze.

1.3.4 Literature review: Future perspectives

The evolution of the use of biosignals in different aspects of life in the coming years has very high expectations. This technology eruption and possible applications will experience the usual hype cycles in technology passing through peak, disappointment, and recovery

Figure 1.3: Different Applications of BITalino [24], ranging from heart rate monitor in a bicycle handlebar to an ECG sensor located in a game controller.

patterns [26]. The World Health Organization estimated in 2012 that around 70% of total global deaths correspond to the category of Non Communicable Disease (NCD) [27]. The use extension of biosignals could represent an important milestone in the prevention and detection of these NCDs. An early detection and treatment of these diseases driven by an established biomedical signal analysis roadmap [28] may reduce the consequences of these NCDs, such as: high healthcare costs, limited ability to work and financial insecurity [29].

Another point to mention is the rapid growth of wearable technologies, which allows easy integration of biosensors. In 2015, the diffusion of wearable techniques was already at an early adopter stage [30]. Therefore, nowadays, that technology is more and more present in our lives, which facilitates the physiological data acquisition process, accelerating the daily use of biosignals.

To summarize all the prior work, the study and processing of physiological biosignals is a well studied topic, which currently progresses in the direction of linking this knowledge with social aspects. That aim of study, together with the rapid evolution of wearable sensors and computational techniques, will allow this study area to grow in the future.

1.4 Thesis overview

1.4.1 Scope

The scope of this M.Sc. Thesis, even though it is part of the WristAngel project for OCD patients, it is exclusively restricted to non-patients. The access of pediatric patients data is very difficult and confidential due to its nature and will not be used.

Another point to remark is that no other biosignals than the ones measured by the Empatica E4 Device are measured nor used in the project. More recent devices from Empatica are available in the market at the time of this project release, but are not used.

This project is part of the study program of the author, and corresponds to a workload o 30 ECTS, which is equivalent to 840 hours of work. The project is developed during a time span of approximately 6 months, which correspond to around 30 hours of work per week. The first month has been used on the Project Definition Report (PDR). This document

includes a broad project description, the intended learning objectives and a project plan stating activities, milestones, deliverables and risks of the project.

2 Datasets

There are 2 sources of data: the biosignals recorded by the Wristband, and the participants surveys filled after every phase.

This section does not show in-depth descriptive analytics of the data used in the project, as those are considered parts of the Section 4 as the data is a result of the experiments. Consequently, the information is placed and described there.

2.1 Bio Signals

Most of the signals of biological nature analyzed in this project are driven by the human Autonomic Nervous System (ANS), which produces responses after given stimuli is received.

2.1.1 Measuring device: Empatica E4 Wristband

To acquire the data in the experiments, the participants wear an Empatica E4 Wristband. This wearable device contains several sensors [31]:

- Photoplethysmogram (PPG) Sensor: measures the Blood Volume Pulse (BVP), allowing to extract the HR.
- 3-axis Accelerometer: motion is measured.
- EDA Sensor: Changes in electrical properties of the skin.
- Infrared Termopile: measures skin temperature.
- Internal Real-Time Clock: high accuracy clock.
- Event button: marks beginnings and ends of events.

Hence, the following signals are possible to be extracted:

- EDA: Electro Dermal Activity measured in microSiemens (μS)
- HR: Heart rate frequency measured in Hz.
- TEMP: Temperature sensor data measured in Celsius (°C)
- ACC: Accelerometer values in x, y and z directions, measured in g.
- BVP: Blood Volume Pulse.

Figure 2.1: Lateral image of Empatica E4 device [31], like the one used in the experiment for data acquisition.

The Empatica E4 device stores up to 60 hours of data. However, this amount of storage is not necessary for the experiment, as each of the sessions lasts around 90 minutes. After every session, the device is connected to a computer and the raw data from the Empatica E4 is transferred via the E4 Manager Desktop App.

The device is held tightly in the wrist of the participant, preferably in the non-dominant hand. The inner area where sensors are located must be tightly in contact with the skin constantly to avoid data acquisition errors.

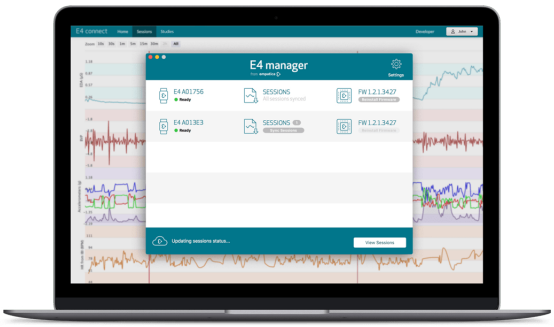


Figure 2.2: Empatica E4 Manager Desktop App, which allows visualizing the data recorded in the devices.

2.1.2 EDA - Electro Dermal Activity

The EDA is a property found in humans which show variable electrical conductance in the skin. The term is often mentioned as galvanic skin response (GSR), electrodermal response (Electro Dermal Response (EDR)), psychogalvanic reflex (Psychogalvanic Reflex (PGR)), skin conductance response (Skin Condutance Response (SCR)) and skin conductance level (Skin Conductance Level (SCL)). However, for sake of simplicity and, from now on it will be referred as EDA to understand the overall phenomena.
The EDA is a very useful signal from our body, as it provides a objective source to understand the emotional state of a person. As the Guide for Analysing Electrodermal Activity states [32], the *"EDA is arguably the most useful index of changes in sympathetic arousal that are tractable to emotional and cognitive states as it is the only autonomic psychophysiological variable that is not contaminated by parasympathetic activity."*.
The acquisition rate of this signal is set at ‘4 Hz.

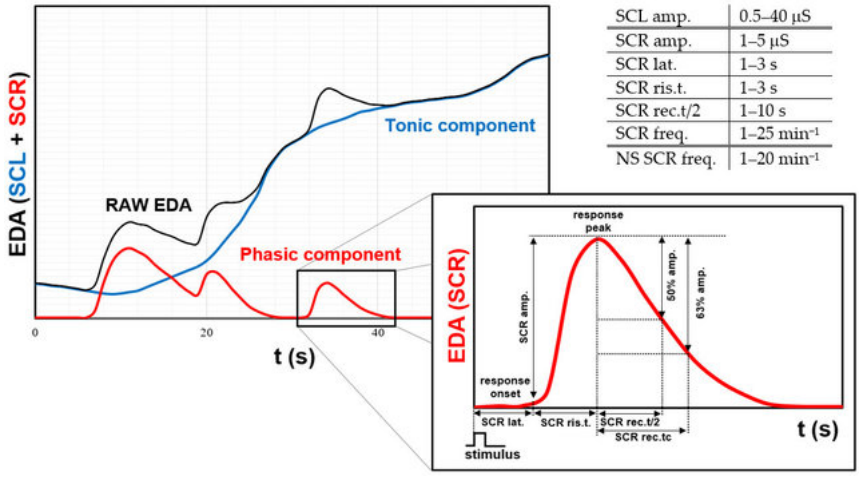
The EDA can be divided into 2 main components: the tonic EDA and the Phasic EDA

Tonic EDA - Skin Conductance Level
Tonic EDA is associated to slow acting components, and is mainly associated with the Skin Conductance Level. It is a moving signal, whose changes are considered slow. It is very common to standardize the Tonic EDA respect to a resting phase to correct for inter-individual variance.

Phasic EDA - Skin Conductance Response
This biological property is strongly related with the fast changes due to specific stimuli. It is associated to a sudden increase of the SCR around 3 seconds after the stimulus (latency in the signal). The threshold required to be considered a phasic activation is in the order of 0.05 μS.

2.1.3 HR - Heart Rate

Heart rate is the number of heart contractions or beats per unit time. It is measured under well-defined conditions and is expressed in beats per minute at the level of the peripheral arteries and in beats per minute at the level of the heart.



SCL amp.	0.5–40 μS
SCR amp.	1–5 μS
SCR lat.	1–3 s
SCR ris.t.	1–3 s
SCR rec.t/2	1–10 s
SCR freq.	1–25 min^{-1}
NS SCR freq.	1–20 min^{-1}

Figure 2.3: Typical EDA response [33], splitting the Raw EDA into their tonic and phasic components.

The HR is extracted from the BVP signal with a frequency of 1 Hz.

2.1.4 TEMP - Skin Temperature

Temperature of the epidermis surface of the participants, measure in the wrist. Historically, there are some body areas where body temperature has been measured, such as tongue or armpit. However, wrist is a suitable place to measure temperature as it is more responsive body part to overall thermal sensation [34].

2.1.5 BVP - Blood Volume Pulse

It is the blood volume change measured across the tissues in contact with the sensor. The BVP is widely used as the original source to derive the Heart Rate in several devices. The BVP is measured via the PPG sensor. This sensor emits an infrared light which is reflected by the tissue. The red blood cells absorb the radiation while other cells reflect it, calculating the difference.

2.1.6 ACC - Acceleration

Acceleration of the sensor moved in the wrist. This data measures the movement in the hand of the participant of the experiment. It is measured in the 3 axes.

2.1.7 BioSignals: Data Generated

After every session, the data from the E4 Empatica device is extracted and a zip file with raw data is downloaded. The raw data is composed of a set of 8 files, shown in the list below:

- **ACC.csv** Data from 3-axis accelerometer sensor
- **BVP.csv** Data from photoplethysmograph.
- **EDA.csv** Data from the electrodermal activity sensor
- **HR.csv** Average heart rate
- **IBI.csv** Time between individuals heart beats. Not used in this project. Derived from the diastolic points (local minima of the PPG)
- **tags.csv** Event mark times.Each row corresponds to a physical button press on the device; the same time as the status LED is first illuminated.

- **TEMP.csv** Data from temperature sensor
- **info.txt** Information file about the signals.

All the .csv files in follow the same format: The first row is the initial time of the session, expressed as unix timestamp in UTC. The second row is the sample rate expressed in Hz.

There is a set of files generated per round per participant, hence, 4 different sets in total per participant.

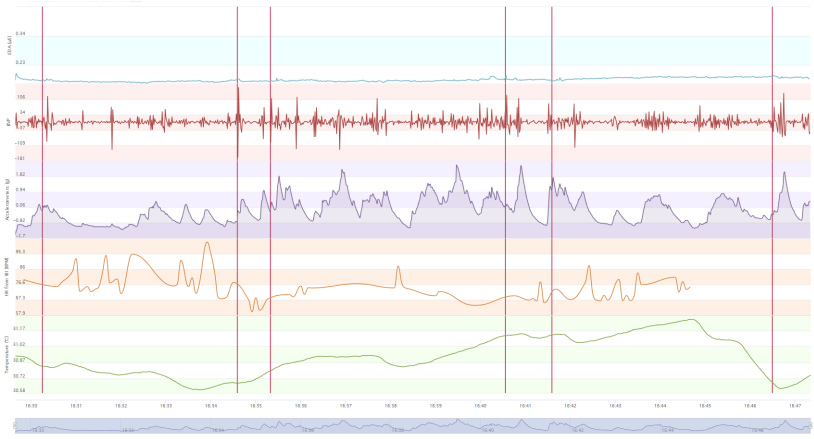


Figure 2.4: Visualization of the different biosignals recorded during the data acquisition experiment.

2.2 Participants Survey

The second data source of the experiment is the set of 12 questionnaires that the participants of the experiment fill at the end of every phase.

The participant answers every round on a Likert Scale (from 1 to 5) to their current state from a list of feelings and emotions. These feelings are: Upset, Hostile, Alert, Ashamed, Inspired, Nervous, Determined, Attentive, Afraid, Active.

In every phase, there is an extra question ”How frustrated are you feeling right now?”, which is answered in a scale from 0 to 10.

At the end of the second phase of each round (task phase), the user needs to answer the question ”How difficult did you find the task?”, which is answered in a scale from 0 to 10.

The questionnaire is located at Appendix A.

2.3 Data privacy

The participants sign an informed consent letter at the beginning of the experiment, where they agree that:

- Data will be collected, pseudo-anonymized and stored for a scientific purpose as described in the information letter – i.e. data is stored with a key which only the researches know.
- Data can be made accessible in a public research database in a pseudo-anonymized version with a random key which is not stored by the researchers. The data include physiological signals and emotion-questionnaire answers registered during the experiment. Name, age, and gender are NOT included.

Figure 2.5: Six GDPR principles to ensure accountability.

- Data collected before withdrawal from the experiment can still be used for the purpose of the experiment.
- Data is protected according to Danish law and European guidelines (GDPR;EU 2016/679).

Overall, the six General Data Protection Regulation (GDPR) principles to ensure accountability are followed, with the final aim of being compliant with all the data regulations. A schematic representation of these principles are seen in Figure 2.5.

3 Methods

The 3rd section of the thesis presents the different processes, methods and tools used in this project. It includes information about: the data acquisition setup and process; the data processing techniques; the technical explanation of the correlation techniques; the ML fundamentals used for the data.

3.1 Experiment design, set up and measurements

3.1.1 Experiment Design

Participants are divided in multiple teams of two and compete in a puzzle competition. Four teams will be competing in an alternating fashion, while the remaining teams will be in either the pre- or post-task resting phase. The competition consists of four rounds and has a duration of 76 minutes. The task of the competition is to solve as many tangram puzzles as possible in total during the four rounds. The task is designed to be difficult. A solved puzzle yields one point and the team with the most points at the end of the competition wins.

A round consists of 1+3 phases: wristband calibration (1 minute), pre-task resting (5 minutes), in-task puzzle solving (5 minutes), post-task resting (5 minutes). 60 seconds at the end of the pre-task, task and post-task phases are reserved to answer the emotion evaluation questionnaire. A timeline of a round is seen in Figure 1. The phases are indicated in the upper part of the timeline.

A team is handed a 7-piece tangram puzzle and a set of sketches of various puzzles. The participants of the teams have one role each: an instructor and a puzzle-solver. The puzzle-solver is the only one who is allowed to touch and assemble the puzzle pieces. The instructor is the only one who is allowed to look at the puzzle sketches. As a team they help each other to solve the puzzles. At any point, a puzzle can be skipped and another can be chosen to be solved. Team roles stay fixed throughout the competition.

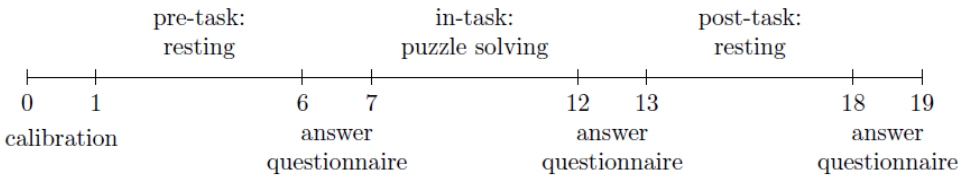


Figure 3.1: Timeline of a round of the puzzle competition depicting how phases and questionnaire answering periods are arranged. At the end of each round, the participants are to turn off their Empatica E4 wristband. Numbers, $[0:19]$, below the vertical line indicates the time in minutes since the beginning of the round.

For each round, at the end of the pre-task, in-task and post-task phases, participants will be asked to rate their feelings of a variety of emotions and how difficult it was for them to solve the task using the following questionnaire:

1. On a scale from 0-5, where 0 is not upset at all and 5 is extremely upset, how upset are you feeling right now?

2. On a scale from 0-5, where 0 is not hostile at all and 5 is extremely hostile, how hostile are you feeling right now?
3. On a scale from 0-5, where 0 is not alert at all and 5 is extremely alert, how alert are you feeling right now?
4. On a scale from 0-5, where 0 is not ashamed at all and 5 is extremely ashamed, how ashamed are you feeling right now?
5. On a scale from 0-5, where 0 is not inspired at all and 5 is extremely inspired, how inspired are you feeling right now?
6. On a scale from 0-5, where 0 is not nervous at all and 5 is extremely nervous, how nervous are you feeling right now?
7. On a scale from 0-5, where 0 is not determined at all and 5 is extremely determined, how determined are you feeling right now?
8. On a scale from 0-5, where 0 is not attentive at all and 5 is extremely attentive, how attentive are you feeling right now?
9. On a scale from 0-5, where 0 is not afraid at all and 5 is extremely afraid, how afraid are you feeling right now?
10. On a scale from 0-5, where 0 is not active at all and 5 is extremely active, how active are you feeling right now?

Additionally there are 2 other questions which are answered outside the table of the previous 10 feelings.

1. On a scale from 0-10, where 0 is not frustrated at all and 5 is ”I could not be more frustrated”, how frustrated are you feeling right now?
2. (Question only in phase 2): On a scale from 0-10, where 0 is not difficult at all and 10 is extremly difficult, how difficult did you find the task?

3.1.2 Set up

The data acquisition experiments were performed in DTU Compute facilities. There are 2 different sources of data: existing data from previous experiments, and new data acquired by the author of this report. The data below summarizes the 2 data sources.

Previous experiments

- Number of participants: 14
- Affiliation: Members of the section of Statistics and Data Analysis at DTU Compute, affiliated to WristAngel project
- Dates: 26/10/2022 (8 participants) and 28/10/2022 (6 participants)
- Location: DTU Compute Building 324

New experiments

- Number of participants: 14 (8+6)
- Affiliation: Members of DTU with no affiliation to WristAngel project
- Dates: 17/12/2021 (8 participants) and 17/08/2022 (6 participants)
- Location: DTU Compute Building 324

In none of the experiments the participants were compensated economically, but they were provided light snacks and drinks. Most of the participants were active students/researchers at DTU, so the study group is very homogeneous in terms of age.

3.1.3 Measurements

The data gathered from the participants surveys is treated and analyzed to understand the baselines and levels for every feeling. The average and standard deviation of values are calculated, understanding the differences along phases, rounds, roles and correlation among feelings.

It is very important to remark that there are multiple measurement per participant, as there are 12 different phases per participant. Therefore, the collection of measurements is not random and are not mutually independent. **Hence, the dataset is not independent and identically distributed (iid).** This implies that in exercises where all data is analyzed, no t-test can be conducted.

However, we can compare individuals assuming that each of the individuals is mutually independent.

3.2 Data processing

3.2.1 Raw Data

This section deals with the processing of the data since obtained from the E4 Empatica until it is transformed in a workable format.

Steps:

1. Excluding IBI measurements and sessions shorter than 1 minute.
2. Division of the data by phases: The data per round is divided in 3 phases with the information extracted from the tags.csv.
3. Store the new csv in ordered folders by phase, round and device.

The raw data input is transformed into data per phase of around 5 mins of duration, which are defined by every time the participant marked beginning and end of a phase.

3.2.2 Features extraction

In this step, some of the signals are analyzed and different metrics are obtained. It is very important to remark the nature of these features. As explained in section 1.3.1, there is an evolution of the feature extraction in the biosignals area. The features stated below are all comprised in the Time and Frequency domains, without including and sparse domain feature extraction. This implies that no high level decomposition is performed, and certain features remain hidden and not discovered in this project.

EDA

EDA processing is focused on the split between the tonic and phasic EDA values, which were described in 2.1.2. This process is made with the Python packaged Neurokit [35]. This package allows advanced biosignal processing and has a complete section dedicated to the analysis of EDA.
The decomposition into Phasic and Tonic is performed with the function. The sampling rate of this signal has a frequency of 4Hz.

The EDA Phasic and EDA Tonic are analyzed both in the time and frequency domains, in contrast with Temperature and Heart Rate which are only analyzed in the time domain. The analysis in the frequency domain are obtained with the Fast Fourier Transformation

(FFT). The FFT allows decomposing the time series analyzed to the frequency domain. More specifically, the FFT computes the discrete Fourier Transformation of the 2 EDA components. The DFT can be expressed as:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \tag{3.1}$$

The expression in Equation 3.1 can be expressed as a finite sum [36], like the one displayed below:

$$F(\omega) = \sum_{i=0}^{N-1} y_i(i\Delta t) e^{-j\omega(i\Delta T)} \Delta T \tag{3.2}$$

where N is the length of the Time Series and $y_i(i\Delta t)$ is the actual data value at time iΔt.

The variables can be summarized in the following list:

- EDA
  - Symp: Power of EDA in the 0.045-0.25 Hz range. According to Posada-Quintero, EDA Symp is a good indicator of orthostatic, physical and cognitive stress [37]. Derived via Neurokit package with eda sympathetic function.
- EDA phasic
  - Number of peaks - via Neurokit edapeaks() function
  - Average Rise and Recovery Time - via Neurokit edapeaks() function
  - Hjorth mobility and complexity
    - Statistic properties of signals usually applied in ElectroEncephalography, developed by Hjorth in the 1970s. [38]
    - via AntroPy package
  - Frequency and Time: mean and standard deviation
  - Frequency and Time: min and max
  - Frequency spectral entropy
    - Spectral power distribution of the signal. The value is based in Shannon entropy. [39]
    - via AntroPy package.
    - Method used: Welch
- EDA tonic
  - Frequency and Time: mean and standard deviation
  - Frequency and Time: min and max slope
  - Frequency and Time: min and max

HR (Only in Time Domain)

- Mean HR per phase
- Std deviation of HR per phase
- Min/Max HR per phase
- Gradient HR
- Min/Max gradient per phase

TEMP (Only in Time Domain)

- Mean TEMP per phase
- Std deviation of TEMP per phase
- Min/Max TEMP per phase
- Gradient TEMP
- Min/Max TEMP gradient per phase

3.2.3 Rolling moving averages

The rolling moving average calculated the unweighted mean value of the previous k values. This is particularly interesting to smooth time series and focus more on the general behavior of the series rather than specific peaks in the graph. A visual representation of this is shown in Figure 3.2.

$$RMA_k = \frac{p_{n-k+1} + p_{n-k+2} + ... + p_n}{k} = \frac{1}{k} \sum_{i=n-k+1}^{n} p_i \tag{3.3}$$

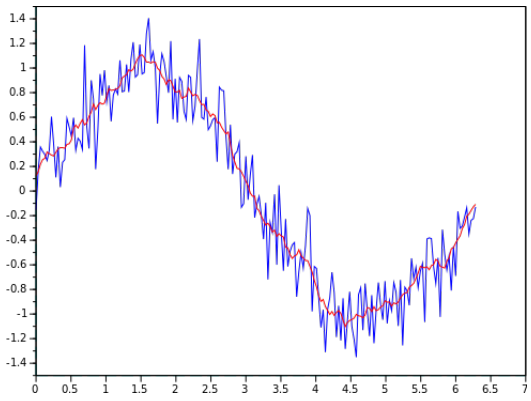


Figure 3.2: Example of a rolling moving average over a dataset. The blue line represents the original signal, while the red line is the smoothed rolling average.

The rolling average of the different graphs are calculated with 2 different time spans: 5 and 30 seconds. With these 2 time averages, we expect to observe both the short and long term behaviors.

3.3 Synchrony

This section deals with the different techniques used in the synchrony analysis of the data. It is important to note that this section strictly deals with time series correlation, while the next section deals with general correlation.

3.3.1 Cross Correlation (CC)

CC is a type of measure of association, used as a common approach of estimating the level to which two time series are correlated.

$$r = \frac{n(\sum xy) - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \tag{3.4}$$

The cross correlation measure allows understanding how two signals, normally considered as 2 independent random variables (except in autocorrelation) interact with each other.

The signals need to have similar amplitude, otherwise the values extracted from the cross correlation are misleading. If the signals are of different nature or have different amplitude, the series need to be normalized. In this project, all the series analyzed with cross correlation are previously normalized, even if they are of the same nature.

The output of the crosscorrelation formula range between -1 and 1. The interpretation of cross correlation is as follows:

- r = 1 The two signals are the same. For values smaller than 1 but still positive, it implies high correlation between signals. The closer the value is to 1, the more correlation exists.
- r = 0 . The signals are not correlated at all, and the information provided by signal X cannot be used to estimate the signal Y.
- r = -1. The signals are very correlated, however negatively. The signals move in opposite directions.

3.3.2 Auto Correlation Function (ACF)

Autocorrelation refers to the correlation between the same sequence of time series. This means how correlated the time series is with itself. This technique is specially useful to identify periodic and seasonal patterns in a signal.

The autocorrelation function assesses the correlation between observations in a time series for a set of lags. The way of understanding it is the duration and intensity of the memory of a time series value.

3.3.3 Partial Auto Correlation Function (PACF)

Partial autocorrelation is similar to the ACF, with the difference that it finds the correlation between the lag residuals and the next lag value rather than just between the lag values.

The use of the residuals allows discarding the already found correlation in the ACF, and analyze the non-found correlation (residuals) to identify patterns.

3.3.4 Time Lagged Cross Correlation (TLCC)

The TLCC measured the correlation between the datasets with a lag implemented in one of the time series. This technique allows us to understand if two time series are correlated, but there is a shift or delay between them.

The lag implemented can be either negative or positive, as it is necessary to analyze if one signal precedes the other or vice versa.

For a given time delay, we can assume that:

$$\Delta t = t_2 - t_1 \tag{3.5}$$

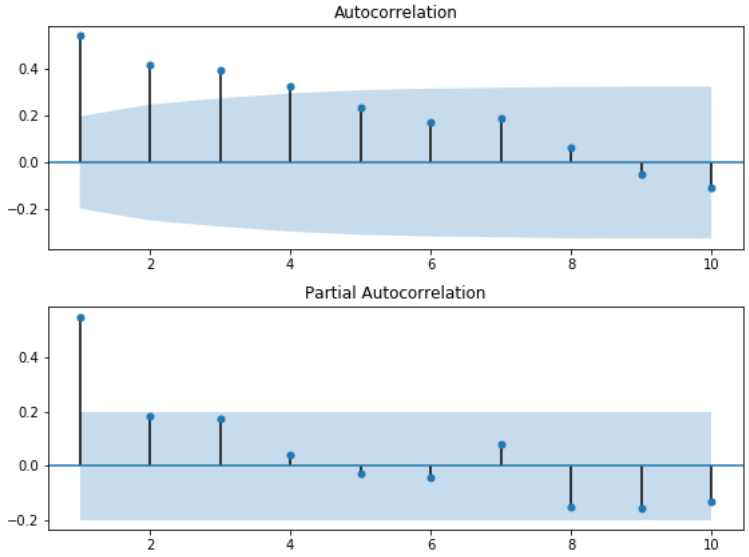


Figure 3.3: Example of an ACF - PACF plot of a time series.

$$\rho_{X,Y}(t1,t2) = \frac{cov(X_t1, Y_t2)}{\sigma_X, t1\sigma_Y, t2} \tag{3.6}$$

Hence, we can find two variables with this technique: the offset and the maximum correlation coefficient located in the lag range. This lag range is assumed to be 30 seconds in the study of all the variables. This is assumed to be a big enough time period.

3.3.5 Dynamic Time Warping (DTW)

DTW allows measuring similarity between time series. These time series can be of different length, although in this projected both signals are approximately of the same length [40].

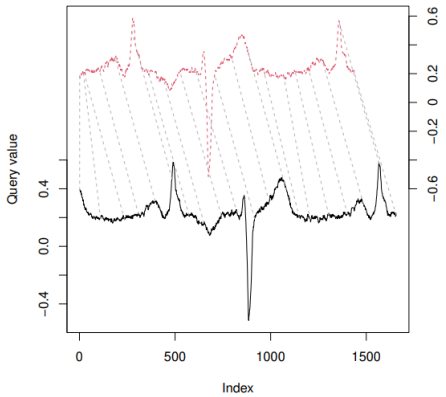


Figure 3.4: DTW: Time series aligning [40]

The objective of this algorithm is to build one-to-many and many-to-one matches, calculating the total distance between the two, and minimizing that distance. DTW outputs the cumulative distance between the two time series. It can be executed in Python and R.

3.3.6 Phase Synchrony

This measure assumes that the studies signals have certain frequency properties, hence have a phase associated to them. Assuming they have, the phase synchrony is a good

measurement of the instantaneous synchronization between the 2 signals. However, not only the instantaneous can be computed, but also the average phase synchronization between the 2 signals.

The values of phase synchrony take values in the range of [0 , 1].

3.4 Other methods for synchrony / correlation

3.4.1 Pearson Correlation

The Pearson Correlation Coefficient, PCC from now on, measures linear correlation between two different datasets. It is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.7}$$

The Pearson correlation assumes a linear association between the 2 sets of data. When nonlinear datasets, like the ones used in the project, the PCC loses reliability.

The correlation takes values between -1 and 1. Values close to 0 imply not correlation, while values close to -1 or 1 imply high correlation (negative and positive, respectively).

It is important to note that, although simplistic, the PCC offers a basic good understanding of the level of correlation between both variables. This is exemplified in 3 different properties:

- Variables may not be normally distributed
- Time series may not follow a linear relationship, as the biosignals usually follow non-linear behavior
- Data may have outliers. The measurement device is not perfect and may record and store outliers

3.4.2 Spearman Correlation

The Spearman Correlation deals with monotonic relationship between variables. This implies that, like the PCC, measures if given a value, the other increases or decreases, but not necessarily at a constant rate (non linear).

$$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X),R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \tag{3.8}$$

3.5 Statistical Tests

3.5.1 Analysis of Variance (ANOVA)

Definition and Information

ANOVA is the acronym for ANalysis Of VAriance, and is a test to compare the similarity among multiple groups of data. The ANOVA is strictly univariate, as there is a single dependent variable in ANOVA. In case of multiple dependent variables in the dataset, the MANOVA test is used. [41]

ANOVA is the extension of t-test or z-test to more than 2 groups.

Hypothesis

- Null Hypothesis: The mean value along different groups are equal: H_0 : μ_1 = μ_2 = μ_3
- Alternative Hypothesis: There is at least 1 group mean different from other groups. H_1 : Not all μ are equal

Assumptions

- Observations are sampled independently
- Continuous variables (if ordinal variables are used, it is preferred to user other tests, such as: Mann-Whitney U test, Kruskal-Wallis test)
- Experiment error is normally distributed

Formulas

ANOVA requires multiple steps to finally get the p value. The

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{3.9}$$

where Y_{ij} is the j-th observation in the i-th group, $\mu + \alpha_i$ is the mean for the i-th group and ϵ_{ij} is the error for the j-th observation in the i-th group

There are 3 sum of squares:

The total sum of squares:

$$SS_T = SS_B + SS_E = \sum (y_{ij} - \bar{y})^2 \tag{3.10}$$

The group sum of squares:

$$SS_B = \sum n_i (y_{ij} - \bar{y})^2 \tag{3.11}$$

The residuals sum of squares:

$$SS_W = \sum (y_{ij} - \bar{y})^2 \tag{3.12}$$

The F value is calculated as the ratio between Mean Squares (sum squares divided by population size)

$$F = \frac{SS_B / n_1}{SS_W / n_2} \tag{3.13}$$

Finally, the p value is obtained from the F value and the number of degree of freedom

How to interpret ANOVA test

The ANOVA test provides 2 main outputs: the F value and the *p* value.

The *p* value is the number describing how likely it is that your data would have occurred by random chance. For simplicity, a single threshold of 0.05 is applied ofr p value.

- p < .05: P value is statistically significant. There are strong evidences against null hypothesis.
- p > .05: P value is not statistically significant. There are strong evidence for the null hypothesis.

3.5.2 2 sample t test

Also known as Student's t-test, is a statistical method that compare 2 the means of 2 different groups. This statistic follows the t distribution [42].

Hypothesis

- Null Hypothesis: The mean value along 2 groups are equal: H_0 : μ_1 = μ_2
- Alternative Hypothesis: Mean of group 1 is higher than group 2 or vice versa . H_1 : $\mu_1 \neq \mu_2$

Assumptions

- Observations follow normal distributions
- Observations are sampled independently
- Continuous variables (if ordinal variables are used, it is preferred to user other tests, such as: Mann-Whitney U test)
- Experiment error is normally distributed
- Variances are unknown (otherwise, Z-test should be used)

Formula

The formula is shown below:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{3.14}$$

Where $\bar{x}_1$ $\bar{x}_2$ are the means of the independent samples, s^2 is the pooled sample variance and n_1 n_2 are the sample sizes.

The p value is found from the t distribution table.

3.5.3 Kruskal-Wallis test

The Kruskal-Wallis test, also referred as KW test, is an alternative test to the ANOVA test. This alternative is used when certain assumptions of ANOVA are not met. In this study, there exists certain non-continuous variables (feelings answered by the participants in the experiment) where ANOVA cannot be applied [43].

This test is a standardize version of the Mann-Whitney test, which can only be used for comparing differences between 2 groups.

The interpretation of the Kruskal-Wallis and ANOVA tests are similar. If p value is lower than the established threshold, there are significant differences between the datasets analyzed.

Hypothesis

- Null Hypothesis: The mean rank value along different groups are equal: H_0 : mr_1 = mr_2 = mr_3
- Alternative Hypothesis: There is at least 1 group mean rank different from other groups. H_1 : Not all mr are equal

Kruskal-Wallis test assumptions

- More than 2 or more independent groups
- Observations from independent group are randomly selected from target populations
- Dependent variable should be continuous or ordinal

Formula

$$H = (\frac{12}{N(N+1)} \sum_{j=1}^{k} R_j^2) - 3 * (N+1) \tag{3.15}$$

where N is the total sample size, k is the number of groups, n_j is the sample size of the j group and R_j is the sum of ranks in the j group.

3.5.4 Mann-Whitney U test

The Mann-Whitney U test is the equivalent of the T-test for ordinal and continuous variables. It is also a reduced version of the Kruskal-Wallis test [44].

Hypothesis

- Null Hypothesis: The median of both groups are equal: $H_0 : Mdn_1 = Mdn_2$
- Alternative Hypothesis: The median of the groups are different . $H_1 : Mdn_1 \neq Mdn_2$

Test assumptions

- Only 2 independent groups
- Observations from independent group are randomly selected from target populations
- Dependent variable should be continuous or ordinal

Formula

$$U_x = mn + \frac{m(m+1)}{2} - R_x \tag{3.16}$$

$$U_y = mn + \frac{m(m+1)}{2} - R_x \tag{3.17}$$

$$U = min(Ux, Uy) \tag{3.18}$$

where U is the Mann-Whitey statistic, m is the samples drawn from population X, n is the samples from population Y, and R_x R_y are the sum of ranks attributed to their populations.

The p value is calculated, comparing the U with the critical value (obtained from Table)

3.6 Machine Learning

3.6.1 Method: Random Forest Classifier (RFC)

Random Forest is a type of ensemble learning method use in machine learning, and can be used both in classification and regression. It is based on the combination of a set of decision trees which form an ensemble method. Ensemble methods, as the name suggest, is the combination of multiple predictive learning algorithms.

The ensemble of these decision trees is done via bagging, also known as bootstrap aggregation, which allows the final random forest method to get better predictions. This is due to generating multiple models from dataset samples, training the model independently and combining the result of all of the models according to their performance. This reduces the overfitting problem (compared to a single decision tree), but leads to longer calculation times.

The classification is based on the Gini index:

[1] https://dsc-spidal.github.io/harp/docs/examples/rf/

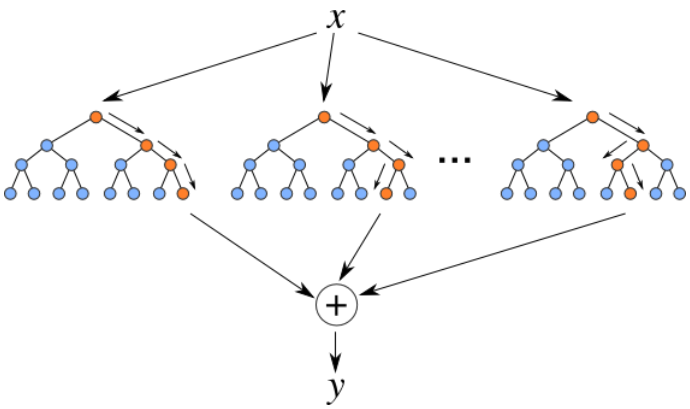


Figure 3.5: Schematic image of a Random Forest [1], showing the decision process from the given inputs (x) to the classification result (y)

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.19}$$

The Gini index allows understanding the occurrence probability of every branch, taking into account the number of classes (c) and frequency of the class observed (p_i).

3.6.2 Metrics

There are 4 metrics used in the evaluation of the model: Accuracy, precision, recall and F1.

Accuracy shows the ratio of the correct classifications divided by the total number of classifications. Accuracy is represented in Formula 3.20.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.20}$$

Precision shows the ratio of the true positives divided by the total number of instances classified as positive. Thus, it evaluates the ability of the model to classify a sample as positive. Precision is represented in Formula 3.21.

$$Precision = \frac{TP}{TP + FP} \tag{3.21}$$

The recall measures how the model detects positive samples, as it is divided by the total number of true values (True positive plus false negatives). Recall is represented in Formula 3.22.

$$Recall = \frac{TP}{TP + FN} \tag{3.22}$$

Lastly, F1 score is a measure used when there is an unbalance of categories and the precision and recall may not provide enough insights.F1 is represented in Formula 3.23

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.23}$$

3.6.3 Confusion Matrix

The confusion matrix is a table widely used in classification problems, which shows the performance of the model by displaying the predictions of the model. These predictions

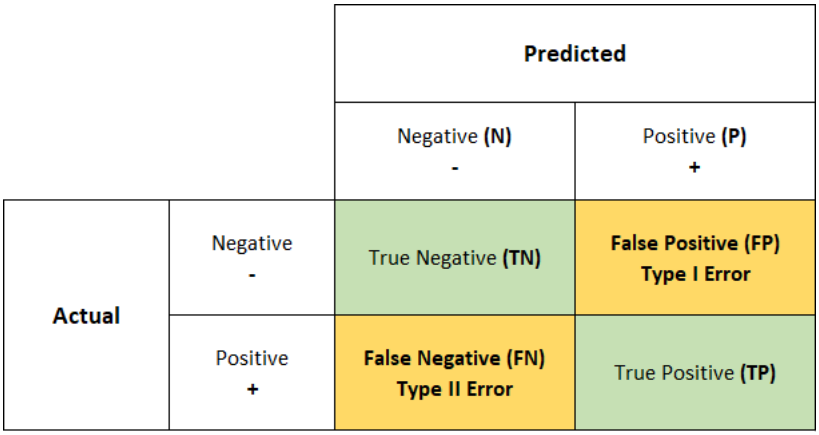


Figure 3.6: Confusion matrix for 2 categories [2], labeled as negative and positive. The green boxes represent correct classification (true negative and true negative), while the yellow boxes represent wrong classified instances.

are placed in the table according to the category they were classified and the one they should have been classified.

Figure 3.6 shows a 2 category classification confusion matrix. This matrix can be extended to n dimensions, which match the amount of categories of the model. In the analysis of this report, there are 2 classification exercises, one with 2 categories (like the one shown in the image), and one with 4 categories. It allows showing visually how model tends to classify the categories, showing some tendencies into a specific category and showing dataset unbalancing.

3.6.4 Feature Importance: Permutation

Feature importance allow us to understand how the input variables affect to the performance of the model.

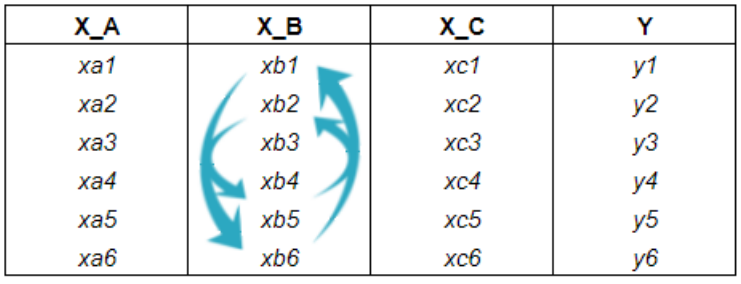


Figure 3.7: Permutation technique [3], in this case applied to column X_B, which works as input to predict variable Y

Permutation is a technique to calculate the importance of the different variables used in the model. These techniques evaluate the performance of the model when the values of the variable evaluated are scrambled. It uses the accuracy metric to evaluate its performance. This technique is represented in Figure 3.7, where variable XB values are permutated to analyze the ability of the model to predict the value y.

Other importance methods were considered for the project, including: impurity decrease

[2] https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5
[3] https://towardsdatascience.com/feature-importance-with-neural-network-346eb6205743

importance and drop column importance. The selection of the permutation feature importance method is its good reliability and efficiency, together with its good applicability to every model. However, its main con is its high computation expense. Although the main advantage of this method is that it does not require to retrain the model in every permutation, like the drop column technique.

3.7 Excluded methods

The last section addressed those methods that were originally in the scope of the project, but were excluded due to different reasons stated below.

3.7.1 Phase Synchrony

The use of phase synchrony in the analysis is discarded due to the nature of the variables studied in the biosignals. The variables are grouped in 2 main groups:

- Biosignal features: The time series per phase and round are summarized in different values that contain information on the analyzed. For instance, the mean value of the heart rate, the maximum temperature during the phase or the number of peaks of the EDA phasic. However, this represents a single value and the phase synchrony cannot be applied to it
- Biosignal rolling averages: this set of data is a time series and not a single point as the features. However, the use of rolling averages smoothes the curve and discards peaks and valleys, focusing more on the local trends of the values. Thus, the phase nature of the signal is lost as it does not have a proper phase. A visual representation of this issue is seen in Figure 3.2.

3.7.2 Dynamic Time Warping

Signals are of similar length, and from TLCC we can see that the offset is not big, where the maximum level correlation (in the time window) are not far in magnitude from the instantaneous CC. Plus, rolling averages (which change the understanding of the signal) are used, and immediate interaction between users lead to not use it.

3.7.3 ACF & PACF

Like in the DTW, the use of rolling averages distorts and eliminates the original signal nature. Also, the ACF and PACF deal strictly with 1 time series, and the information that these methods would provide would not be valuable in the scope of this project (no synchrony involved).

4 Results and discussion

This section shows and discuss the results of the different analysis performed on the datasets described in section 2. For a better understanding of the document, the results and discussion sections are merged in a unique section. Furthermore, the conclusion section also summarizes some of the most important insights of the discussion, which are strongly related with the research goals of the project.

4.1 Excluded signals

This first section addresses those signals that were originally in the scope of the project, but were disregarded due to different reasons.

4.1.1 ACC - Acceleration

The ACC signal is not used in the analysis, even though it was recorded in every single experiment. There are 2 reasons to discard this signal:

- **Not reliable recordings:** Some devices were measuring abnormal accelerations, such as the one observed in Figure 4.1. The acceleration should be constant along 0 in the 3 axes, except in the times where there is movement. However, it is observed that the device measures long periods of constant acceleration in the 3 axes. This does not makes sense as the participants are not continuously accelerating their hands. For simplicity, an example of participant 12 is taken during a resting phase.

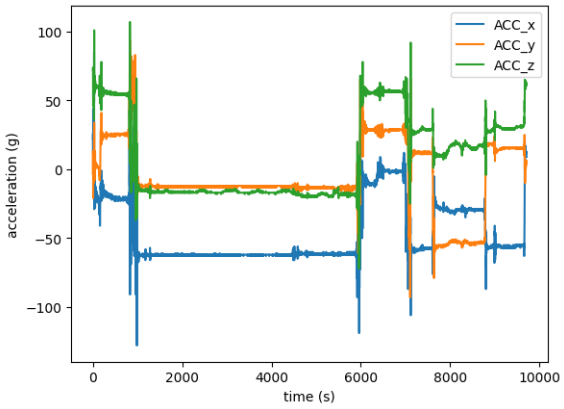


Figure 4.1: Example of abnormal accelerations measure: Acceleration sensor of Participant 12 in Phase 1, Round 1. The acceleration value should be constant near 0 g in all of the axis, however this behavior is not found.

- **Static hands on participants:** It was observed that some of the participants had the hand (in which the device was set up) completely static against the table. Hence, movement of the hand was negligible, even when using the other hand.

4.1.2 BVP - Blood Volume Pulse

The signal of BVP was found to be extremely noisy along all the participants. An example of this behavior is shown in Figure 4.2.

Besides, the main purpose of the BVP was the extraction of the HR signal, which is widely used in the analysis, being the HR also very correlated with the values of the BVP. Hence,

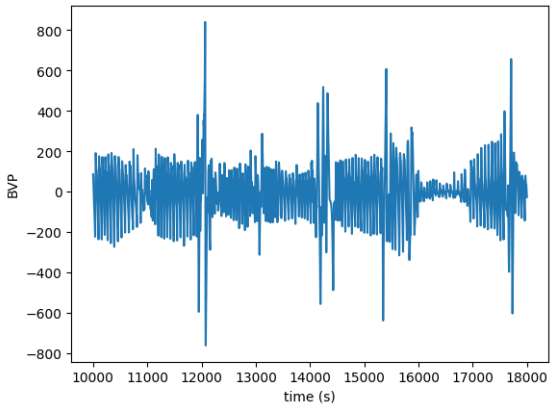


Figure 4.2: Example of BVP measure: Acceleration sensor of Participant 12 in Phase 1, Round 1. Noisy behavior is found consistently along participants and phases.

the BVP signal is not used in the analysis, even though it was recorded in every single experiment.

4.2 Participants

The participants' information is displayed in Table 4.1. Due to privacy issues, no personal information which can identify them is stored, other than its biosignals and the anonymized information related to the experiment (device used, team...)

# Participant	Device	Team	Role	Session	Comments
1	A03701	1	Child	1	
2	A0306B	1	Parent	1	
3	A02FBE	3	Parent	1	Round 1 discarded - Insufficient Data
4	A03857	3	Child	1	Round 1 discarded - Insufficient Data
5	A02B10	4	Child	1	
6	A03072	4	Parent	1	
7	A0388C	2	Parent	1	Round 1 discarded - Insufficient Data
8	A03804	2	Child	1	Round 1 discarded - Insufficient Data
9	A02857	5	Parent	2	
10	A03469	5	Child	2	
11	A02FC2	6	Parent	2	
12	A02FD7	6	Child	2	
13	A02FBE	7	Parent	2	
14	A02B10	7	Child	2	
15	A03469	8	Parent	3	Wednesday morning session
16	A03857	8	Child	3	Wednesday morning session
17	A03701	9	Parent	3	Wednesday morning session
18	A0306B	9	Child	3	Wednesday morning session
19	A03469	10	Child	3	Wednesday afternoon session
20	A03857	10	Parent	3	Wednesday afternoon session
21	A03701	11	Parent	3	Wednesday afternoon session
22	A0306B	11	Child	3	Wednesday afternoon session
23	A0306B	12	Child	3	Friday morning session
24	A03701	12	Parent	3	Friday morning session
25	A03857	13	Parent	3	Friday afternoon session
26	A03469	13	Child	3	Friday afternoon session
27	A03701	14	Parent	3	Error during data acquisition - Not used
28	A03857	14	Child	3	Error during data acquisition - Not used

Table 4.1: Summary of participants along the different data acquisition experiments. Teams 2,3 and 14 experienced difficulties which led to limitation on the use of the data.

As reflected in Table 4.1, there are some recordings which cannot be used due to errors during the data acquisition. These errors are due to:

- Insufficient session length:
  - The error has its origin in the manual recording of the phases by the participants, as they need to push a button to mark the beginning and end of the event.
  - Teams 2 and 3 did not mark correctly the phase and led to too short recordings in round 1. Hence, the values are discarded.
- Device turned off during session
  - The error is produced when the device restarts or turns off automatically due to some error.
  - Team 14 is discarded as Participant # 27 got errors during 2 rounds, and had data which was too short to be analyzed. Hence, the experiment is reduced to 13 teams.

4.2.1 Participants data: GDPR Principles

The use of the participant's data acquired during the experiment is carried out following the six GDPR principles to ensure accountability.

Principle	Requirement	Action
Lawfulness	The use of data must be transparent and fair. The data process and use should be used for the specific purpose agreed by the user.	The participants sign a document before the experiment where they are informed about the type of data stored and its purposes. The data is not shared nor used outside the scope of this project. No external use of this data is allowed.
Purpose limitation	The data collection can only be done for specified legitimate purposes. Explicit consent from participants is required.	As before, the participants sign a document where they are informed about the data collected and the purposed of their use. Without explciit consent, the participant cannot attend the session.
Data minimisation	To limit the amount of data collected. Only collect and retain what will be used in the future.	Participant data is limited to the questionnaire and the signals measured by the Empatica E4. Both data sources are extensively used along the project and data minimization was not required.
Data accuracy	To ensure the data is kept up to date and accessible by users.	Data is not updated after the execution of the 3rd experiment. Participants cannot access their data as the data is anonymized and there is no link recorded between the person and their ID.
Storage limitations	To only keep the necessary data. Delete the data when no longer necessary or anonymize it if stored for long.	Participant personal data, such as name, email or gender, which are used for experiment organizational purposes, are deleted as soon as the experiment is set up. Afterward participant information is anonymized by the use of a participant ID.
Integrity	To ensure a proper data safeguarding. Processors must protect the data against processing or loss.	All the data is saved locally in a local computer, saved by password. This local computer is only used for educational purposes by the researcher. Additionally, the physical data (questionnairs and declaration of consent), are stored in a closed office at DTU compute.

Table 4.2: The 6 GDPR principles to ensure accountability, together with the actions carried out to ensure that the requirements are met.

The requirements of the principles are shown together with the actions performed in the data acquisition process in Table 4.2. Overall, the 6 principles have been followed, and their requirements have been met when possible. The only deviation that exists is in the Data Accuracy requirement, as the users do not have access to their data. Their lack of access to the data is explained by the anonymized participant identification, which are linked to an ID number, and there is no document that traces the link between participants and ID.

4.3 Participants feelings

This section analyzes the responses from the participants after every phase. The questions and responses were described in Section 3.1.1.

The box plot shown in Figure 4.3 represents the distribution of feelings by the participants. The scale for ”frustrated” and ”difficulty” ranges from 0 to 10, while the scale for the rest is from 1 to 5. Certain feelings have very low distribution of values, such as ”hostile” and ”afraid”. Another noticeable thing is the low number of outliers, showing the majority of answers are insider the whisker.

The difficulty is a variable only answered by the participants after the task phase, not during the pre and post task resting phases. Positive feelings like inspired, determined, and attentive present the larger mean values, with the negative being presenting lower values.

These results induce the idea that the participants of the experiment answer the survey

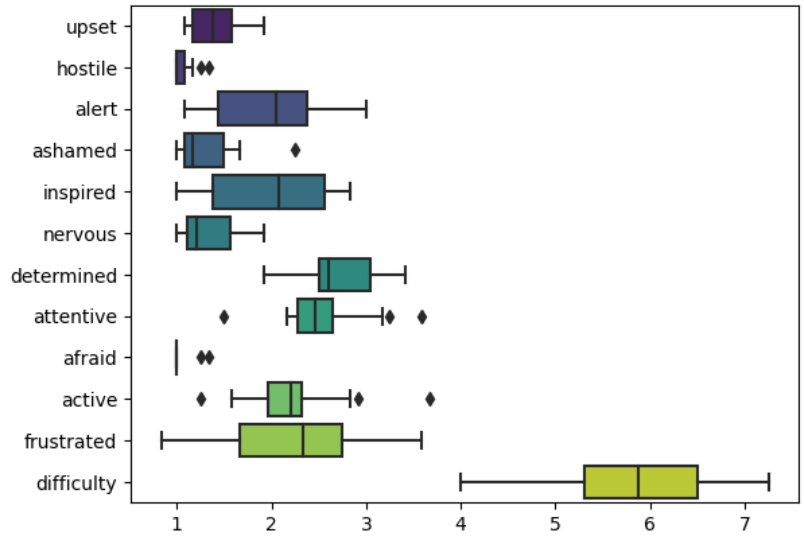


Figure 4.3: Participant feelings: Box plot with the distributions of feelings answered by the participants, including outliers.

in a truthful way, as certain feelings such as afraid or hostile have very low values. This is expected as the participants are not subjected to any kind of penalty nor hostile environment during the whole experiment. Also, no hostile/afraid behavior was shown by any of the participants during the experiment.

In Figure 4.4 the correlations between feelings are shown. Overall, we can find positive correlation among all the feelings, except for some specific cases (alert-hostile, alert-afraid). However, these cases with negative correlation occur with feelings with a very small variance. As expected, the highest correlation occurs among positive-positive and negative-negative feelings, such as active-determined, inspired-determined, attentive-determined, frustrated-upset, afraid-hostile.

4.3.1 Phase Analysis

The participants go through 3 different phases (pre-task, task, and post-task), collaborating to do the tangram puzzle in the 2nd phase and resting in the other two. Hence, we can group the pre-task and post-task phases and compare the feelings of the participants in respect to the task phase.

Figure 4.5 shows the different nature of some feelings in the phases. Overall, the feelings are reported stronger right after the task phase,

The frustration, active, attentive, determined and alert feelings are noticeably higher in the task phase than in the rest phase. This is an expected result that further validates the dataset, as participants are likely to experience more active-related feelings during the task phase than during resting periods.

4.3.2 Round analysis

The experiment consists of 4 rounds, like the one shown in Figure 3.1. It is interesting to analyze if the attitude of the participants changes with time (i.e., with rounds). From the correlation matrix, we can observe that the relation between the round and the feeling levels in negligible, and the participants do not seem to be affected by the round.

The difficulty associated with the task phase does not seem to be strongly influenced by the round either. A plausible reasoning behind this is the freedom of the parents to choose

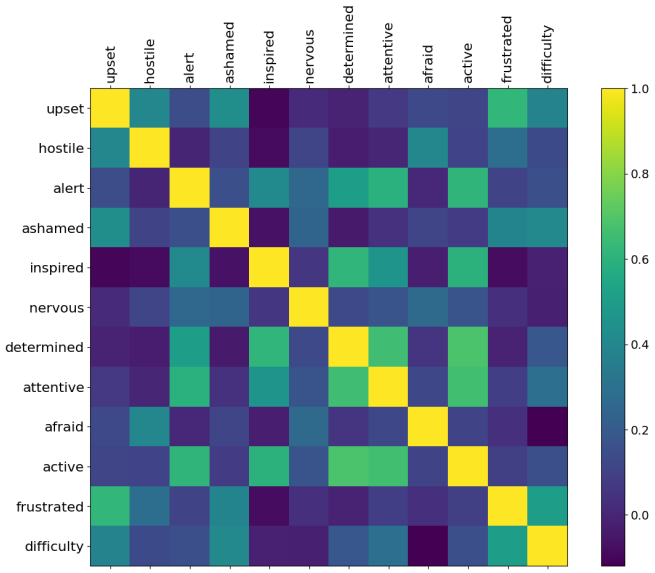


Figure 4.4: Participant feelings: Matrix with the correlation of feelings answered by the participants. Some of the feelings are very positively correlated, shown in brighter colours.

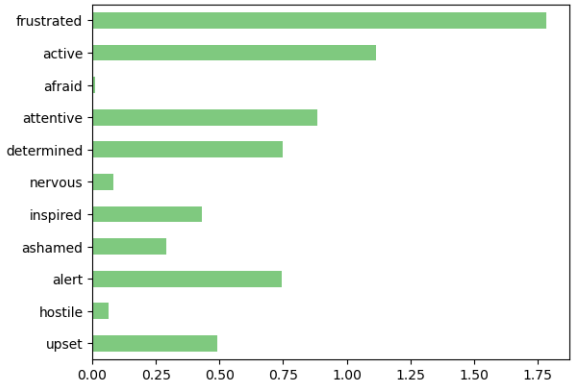


Figure 4.5: Average difference of feeling of participants between the task and rest phase. Value is calculated as the difference of feeling answered between task and rest phase. Positive values imply higher values during task phase than in rest phase.

the tangram they want, independently of their difficulty. Hence, the difficulty is completely dependent on the tangrams chosen in the given round.

Figure 4.6 shows the lack of clear influence of the round on the difficulty perception. From the three first rounds, we can observe a difficulty perceived growth with the rounds. However, this tendency is broken in the last round.

Also, for Figure 4.4, the correlation between the round variable and the feelings is very close to 0, displayed by dark colors in the correlation matrix.

4.3.3 Role Analysis

The participants play 2 different type of roles (parent and child), which are kept along the experiments. This subsection deals with the differences seen by the roles.

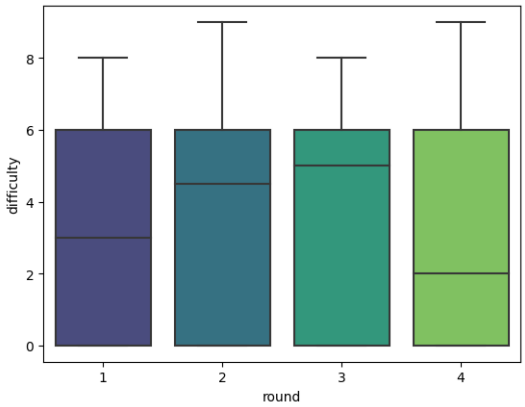


Figure 4.6: Participant feelings: Box plot with the difficulty level reported along the different rounds. The difficulty experienced by participants is similar along rounds

Figure 4.7 shows the average difference of feelings between the different roles.

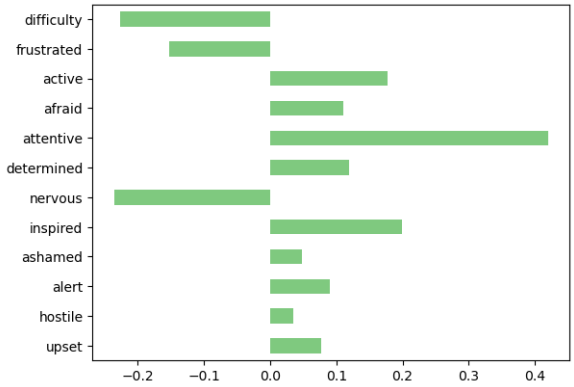


Figure 4.7: Average difference of feeling of participants between the parent and the child. Value is calculated as child minus parent. Hence, positive values imply higher feeling level for children. Difficulty, frustration and nerves are more felt by children/puzzle maker.

It is interesting to observe the higher intensity for parents of feelings in all the sentiments except in children, frustration and nerves. The parent feels higher difficulty, frustration and nervous. However, the child feels specially more active, inspired and attentive.

The negative values (sentiments stronger in parents than in children) are negative. This is an interesting discovery, as it seems that the one who gives commands is subjected to higher frustration and nerves associated to the higher difficulty perceived.

4.4 Similarity Validation

Three stages of experiments are held for the data acquisition process. The data from the 2 first experiments already existed and was provided before starting this project. The third experiment was set up and performed in the scope of this report in an effort to expand and improve the existing database. However, it is important to understand the similarity of these datasets, and test the results' reproducibility. The analysis of this section directly addresses the second research question *Can we generate reliable and useful biosignal*

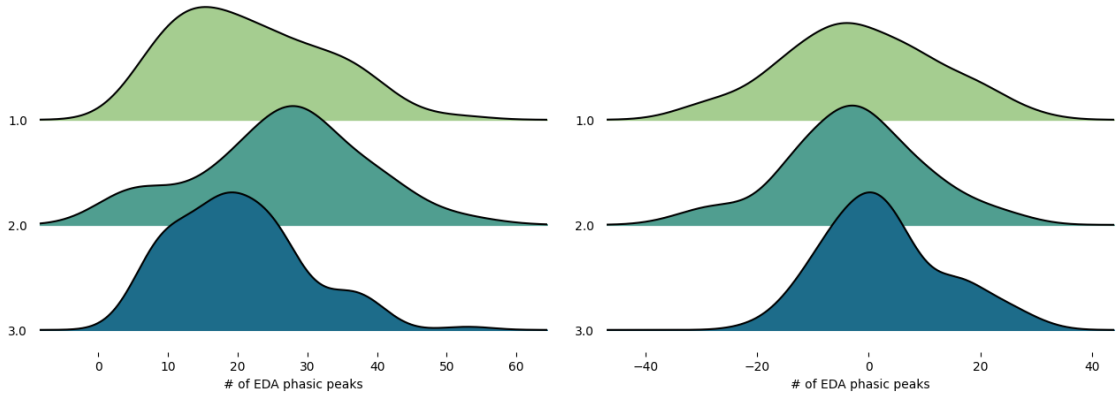
data for the synchrony analysis? Is data consistent along the different experiments?, which intended to analyze the data consistency along the different experiments.

To do this activity, the similarity validation is performed via visual inspection and statistical tests.

4.4.1 Visual analysis

For adequate visual understanding, the values studied along the different experiments are plotted via the JoyPlot package, which displays the density plot in different axes. The information obtained from this plot is similar to the box plot analysis, with the advantage that the distribution of values is visualized directly, without the shape of the values being misrepresented.

Example 1: Phasic EDA number of peaks
The first example is the Number of peaks measured in the EDA Phasic.



(a) Original dataset (b) Dataset respect baseline

Figure 4.8: Variables along experiments: Phasic EDA number of peaks. The distribution of number of peaks is similar in the 3 experiments, by visual inspection.

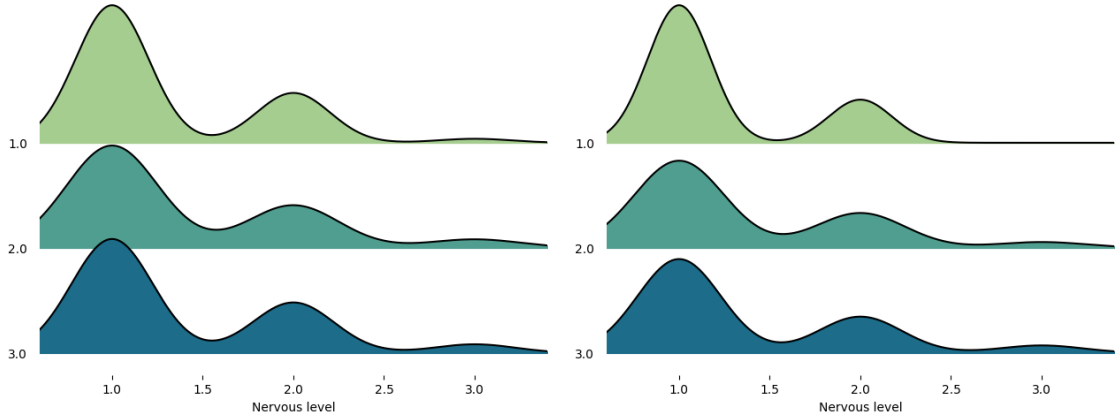
The graph, once compared with the baseline of the phase 1, seems to follow a normal distribution. The shape and scale of the graph is consistent among experiments, showing reproducibility of results via visual inspection for this variable.

Example 2: Nervous
The second example is the Nervous level of participants answered in the surveys. From visual inspection in Figure 4.9, we can see that the distribution of values is very similar along the different experiments, both in the original and normalized datasets.

For this specific variable, we can see a multimodal distribution. This is common in most of the feelings answers, as the answers in the rest phase and task phase are different. The left peak, with lower nervous level, corresponds to the rest phase. The lower distribution around a nervous level of 2.0, corresponds to the task phase answers.

This type of behavior is observed in the majority of feelings answers. Not only they don't follow a normal distribution, but do not follow any distribution. However, the visual inspection is positive for this type of data as the behavior is constant along experiments.

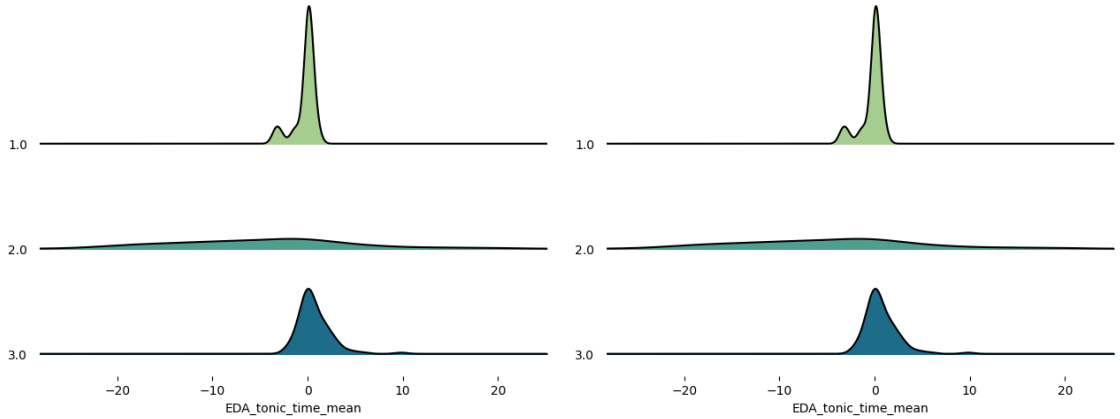
Example 3: EDA tonic values
The analysis of the EDA tonic variables is exemplified in Figure 4.10 and Figure 4.11. It is very clear that during the session 2, there is some apparent disturbance in the data, as



(a) Original dataset (b) Dataset respect baseline

Figure 4.9: Variables along experiments: Nervous level distribution along experiments reported by participants. Both datasets show very similar distributions along the 3 experiments.

the mean values of the tonic EDA are unusual. Usual values of the SCR amplitude is 5 μS.

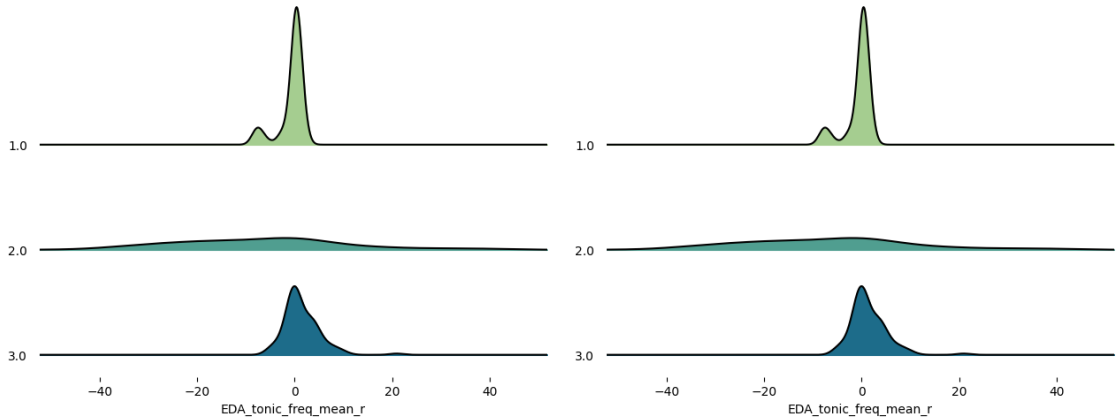


(a) Original dataset (b) Dataset respect baseline

Figure 4.10: Variables along experiments: EDA tonic mean value distribution (time domain). The distributions in experiment 1 and 3 are similar, while experiment 2 is very different from the rest.

The analysis is performed in the mean EDA value for both the time and frequency domain.

This issue is repeated in most of the EDA tonic values. In this point, it is convenient to analyze the possible reasons behind this issue. The EDA values are a measure of the skin conductance during the experiment. The skin conductance is affected by several factors, directly driven by the current emotional and physical state of the person. One of the manifestations of this state is the dermis transpiration level. Sweat affects electrodermal activity measures, increasing the spectral power of the EDA to higher frequency bands than normal state [45]. The second experiment was performed during summer, while the other sessions were performed during fall, where ambient temperature and humidity levels are usually very different. This is a plausible explanation of the values, which needs



(a) Original dataset (b) Dataset respect baseline

Figure 4.11: Variables along experiments: EDA tonic mean value distribution (frequency domain). Like in the mean value, the experiment 2 distribution is very different from the others.

to be taken into account for future works.

4.4.2 Statistical Tests

In this subsection, different statistical tests are performed to analyze the hypothesis that the different datasets along experiments share a common nature. Univariate statistical tests are used, repeating the test along the set of variables (features) coming from the biosignals. The plots displayed in the next subsections are the p value of the set of features analyzed.

Like in the previous section, there are 2 datasets analyzed: original dataset (features) and dataset respect baseline (features standardized after subtracting the rest value of round 1 phase 1).

It is important to recall at this point that our dataset does not respect the Independent and Identically Distributed (IID) principle, as there are multiple records from the same individual in the experiment. Hence, the use of the statistical tests here needs to be interpreted under that assumption, with illustrative purposes mainly rather than strictly statistical analysis.

For simplicity, a common *p* value threshold of <.05 is used for every test. This value shows strong evidence against the null hypothesis H_0, as the probability that the null evidence is correct is lower than 5%.

Another very important element to mention is that the rejection of the null hypothesis does not imply that the alternative hypothesis H_1 is true [1] . Hence, other methods, such as visual inspection, are encouraged to reach a conclusion about the datasets' similarity.

Continuous variables: ANOVA

The first analysis is the ANOVA, which uses the F-Test to assess how the different experiments differ from each other. Only the continuous variables are taken for the test, therefore the answers from the participants experiments are discarded in this section due

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017929/

to their discrete nature. The distribution of p values on the original and standardized (compared with baseline) datasets are shown in Figure 4.12

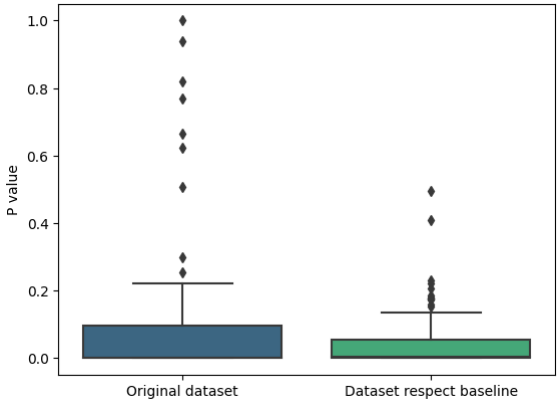


Figure 4.12: Variables along experiments: Continuous variable validation via the ANOVA test. The box plot shows the low p values obtained in the test, indicating a rejection of the null hypothesis.

The Table 4.3 presents the % of variables that lie above or below the threshold established of p value, then, accepting or rejecting the null hypothesis.

Dataset	**p< .05**	**p> .05**
Original	70%	30%
Baseline	75%	25%

Table 4.3: Variables along experiments. ANOVA test *p*-values distribution: Only a 25/30 % of variables accept the null hypothesis. This implies that the continuous variables along datasets do not seem to be similar.

It is observed that the majority of variables, both from the original features and normalized features have p value below the threshold, rejecting the null hypothesis. However, it is important to recall that the ANOVA test show insignificant *p*-values when at least 1 of the datasets compared does not follow the trend of the others.

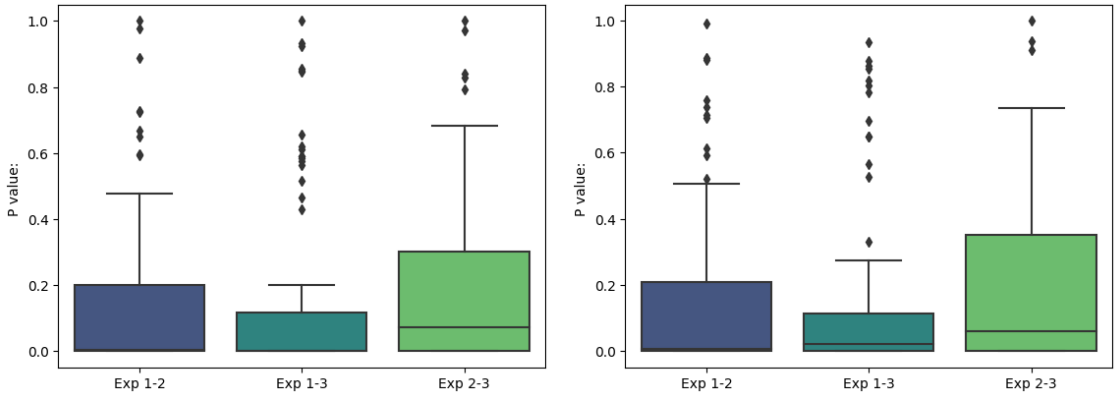
Therefore, the same test must be performed by couples between experiment datasets. The test to be used is not ANOVA, as ANOVA is strictly used for more than 2 datasets. Thus, the reduced version for 2 datasets, the *t*-test, is used.

Continuous variables: *t*-test

The *t*-test analyzes the similarity between 2 univariate datasets. The datasets are compared by pairs between experiment 1,2 and 3, so in total we will find 3 comparisons. The distribution of p values are shown in Figure 4.13. The notation Exp 1-2 corresponds to the t-test between the values of experiment session 1 and experiment session 2.

Like in ANOVA, it is helpful to identify how many of the variables are found to accept or reject the null hypothesis of experiment session similarity.

The results of the T-test show a higher percentage of values above the p value threshold than the ANOVA test. This is because it excludes one of the datasets from the test, which may differ from the other and fail the test.



(a) Original dataset

(b) Dataset respect baseline

Figure 4.13: Variables along experiments: Continuous variable validation via T test. Comparing individually the datasets improve the similarity of continuous variables.

Dataset	Experiment Session	p<0.05	p>0.05
Original	Exp 1-2	55%	45%
	Exp 1-3	63%	37%
	Exp 2-3	48%	52%
Baseline	Exp 1-2	63%	37%
	Exp 1-3	63%	37%
	Exp 2-3	49%	51%

Table 4.4: Variables along experiments. T test p-values distribution comparing the datasets by pairs. Between 37 % and 52 % of continuous variables accept the null hypothesis. These values represent an improvement respect the ANOVA test of previous section.

Experiments 1 and 2, the ones performed before the execution of this thesis, have around a 45% of similarity (percentage above the p = .05 threshold), while experiments 1-3 have a slightly lower amount of similarity. However, experiments 2 and 3 show the highest level of variables with p value above the threshold.

This is an interesting insight as from visual inspection we could see that the EDA values of Experiments 1 and 3 were actually similar, but not the ones from 2. This is an evidence why the validation of datasets should include statistical tests and not only visual inspection. Overall, around a 50% of validation among pairs of datasets is obtained. This value may seem low, but it must be taken into account the difficulties setting up and performing an experiment in a controlled environment which lead to repeatable results.

Also, the author of this report only performed experiment 3, and was not involved in the rest of experiments, making it difficult to recreate the same conditions in all of them.

Discrete variables: Kruskal-Wallis test

A different test is required to do a similar for the set of discrete variables, like the analysis in section 4.4.2. The use of a different test is explained by the suitability of ANOVA test only for continuous variables and not discrete. For that reason, the Kruskal-Wallis test,

which allows discrete variables, is applied here.

The discrete variables analyzed are the frustration level, difficulty experienced and the 10 Likert scale feelings stated in the participants questionnaire. For more information on these variables, please consult Section 3.1.1, where they are named and explained.

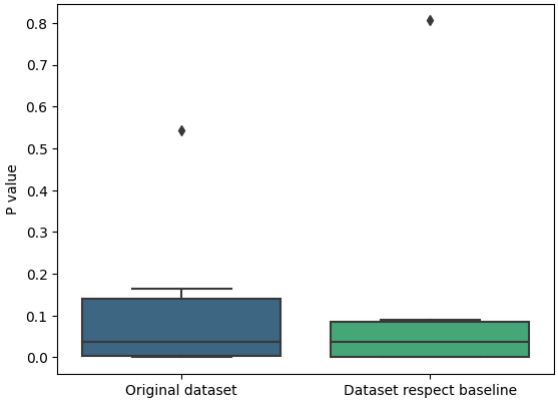


Figure 4.14: Discrete variable validation: Kruskal-Wallis test

The discrete variables show a higher level of similarity among datasets in respect to the continuous values. There are almost no outliers in the boxplot seen in image 4.14, suggesting that the distribution of similarities is similar along variables.

As seen in Table 4.5, the values which do not reject the null hypothesis H_0 correspond to the 50% of the whole dataset. This implies that half of the discrete values mean rank value are equal, while the other 50% have a different mean rank value.

Dataset	p< .05	p> .05
Original	50%	50%
Baseline	50%	50%

Table 4.5: Kruskal-Wallis test p-values: Discrete variables

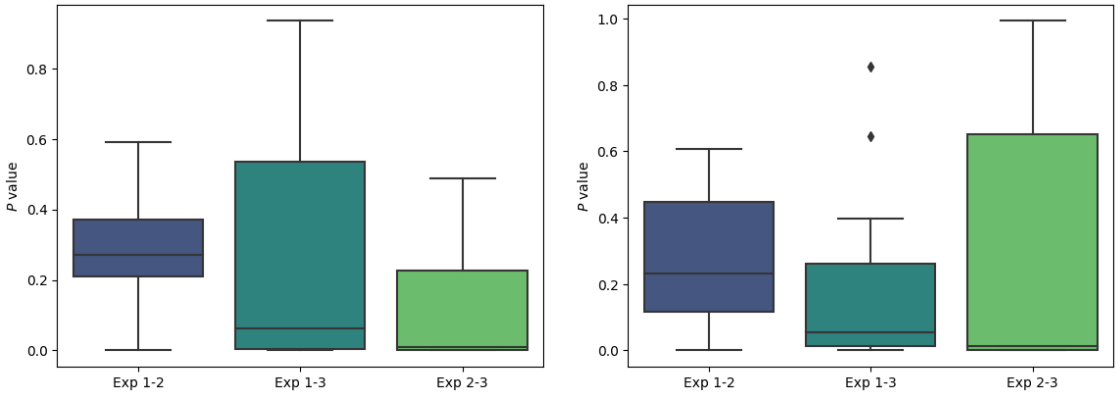
It is important to remind the reader that these discrete variables are completely subjective and dependent on the perception of the participant. Several factors can affect the answers of the participant, and differ the answers among experiment sessions.

Discrete variables: Mann-Whitney U test

The last statistical test of the section is the Mann-Whitney test, which is equivalent to the reduction of the Kruskal-Wallis test to a comparison of only 2 datasets. Like in the continuous variables, the test is applied by pairs on the 3 different experiment sessions.

Figure 4.15 shows the distribution of Mann Whitney U test *p*-values for the discrete variables along the experiment pairs. It is noticeable the increase of values above the threshold, which accept the null hypothesis and conclude that the mean rank of the variable along datasets is similar. This is very notorious for the experiments 1 & 2, with a median above 0.2 and most of the values above the p value threshold. The experiment 2 & 3 pair show however a median value lower than the *p* value threshold of .05. This suggests that the majority of discrete variables have rejected the H_0 hypothesis for this specific pair.

Checking the percentages of discrete variables above the threshold from the table above, we can observe a majority of discrete variables with similar mean rank value in the pairs



(a) Original dataset

(b) Dataset respect baseline

Figure 4.15: Discrete variable validation: Mann-Whitney U test along experiments

Dataset	Experiment Session	p< .05	p> .05
Original	Exp 1-2	17%	83%
	Exp 1-3	17%	83%
	Exp 2-3	66%	34%
Baseline	Exp 1-2	17%	83%
	Exp 1-3	17%	83%
	Exp 2-3	66%	34%

Table 4.6: Mann-Whitney U test p-values: Discrete variables

experiment 1-2 and Experiment 1-3, with over 80% of values which accept the null hypothesis. It is interesting to see that, even though the *p* value distribution of Exp 1 & 2 is by average higher than 1 & 3, the % of values above the threshold is the same. This is because it does not matter if a p value is .06 or .6, both of them compute in the same way (as accepting H_0) in the final percentage. The experiment 2 & 3 validation shows very different results from the other 2 pairs. Around 2/3 of the discrete variables rejected the null hypothesis, indicating that only 1/3 of them have similar mean rank in both datasets. This was expected from the boxplot of 4.15, with a very low median *p* value below .05.

It is interesting to understand why the participants expressed different feelings in experiment 3 respect to 2. As mentioned before, the level of subjectivity of the responses is very high, and several non-measured factors influence the answers.

4.4.3 Validation on experiment 3 - different sessions analysis

This section strictly deals with the similarity analysis between the different sessions held in Experiment 3. This experiment had 4 sessions, 2 held on a Wednesday (morning and afternoon) and a Friday (morning and afternoon). Due to the data shortage on the sessions held on Friday (morning session only had 1 team, and data from team 13 was discarded due to corruption), the similarity analysis is applied between the Wednesday sessions, which had 4 participants each.

This experiment 3 corresponds to the one performed by the author of the report, and it evaluates the consistency of the data of sessions performed in the same day under same controlled conditions.

For convenience, the ANOVA statistical test is used for the continuous variables, as in the previous section.

Dataset	p< .05	p> .05
Original	12%	88%
Baseline	64%	36%

Table 4.7: Variables along experiment 3 sessions. ANOVA test p-values distribution:Around 90 % of variables reject the null hypothesis. This implies that the continuous variables along Wednesday sessions are similar in Experiment 3.

Table 4.7 shows the distribution of values that accept or reject the null hypothesis of the test. Using the original dataset alone, the *p* values are mainly above .05, with an 88% of variables showing similarity along datasets. This is a really high values that imply the high level of similarity, much higher than in the tests performed for different experiments.

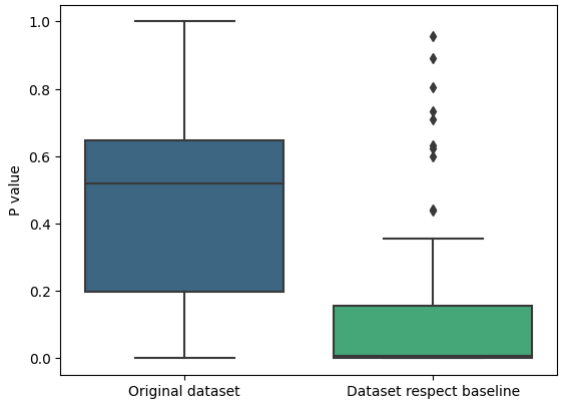


Figure 4.16: Variables along experiment 3: Continuous variable validation via the ANOVA test. The boxplot shows the high p values obtained in the test, indicating the similarity between variables along sessions (Wednesday sessions only.)

Figure 4.16 shows the distribution of *p* values after the ANOVA test. The majority of *p* values in the original dataset are above 0.05, while the normalized dataset (dataset respect to the baseline, which is the phase 1 of round 1), show lower values. These lower values suggest a bigger amount of dissimilarity between datasets. Taking only the original dataset, the results show a big inter-session similarity, where conditions were kept controlled.

Like in the previous subsection, different tests are applied according to the nature of the variable. The Kruskal Wallis test is applied for the discrete variables.

Figure 4.17 and Table 4.8 shown the results of the test on the discrete variables along the experiment 3 two first sessions. The results, in line with the continuous variables' analysis, show a very high similarity along datasets as the null hypothesis is not rejected. The null hypothesis states that the mean rank value along different groups are equal.

4.4.4 Similarity Validation Summary

Variability of the data in the experiment is a factor to be taken into account in the analysis. It has been shown above that data differ from one experiment from another, and a single

Dataset	p< .05	p> .05
Original	30%	70%
Baseline	30%	70%

Table 4.8: Variables along experiment 3 sessions. Kruskal Wallis test *p*-values distribution:Around 70 % of discrete variables reject the null hypothesis. This implies that the discrete variables along Wednesday sessions are similar in Experiment 3.

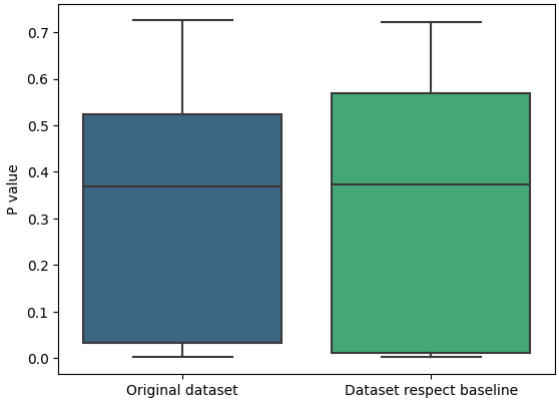


Figure 4.17: Variables along experiment 3: Discrete variable validation via the Kruskal Wallis test. The boxplot shows the high p values obtained in the test, indicating the similarity between variables along sessions (Wednesday sessions only.)

statistical or visual analysis is not a reliable proof for similarity. The complexity of the problem resides on the lack of convention or strict standards in the set up.

It is convenient to mention here how high the similarity along sessions in Experiment 3 is. This is the experiment performed in the scope of this report (the previous experiments were already performed), and show a very high similarity on the sessions with more participant data. This proves the high level of reproducibility when performing the experiment under constant controlled conditions.

The nature of the biosignals is very variable, as it includes a lot of heterogeneity because it does not only depend on the conditions and circumstances in which the experiment was performed, but also on the physiological state of the person. Experiments were performed in different seasons of the year, different rooms, different hour of the day and different amount of participants present in every session, for instance. This amount of differentiating factors may have played a key role in the similarity of the values.

Another important factor to take into account is the lack of independent data sources in the project. There are 28 participants (26 if we exclude the last team due to error in data acquisition), which may not be a population big enough to do this type of similarity analysis.

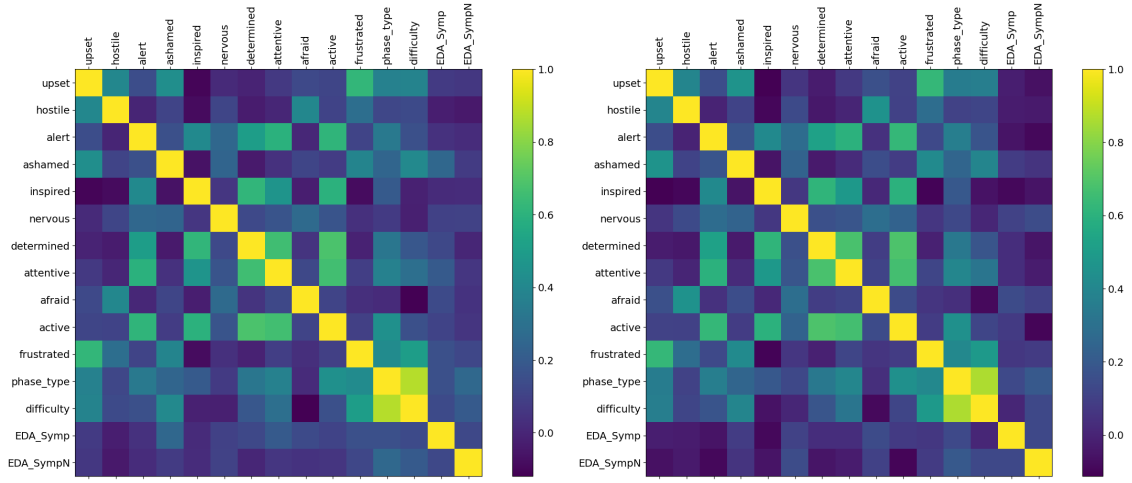
Furthermore, as we have mentioned in previous chapters, the features are constrained to time and frequency domain, without performing a feature analysis comparison of sparse features. This sparse features tend to find further insights via decomposition of the signals, focusing more on small-time could eliminate the heterogeneity that a whole signal feature implies.

4.5 Feelings - Biosignal coupling

This section analyzes how the different parameters extracted from the biosignals are related to the feelings reported by the participants.

4.5.1 Feelings - EDA

This section shows the correlation between the feelings answered in the participants survey and EDA (without decomposing) features.



(a) Original correlation

(b) Correlation respect to baseline

Figure 4.18: EDA - Feelings correlation tables.

No apparent correlation between the Sympatethic activity and the feelings. The Sympatethic activity is supposed to be a good cognitive stress indicator, but none of the feelings show significant correlation with the feature.

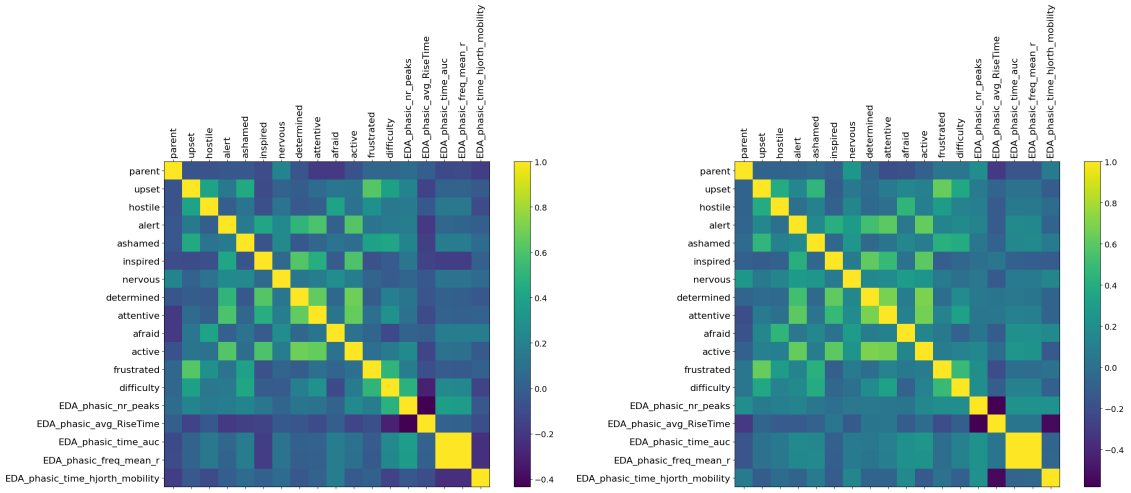
4.5.2 Feelings - EDA phasic

This section shows the correlation between the feelings answered in the participants survey and Phasic EDA features.

Phasic EDA (SCR) features show certain negative correlation of the number of peaks and the average recovery time with the difficulty level in the original dataset (without comparing with the baseline). The rest of SCR values do not show big correlation.

When comparing the correlation respect to the baseline, we can observe that the correlation of the average recovery time with the feelings is very noticeable, as most of the correlation values are very negatively correlated with them. The SCR recovery and rise time have a very high level of inter-individual and intra-individual variability [46]. The SCR is related to the sweat levels in the dermis, as it influences the skin conductance until it is reabsorbed or diffused away [47]. The reason behind this is the strong link between sweat secretion and a set of regulatory processes involved [48]. Other factor that drive this are the hydration level of the person.

In regard to the number of peaks, this value is considered as a good input of emotional arousal level[48]. This can be seen in the correlation matrix, where the phasic number of peaks is correlated with the phase type (rest vs task) and with the difficulty perceived by the experiment participants.



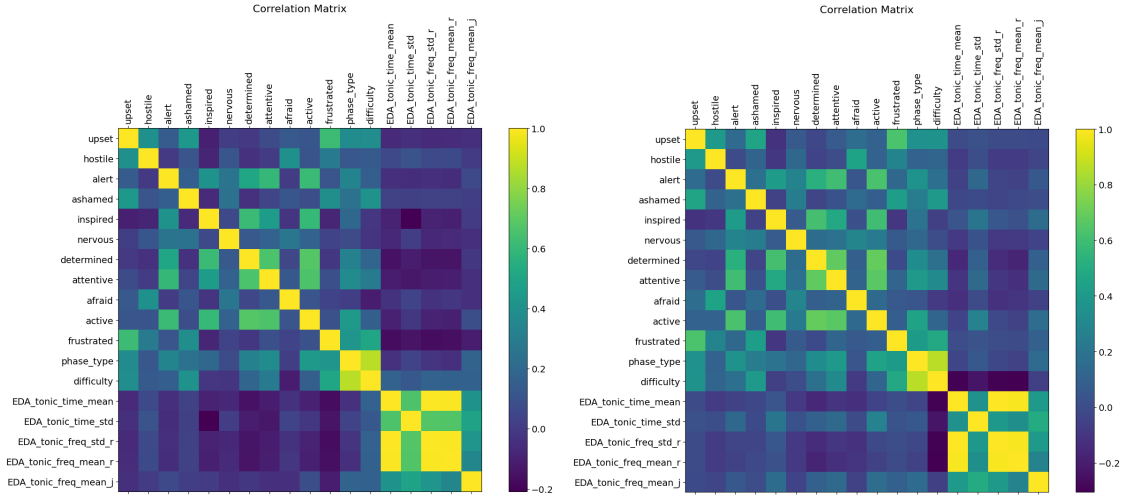
(a) Original correlation (b) Correlation respect to baseline

Figure 4.19: Phasic EDA - Feelings correlation tables.

Hence, we can see that these two variables act as good indicators of the current state of the person, as shown by the previous research and the correlation matrixes .

4.5.3 Feelings - EDA tonic

This section shows the correlation between the feelings answered in the participants survey and Tonic EDA features.



(a) Original correlation (b) Correlation respect to baseline

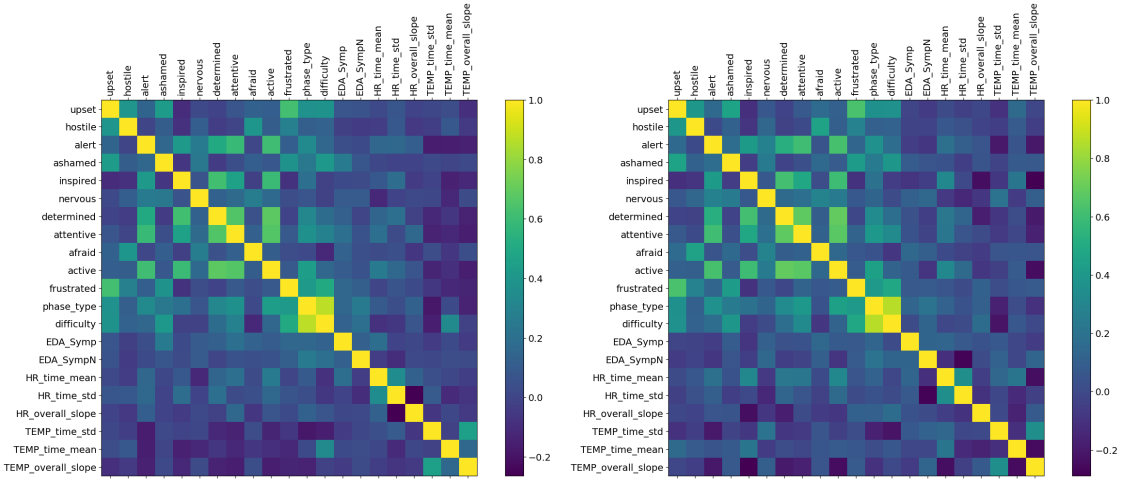
Figure 4.20: Tonic EDA - Feelings correlation tables. The tonic variables are correlated between each other, but no correlation is found with feelings.

The original dataset does not show a high level of correlation, with most of the EDA variables being negligibly correlated with the feelings. The introduction of the normalized dataset in respect to the baseline shows higher correlation levels, however these are still not significant.

All of the variables of the Tonic EDA are strongly correlated between each other, showing bright colors in the lower right area of the matrix.

4.5.4 Feelings - HR & Temp

This section shows the correlation between the feelings answered in the participants' survey and HR and Temperature features. No strong correlation is shown in any of the HR or temperature with the feelings, except for the mean temperature values with the difficulty perceived by the participants.



(a) Original correlation

(b) Correlation respect to baseline

Figure 4.21: Temperature and HR - Feelings correlation tables. No apparent correlation between HR and feelings are found, as expected from previous works [49].

From a preliminary point, some correlation between the HR and feelings was expected, as it is a natural idea to think that these 2 elements are correlated. This relationship was also pointed in papers described in the literature review [19] [20] However, there are other papers [49], which state that the heart rate is rarely related to emotional state, with only a minority of subjects showing association between these values in experiments. This is aligned with the results shown in Figure 4.21.

4.6 Participants Synchrony

The participants' synchrony section analyzes the level of biosignal cross correlation (CC) during the whole recording along the different phases and rounds. It is important to mention that the synchrony is calculated using a 30 seconds rolling average for the variables, smoothing the curve and pointing towards the understanding of the trend. The only biosignal which has a 5 seconds rolling average is the EDA Phasic, due to its short time nature (usual rise and recovery time are noticeably shorter than 30 seconds).

Every biosignal is analyzed individually in their respective sections, and all of them include 2 different analysis:

- **Cross Correlation distribution analysis**: The CC is analyzed measuring its distribution of values (between -1 and 1) in different phases. These phases correspond to resting in phase 1 and activity in phase 2. Phase 3 is not included as it is more a recovery phase after the puzzle than an actual resting phase. In order to analyze if the synchrony between participants exists for the analyzed variable, the CC of

2 participants who are not teamed up together (also referred to as non pairs) are also analyzed. This non-teamed graphs allow us to see visually how the CC is for non-teamed up participants behave versus actual teams. These non-teamed up participants were present at the same time in the same experiment sessions, and maybe they were subjected to similar stimuli at similar times.

- **Time Lagged Cross Correlation analysis:** The TLCC, described in Section 3.3.4, allows understanding how the instantaneous CC changes if a lag is applied. This activity introduces the idea that some signals may be lagged respect to others, as the peak in Cross Correlation may occur introducing a lag in the time series. There are 2 variables obtained and displayed in this analysis: the time offset at which the CC peak occurs vs the instantaneous; and the magnitude of difference in absolute cross correlation between the instantaneous and the peak CC. The time window for the TLCC analyzed ins 30 seconds before and after the instantaneous value.

4.6.1 HR - Heart Rate

The first variable to be analyzed is the Heart Rate.

HR: CrossCorrelation distribution

Figure 4.22 shows the cross correlation of the HR values along phases for team-up and non-teamed up participants. The most noticeable thing observed in the graph is its particular shape with 2 maxima found around -0.6 and +0.6, and a minimum located around the null cross correlation. This idea has two implications:

- **Low density in null correlation:** The density plot suggests that crosscorrelation values tend to be different from zero, either in the positive or negative direction. This lower amount of values in the null cross correlation area implies that there is always a certain synchrony or correlation in the trends of the participant biosignals.
- **Bimodal behavior:** The shape indicates that the CC values reach 2 maximas, showing a bimodal behavior. The negative CC implies that, the values of one of the participant follow the opposite trend to his/her teammate. For example, the parent could be trying to explain how to do the puzzle to the child, but the child may not understand clearly the instructions and would get some emotion arousal transformed into its cardiac frequency. On the opposite way, a positive correlation leads to the idea that both members of the team heart rate tend to increase/decrease together.

It is observed that the CC values of the participants who are teamed-up follow a similar curve in the positive area of Cross Correlation, while the values in the negative area differ. In CC=0, the density associated to Phase 2 is lower than Phase 1. This insight implies that the overall level of correlation (either positive or negative), is higher when the participants are doing an activity together than when they are resting. The absolute level of cross correlation in phase 2 was expected to be higher than 1, and this graph shows this higher level.

The comparison between pairs and non pairs show that the CC values for the teamed up members tend to be more positive, while the non pairs density is more shifted to the left-hand side, where the negative values are. However, the absolute values are very similar. This is a surprising piece of information, as 2 people with apparent no interaction between them have similar absolute levels of Cross Correlation in the HR measurements. It must be stated, though, that the non-pairs were located in the same room and performing the experiment at the same time, hence a certain level of correlation was expected due to environmental similarity.

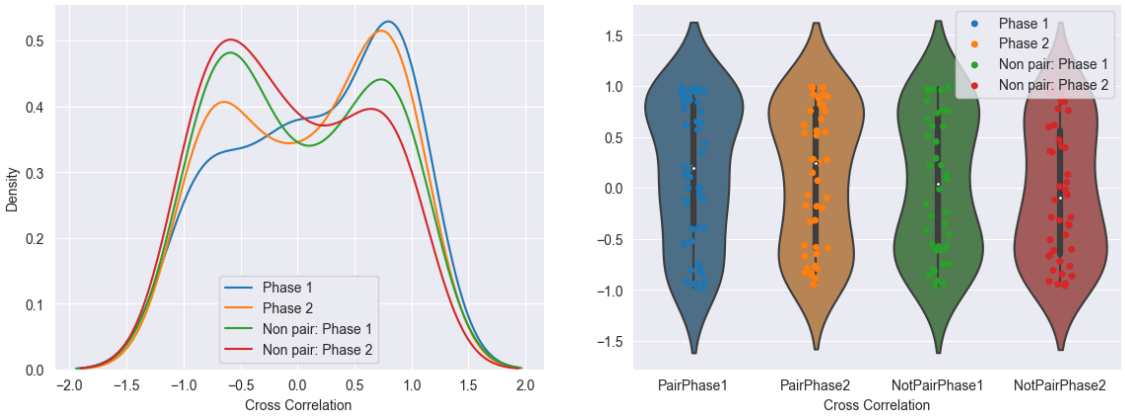
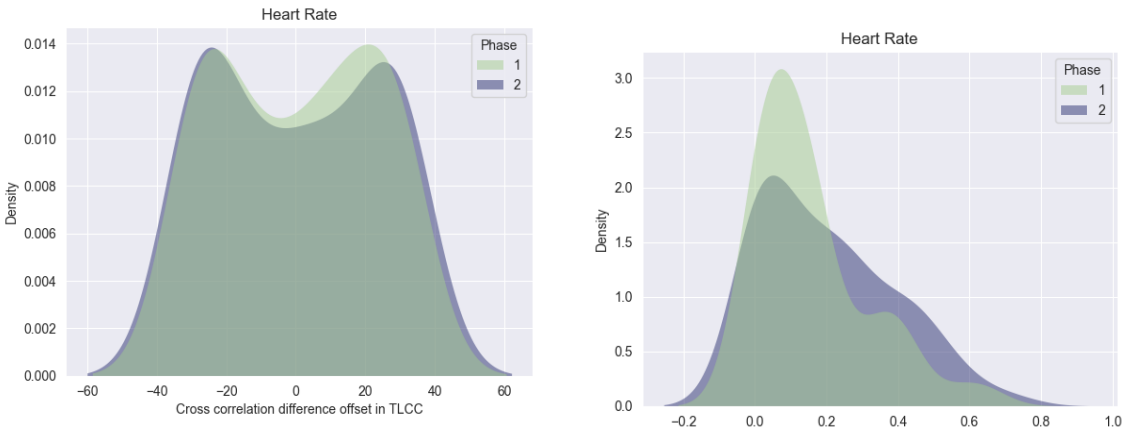


Figure 4.22: Heart rate cross correlation distribution along phases. Real pairs have larger cross correlation levels located on the positive side of the spectrum, while there is a trend to drop the density near the null correlation area.

HR: Time lagged CrossCorrelation

The TLCC analyzes the different Cross Correlation values when lags are applied.



(a) HR TLCC difference offset with 30 seconds window

(b) HR TLCC Cross Correlation difference

Figure 4.23: Heart Rate TLCC analysis: Distribution by phases of Cross Correlation absolute differences (instantaneous vs absolute in the time window) with their respective time offsets

From Figure 4.23, we can see that::

- Similar offsets distribution along phases (a): The two density curves that show the offset of the HR along phases are similar, meaning the cross correlation peak is not dependent on the phase (resting vs performing the activity together). The offsets peak in values very close to the time window limit (30 seconds), hence the absolute cross correlation may be outside the time window.
- The absolute difference in TLCC is lower in Phase 1 (b): The density of Phase 1 is greater in the area with a very low difference in CC. This suggests that during phase 1, the different between instantaneous cross correlation and the maximum

cross correlation in the time window are similar. The time window corresponds to 30 seconds before and after the instantaneous measure.

These 2 plots allow us to understand that even if the offset distribution is similar, the relative absolute difference is small in Phase 1. This implies that the maximum correlation level in Phase 1 is not very different from the one analyzed instantaneously. The opposite reasoning applied to the phase 2, where a higher correlation difference is observed, suggesting the idea that the instantaneous correlation is far (in magnitude) from the correlation peak.

4.6.2 TEMP - Temperature

The next variable to be analyzed is the Temperature.

TEMP: CrossCorrelation distribution

Figure 4.24 shows the cross correlation of the HR values along phases for team-up and non-teamed up participants. Once again, we can observe the bimodal shape seen in the heart rate.

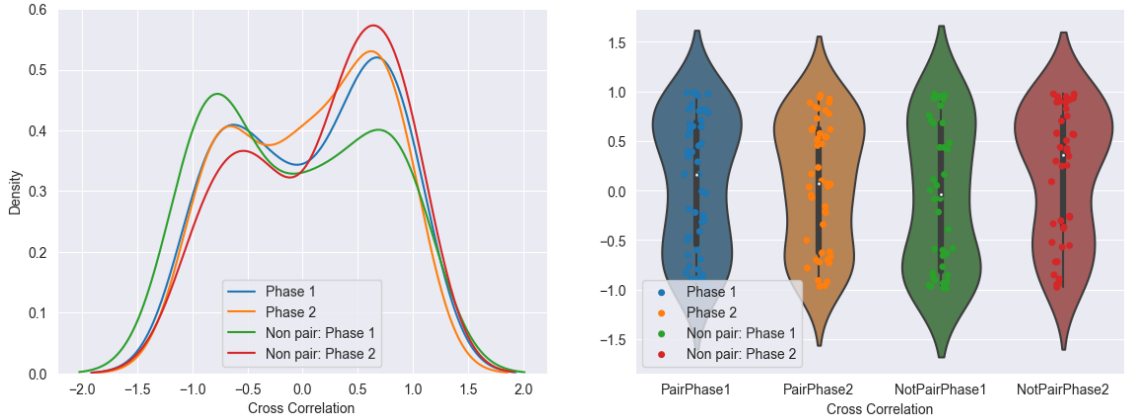


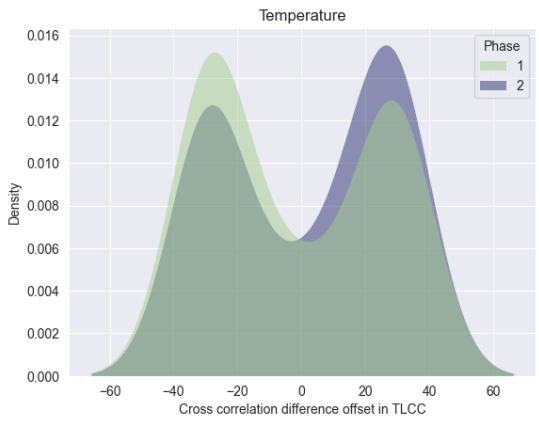
Figure 4.24: Temperature cross correlation distribution along phases. The temperature provides opposing behavior to the HR, with lower null CC (near 0) in Phase 1. Overall, results are similar among groups and temperature is not an optimal indicator of synchrony.

Temperature shows similar curves for both phase 1 (blue) and 2 (orange) for real paired participants CC. This similarity was also found on the previous section analyzing the HR. A big density peak in values around 0.6 are found for both of them, and the same curve is also appreciated in the negative side of the curve. The values in the area of null cross correlation present that the density in Phase 2 is more substantial than in Phase 1. This higher Phase 2 value manifests more null correlation between temperature teammates during the puzzle phase, and more significant synchrony when not performing the activity together (but resting together). This difference is not huge, but it is still necessary to be mentioned.

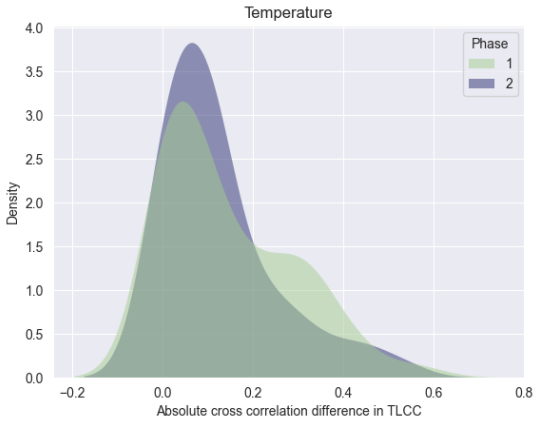
The curves between real and fake teammates also have the same shape, and all the curves have a lower density near the null CC area than the Phase 2 (orange line). This is surprising, as a bigger biosignal correlation was expected in Phase 2 between teammates. Overall, results are somehow similar among groups and temperature is not an optimal indicator of synchrony.

TEMP: Time lagged CrossCorrelation

The TLCC analyzes the different Cross Correlation values when lags are applied for the temperature signal.



(a) Temperature TLCC difference offset with 30 seconds window



(b) Temperature TLCC Cross Correlation difference

Figure 4.25: Temperature TLCC analysis: Distribution by phases of Cross Correlation absolute differences (instantaneous vs absolute in the time window) with their respective time offsets. Lower CC difference is observed in the Phase 2, with the absolute CC values in the time window not far from the instantaneous CC.

Plots from Figure 4.25 (a) show a similar distribution in offsets for the TLCC, with the offset tending to be more positive in the Phase 2 than in the Phase 1. This does not have a lot of implications, as it is more important the shape of the plots than the direction in which it is shifted. The reason behind this is that this plot helps to understand the usual offset of values, which proves to be far from 0 (instantaneous correlation). This may be explained by the slow-change nature of the temperature, which may require minutes until trend changes.

The absolute cross correlation values shown in Figure 4.25 (b) show that the absolute Phase 2 (blue) cross correlation difference is distributed mainly in the range between 0.0 and 0.2 , while the values for the Phase 1 are present until 0.4. This fact suggests that the shown levels of instantaneous Cross Correlation are not far from the maximum level in the time window, and instantaneous synchrony can be considered closed to the maximum.

4.6.3 EDA Tonic - SCL

The next variable to be analyzed is the EDA Tonic, also referred as SCL. The EDA tonic has a very similar shape to the EDA, with the difference that the EDA contains certain peaks which correspond to the EDA Phasic values (SCR).

EDA Tonic: CrossCorrelation distribution

Figure 4.26 shows the cross correlation of the HR values along phases for team-up and non-teamed up participants. Tonic EDA shows very different density curves along phases for the pairs (in the same team). Looking at the blue curve which indicates Phase 1, shows that the majority of the CC values are located at a positive correlation around 0.6, with the lowest density in the null cross correlation around 0. This implies that the correlation of tonic EDA between participants is much higher in resting phases than in the puzzle activity, which is seen in the orange curve that represents Phase 2.This Phase 2 curve has a similar cross correlation density around 0 (no correlation) than near 1, implying that there are similar cases where there is no correlation than those who actually have.

The reason behind this unexpected behavior may be related to the information that tonic

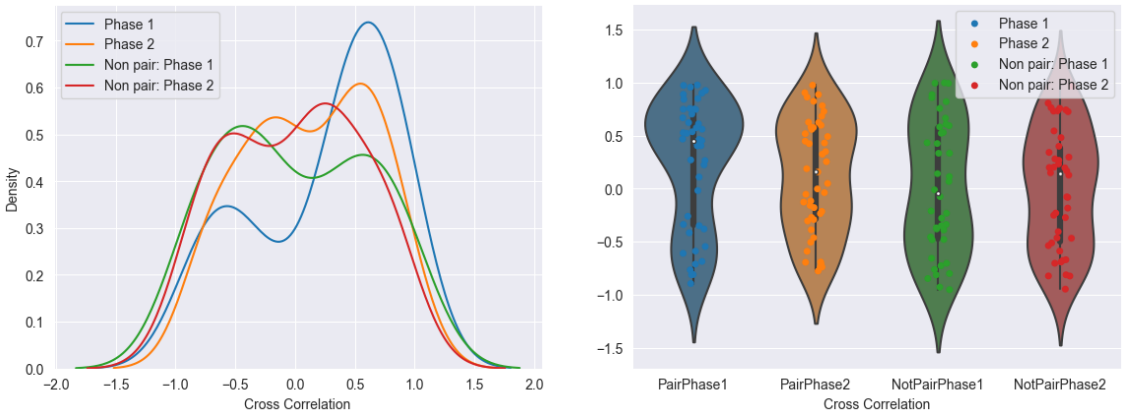
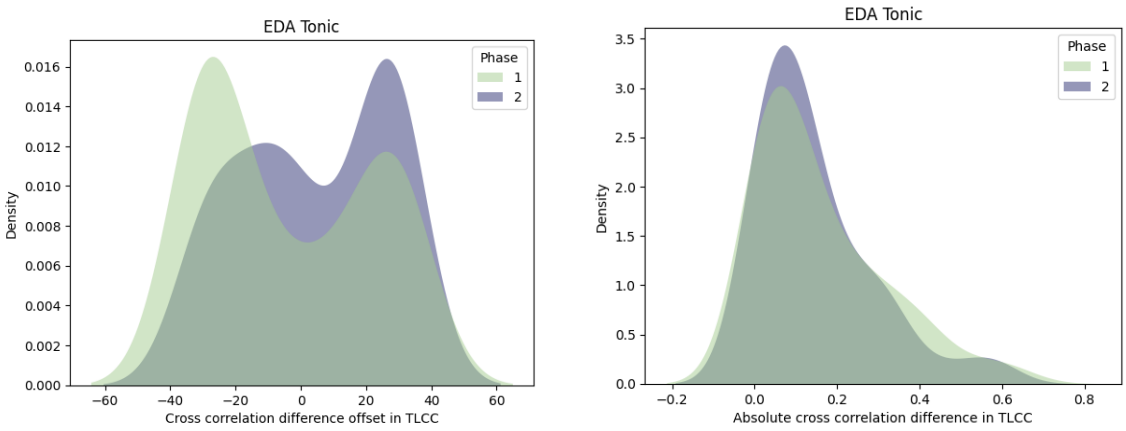


Figure 4.26: Tonic EDA cross correlation distribution along phases. Phase 1, represented in blue, shows low null CC and higher peak of positive CC, showing common pattern in lower level of emotional arousal in rest phases.

EDA provides. The EDA (not only the phasic EDA)is a good indicator of emotional arousal of the subject. Hence, in a resting phase after a collaboration activity between participant, it is likely that the level of excitation of the participants tend to lower and show similar patterns.

For this specific biosignal, there is a more visible difference between the real and non-real pairs in terms of synchrony. It is observable that the positive CC values are more present in the real pairs, and that the curve of Non Pairs Phase (2) is similar to a normal distribution centered in 0 (null cross correlation).

EDA Tonic: Time lagged CrossCorrelation



(a) Tonic EDA TLCC difference offset with 30 seconds window

(b) Tonic EDA TLCC Cross Correlation difference

Figure 4.27: Tonic EDA (SCL) TLCC analysis: Distribution by phases of Cross Correlation absolute differences (instantaneous vs absolute in the time window) with their respective time offsets.

SCL is the first biosignal analyzed which show different distributions of Cross Correlation offsets along phases, as seen from Figure 4.27 (a). The offset in Phase 2 (blue) has a higher density in the area near 0, where the cross correlation happens closer to the

instantaneous synchrony. This implies that overall, the maximum CC value is close to the instantaneous time. This is a very interesting insight, as it suggests that this variable can be used as a measure for the instantaneous synchrony in certain conditions when doing an activity instead of resting.

From Figure 4.27 (b), the absolute CC difference is also slightly more concentrated around 0 for Phase 1, like the one seen in the Temperature. This is another insight that suggest that Phase 2 values instantaneous synchrony represent similar values to the maximum synchrony in the time window. This is valuable to assume that, measuring the instantaneous EDA Tonic cross correlation, we have a good understanding of the overall correlation of the signal.

4.6.4 EDA Phasic - SCR

The last variable to be analyzed is the phasic EDA also known SCR, which corresponds to peaks in the skin conductance for a small interval of time. It is very important to remark that this signal is usually stable at 0, rising and decreasing for some seconds after stimuli. Another element to mention is the use of a 5 seconds rolling aggregate in this signal, in contrast with the 30 seconds period used in the other signals.

EDA Phasic: CrossCorrelation distribution

Figure 4.28 show the non lagged Cross Correlation of the SCR. The most noticeable thing from this figure is the density function, which follows a Normal Distribution alike shape over 0. This is an expected value, as the SCR is a signal which is usually flat at 0 most of the time. Hence, when calculating the Cross Correlation along the whole signal, a big % of the time the signals are flat, leading to a CC of 0.

In addition, there are other 2 insights to analyze, which provide a lot of information:

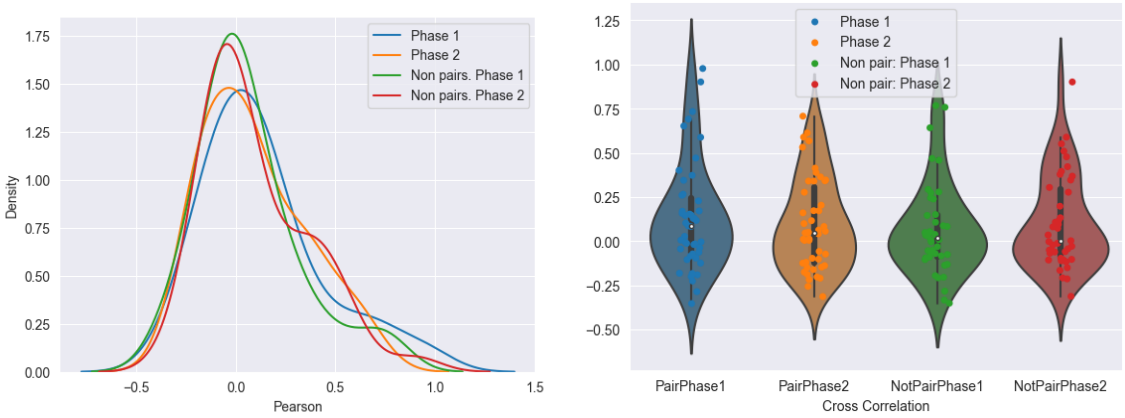


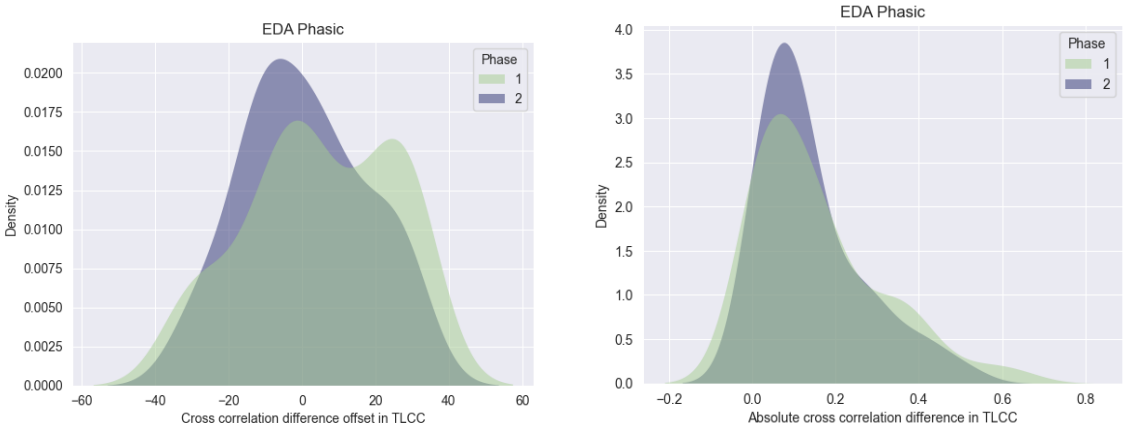
Figure 4.28: Phasic EDA cross correlation distribution along phases. The SCR has a lower CC density of real couples, represented by lower peak of blue and orange lines, implying higher synchrony of real couples under stimuli.

- **Lower density on CC = 0 for real couples:** The height of the distribution of both phases for real couples is lower, implying there is more CC different from zero captured. This is a key insight that proves that under the same stimuli, the level of signal synchrony of real couples is higher than in non-couples.
- **Similar behavior of phases in the right tail of the distribution:** Phase 2 of both the pair (orange) and not pair (red) seem to follow a different tail than the others.

This could be due to a higher number of activations of the SCR as the participants are more subjected to stimuli in the 2nd phase of the experiment.

To sum up, the importance of the information provided by the SCR is vital to consider this variable as a key indicator of biosignal synchrony.

EDA Phasic: Time lagged CrossCorrelation
Like in the previous signals, the last analysis corresponds to the TLCC.



(a) Phasic EDA TLCC difference offset with 30 seconds window

(b) Phasic EDA TLCC Cross Correlation difference

Figure 4.29: Phasic EDA (SCL) TLCC analysis: Distribution by phases of Cross Correlation absolute differences (instantaneous vs absolute in the time window) with their respective time offsets.

Figure 4.29 show the TLCC plots obtained. There are different shapes of the offset distribution (a) in respect to other variables (HR. TEMP and SCL). This graph provides valuable information, showing that the offset is peak is around 0. This idea indicates that, the maximum correlation usually occurs without introducing a lag in the analysis. This proves that the synchrony in this signal is very valuable, as most of the information transfer (correlation) occurs without significant delays. The values in the right and left from zero can be considered noise due to bad timer alignment (some people may have pushed the event button slightly earlier or later).

In Figure 4.29 (b), once again we can observe how the density plot is higher around 0 for Phase 2, showing that the instantaneous CC is not different in magnitude from the absolute of the time window.

4.6.5 Feelings synchrony

This section evaluates how synchronized the answers of participants are in after the resting or the task phases.

The evaluation of the synchrony in terms of feelings is performed getting the difference of feeling in every phase and round between participants. In other words, the feeling level answered by the participant is subtracted to the one answered at the same time by its teammate . This difference in feelings is considered as an asynchrony, as divergence in answers suggest emotional states that differ. Only the magnitude of this difference is used, hence the absolute value is taken.

As the mean value of feelings is different among feelings, the coefficient of variation is

used. This coefficient is the ratio of the standard deviation to the mean ($\sigma \backslash \mu$), and provides information about the variability in relation to the mean of the population.

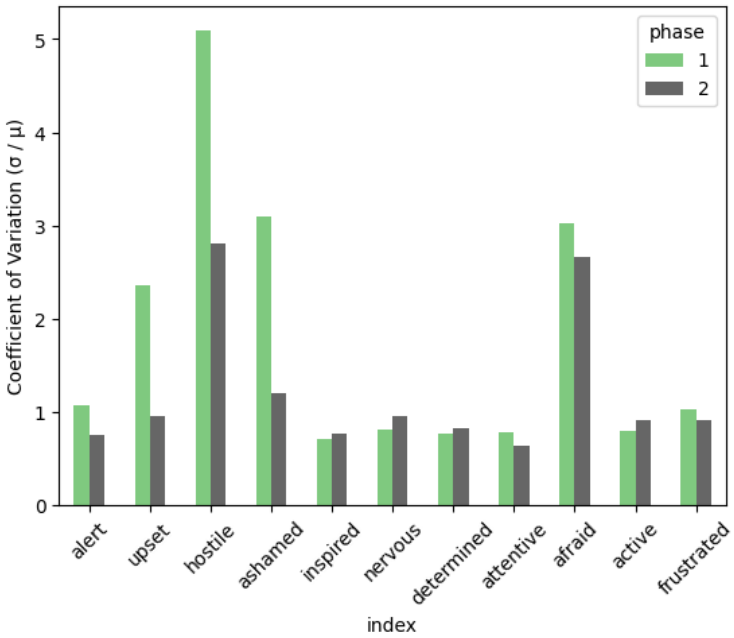


Figure 4.30: Coefficient of variation (CV) of feelings difference between pairs. Low values indicate high level of synchrony and similarity between pairs (low dispersion). The feeling of synchrony is consistently higher in the task phase (gray).

Figure 4.30 shows the coefficient of variation (CV), displaying a lower dispersion during the task phase (phase 2), which is represented in gray. This difference of feelings is noticeable in the alert, upset, hostile and ashamed answers. The rest of variables is very similar along phases.

This suggests that the feeling level shows a bigger level of synchrony after performing a task together than during the resting phase. This result suggests that not only the unbiased sources of information (biosignals time series and its features) are valid synchrony measures, but also the more subjective variables.

A variable to mention with a low Coefficient of Variation (CV) is the frustration level difference between the teammates. Even though in Figure 4.30 one of the roles had a significantly lower frustration level than the other, we can see that in reality, the dispersion is low. This low dispersion shows a common emotional state between participants, with a very similar level after rest and task phases.

4.6.6 Synchrony summary

After the individual analysis of the participants biosignals synchrony, we can summarize that:

- There exists a bimodal behavior in correlation, showing a multimodal distribution with 2 peaks in the density plots, with a valley between them located around null correlation (CC=0), except in the SCR.
- The HR signal shows higher correlation level during active phase (2), while temperature does it in phase (1). This HR correlation insight is in line with some of the

ideas obtained in the literature review [13] and [14]. This is a valuable input related to the general research question.

- The EDA Tonic provides valuable information in terms of correlation for resting phase, showing extremely low values of null CC.
- The EDA Phasic is a great indicator of synchrony, as their activation is made by the same stimuli for teammates, showing clear higher levels of synchrony than those who are not. This insight helps to respond to the 3rd research goal of the project, as this biosignal can be used as a reliable metric for synchrony among participants.
- The instantaneous CC (without lags) is a very good indicator of the maximum synchrony in Phase 2.
- Overall, the synchrony shown via Cross Correlation between participants in the same team is higher than those who are not paired together.
- The synchrony of feelings reported by the participants is higher after the task phase than in the rest phase for a specific set of feelings (alert, upset, hostile and ashamed)

4.7 Predictive Applications

This section analyzes the viability of biosignals use in predictive applications. Two variables are estimated: the role of the participant and the frustration level of the participant.

The analysis comprises the performance of a Random Forest Classifier Model. The RFC is set up with 1000 estimators, Gini criterion for the quality of split and bootstrap used.

4 datasets are used: Feelings + biosignal features of the participant; Only biosignal features of the participant; Feelings + biosignal features of both the participant and the task partner; only biosignal features of the participant and task partner.

The dataset size is comprised by around 300 observations. This number of observations could produce certain overfitting due to the low amount of data.

4.7.1 Role identification

As aforementioned, there are 2 roles during the experiments: the parent (instructor) and the child (puzzle maker). The analysis consists in the training of a model which classified the participant as a parent or child, according to the inputs stated before.

Results

Table 4.9 shows the performance of the model for the different inputs according to the 4 selected metrics.

Metric	Only data of participant		Only data of partner		Data of participant + partner	
	Feelings + Features	Features	Feelings + Features	Features	Feelings + Features	Features
Accuracy	74%	72%	84%	78%	87%	83%
Precision	63%	62%	89%	81%	90%	86%
Recall	87%	82%	77%	72%	81%	79%
F1	73%	70%	82%	76%	85%	82%

Table 4.9: Model performance for participant role prediction scores by different metrics. Datasets with a higher number of input variables (feelings + features and participant + partner) leads to better model performance.

From Table 4.9, we can observe that:

- **Feelings information is valuable:** Even though that questionnaires answers are not an objective answer as they are driven by the perception of the participant, we can see that the inclusion of feelings improves the performance.
- **Partner information improves significantly the role prediction.** The model offers a better performance when data from the partner is included. This is a major insight as it supports the idea that there exists synchrony among participants.

The confusion matrix of the best performing model is shown in Figure 4.31. The confusion matrix show an evenly distributed values for the 2 classes, with 49 parent roles and 47 child roles.

The performance of the model is similar for both categories, with a larger tendency to classify the case as parent than of child. The amount of wrongly classified instances is around 10-15 %, which is a low amount, showing high accuracy.

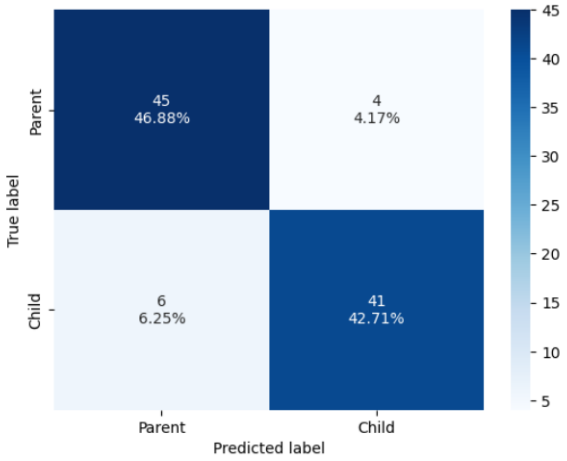


Figure 4.31: Confusion matrix in role prediction. The model is able to classify properly in a well-balanced class dataset, with low amount of false classified values.

Feature Importance

This section analyzes the most important features in the prediction of the participant role. The feature importance is performed via permutation, as described in Section 3.6.4.

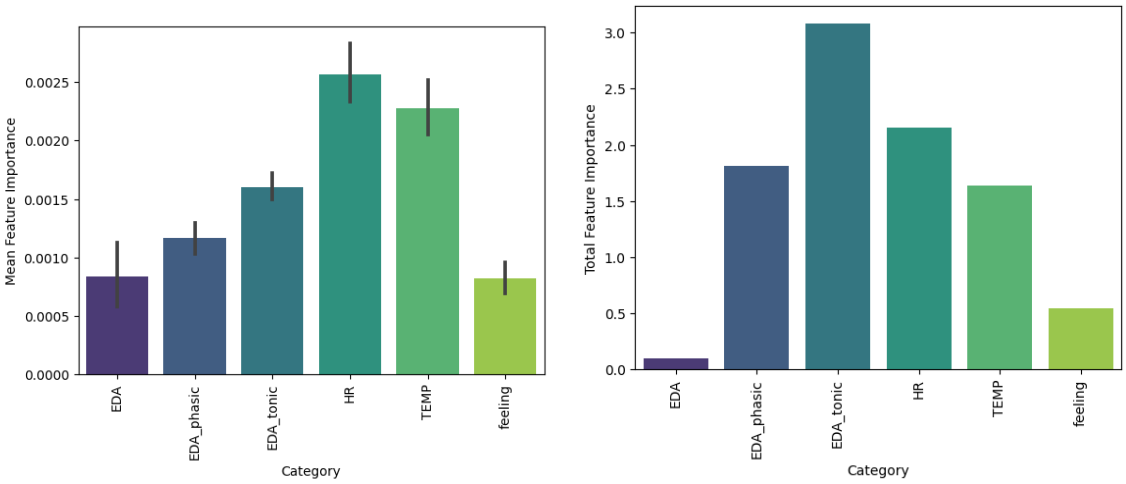
The Random Forest Classifier, as its own name suggests, has a random component which makes the reproducibility of results difficult. Hence, if run only multiple times the feature component analysis, the variables analyzed could have different feature importance. Therefore, in order to avoid this potential error, the permutations are performed 50 times and the importance of every variable is analyzed 20 times. This allow us to understand what are the main variables in the model.

For simplicity, the feature importance is performed in the dataset formed by the data of both participant + partner, which includes both the biosignals features and the feelings. This dataset contains more information than the others, then the feature importance analysis makes more sense to be applied.

Importance	Metric	Source	Importance	Metric	Source
#1	EDA Tonic Std Variation	Partner	**#6**	Temperature gradient	Participant
#2	Temperature gradient	Partner	**#7**	EDA phasic frequency	Partner
#3	Temperature minimum slope	Participant	**#8**	HR minimum value	Participant
#4	HR mean	Participant	**#9**	EDA phasic Hjorth Complexity	Partner
#5	Nervous	Participant	**#10**	Attentive	Partner

Table 4.10: Variables with the most importance in the model for the participant role prediction (4 categories). The feature importance variables are obtained after 50 runs for every variable analyzed.

Table 4.10 shows the top 10 variables for the model that takes Feelings + Features of the participant and partner data. The metric has the source associated to it displayed, if it comes from the actual participant or from the partner of the participant. The top 2 values come from the partner of the participant, a surprising insight as it is the top contributor to



(a) Mean feature importance by category

(b) Total feature importance by category

Figure 4.32: Feature importance value by variable category. The values correspond to the model that takes Feelings + Features of the participant and partner data.

the model was expected to come from the actual participant.

From Figure 4.32, we can see the average importance of the variables of each category to the model. The category with more importance is the EDA variable without the decomposition into phasic and tonic.

The number of variables per category in the model are different. More specifically, we can find 32 Tonic EDA, 26 phasic EDA, 14 Heart Rate, 12 Temperature, 11 feelings and 2 EDA values. This information complements the data of Figure 4.32, which show the total and mean importance of the features.

4.7.2 Frustration level

For an easier classification, the frustration level is categorized in 4 categories:

- Very low: Frustration levels between 0 and 1
- Low: Frustration levels between 2 and 4
- Medium: Frustration levels between 5 and 7
- High: Frustration levels between 8 and 10

The precision, recall and f1 are calculated as the weighted average of the individual values.

Table 4.11 shows the results of the predictive model performance for the frustration level. We can extract that:

- Like in the role analysis, the use of feelings information is valuable to
- The partner data is not very valuable in the prediction of the frustration level of the participant.

The confusion matrix displayed in 4.33 show a very unbalanced distribution of classes, with the biggest majority of values on categories *Very Low* and *Low*. The model tends to predict only in the 2 lowest categories due to this unbalancing (which is also present

Metric	Only data of participant		Only data of partner		Data of participant + partner	
	Feelings + Features	Features	Feelings + Features	Features	Feelings + Features	Features
Accuracy	59%	55%	53%	45%	59%	56%
Precision	53%	52%	46%	40%	53%	50%
Recall	60%	55%	53%	45%	59%	56%
F1	55%	49%	48%	41%	55%	53%

Table 4.11: Model performance for participant role prediction. As in the role prediction, the use of feelings + features improves the model performance. However, the addition of partner data does not represent an improvement.

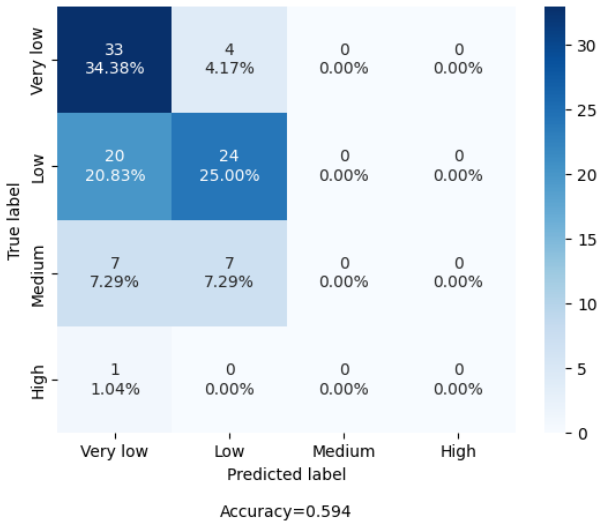


Figure 4.33: Frustration level confusion matrix. The results shown in this matrix come from the model which takes only data of participant and feelings and features.

in the training dataset), with no values predicted in Medium and High, even though there were 14 and 1 value in those categories respectively.

Figure 4.36 shows the uneven categories distribution which led to the results mentioned before. Respondents answered mainly low level of frustration,l which makes sense as this survey was answered after the rest phases too, which usually are associated to lower frustration level.

Balancing 4 frustration categories

Due to the unbalanced classification of previous section, a new set of categories is presented. This categories have around a 25% of distribution of values per category, and are formed by: frustration = 0; frustration = 1; frustration = 2; and frustration bigger or equal to 3.

As seen in Figure 4.35, the categories are well balanced numerically, but the category of ”3 or higher” is composed by a big set of different answers ranging from 3 to 10.

These categories however do not make real sense, as categories 0, 1 and 2 are very similar and are not answered according to common standards among participants. This

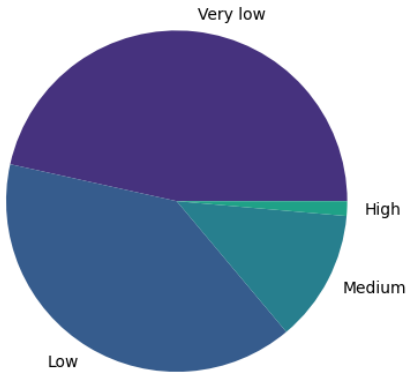


Figure 4.34: Frustration categories distribution in the whole dataset. The very low (0-1) and low (2-4) levels of frustration in a 0-10 scale are dominant.

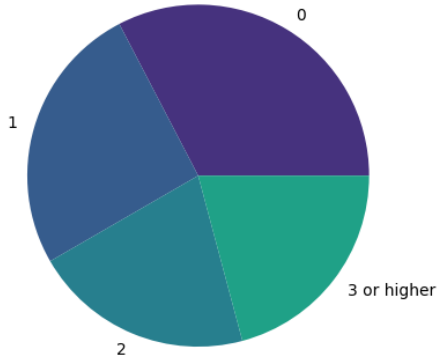


Figure 4.35: Frustration categories distribution in the whole dataset. The very low (0-1) and low (2-4) levels of frustration in a 0-10 scale are dominant.

high subjective component in the answer leads to hgih interchangeability of answers in this category

The confusion matrix gives us a hint on how the classification algorithm works, as it is concentrating the classification in the 2 extremes, either in category 0 or in the above 2 level. The performance in the intermediate levels, such as frustration level of 1 or 2, is very poor, and the values in these categories are categorized in the highest and lowest categories.

Thus, this algorithm may perform better when applying a lower number of balanced categories, which will be discussed in the next paragraph.

Reducing frustration categories

The classification is reduced to a binary classification, with the first category (55% of sample) consisting in very low level of frustration (0 and 1) while the second category is the rest of values (45% of the sample, composed by answers from 2 to 10).

The result of reducing to a 2 balanced classification is very positive, as observed in Figure

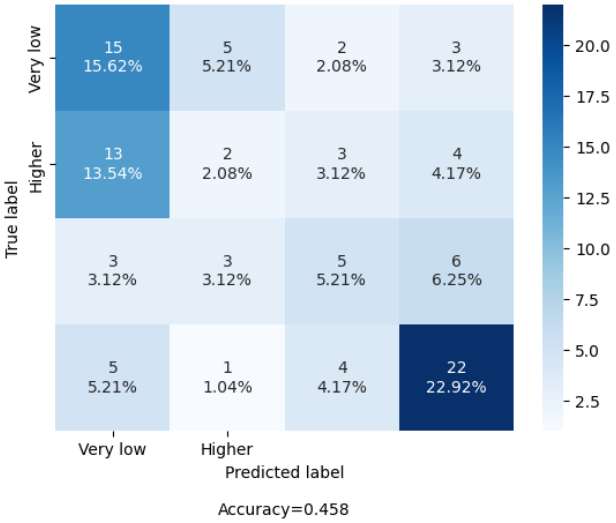


Figure 4.36: Confusion matrix in balanced 4 frustration categories classification. The algorithm tends to classify the values in the extreme categories.

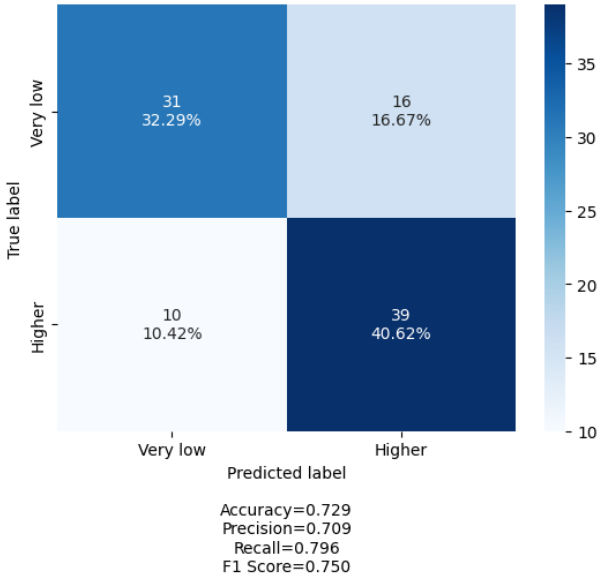


Figure 4.37: Confusion matrix in balanced 2 frustration category classification. The reduction of categories and the better balancing helps the model performance to classify the participant frustration level prediction.

4.37, with all the metrics scoring above 0.7. The proper balancing together with the reduction of categories boosts the classification performance, with most of the values classified correctly. This successful classification is a valuable piece of information to extract, as it suggests that subjective data, such as frustration feeling, can be estimated at a certain level.

The algorithm has certain tendency to classify the instances in the Higher category, and it could be due to the big diversity of frustration levels that hosts. Anyhow, this tendency is not extremely concerning.

4.8 Analysis of Project Definition Report PDR

This last section of chapter 4 deals with the execution of the project in respect to the original PDR. The most updated PDR is added to Appendix

4.8.1 Deviations to the PDR

This chapter addresses the deviations and modifications done in the PDR from the original document submitted in October.

Research question

The first research question was modified to improve the wording. The original question was *Can we correlate collaboration levels with biosignals synchronization?* while the new one is *Can we detect collaboration levels from the synchronization biosignals?.*

The first part of the second question was modified for better understandability of the purpose of the data generation. The original question was *Can we generate reliable and useful data for this purpose?* while the new one is *Can we generate reliable and useful biosignal data for the synchrony analysis?*

The original question number 3 was modified to not take into account the collaboration levels into consideration. The original question was *Which variables show the biggest levels of synchrony among people who collaborate in a given task?* while the new one is *Which variables show the biggest levels of synchrony among participants?*.

The collaboration analysis is still found in Question 1, as it states *"Can we correlate collaboration levels with biosignals synchronization?"* . The evaluation of the question is adressed in Section 5.2.

Data access and use

The original PDR opened the possibility to the use of real OCD patients for analysis, apart from the data of non-patients from experiments. Finally, due to the restrictive nature of the data and the lack of access to it, this source was discarded and not used in the thesis project.

Deliverables

In the deliverables section, there were 2 milestones associated with the generation of a report. These deliverables were:

- Experiment report: set up and analysis on how the experiment was performed.
- Biosignals synchronization identification: small report stating the relevant variables found to have synchronization on the original dataset (Experiments 1 and 2).

These 2 reports were not done as their information was directly added into the Thesis document. There would be a duplicity of content and of effort, so to avoid this, all the information was added in the corresponded chapters of the document.

4.8.2 Risk Analysis identification and impact

The risks identified in the PDR are shown in Table 4.12. No edits are performed in the table, and it is shown as in the original PDR in order to evaluate how well the risks were assessed before the project.

The stated risks are analyzed in the bullet points below, once the project is concluded and we can assess the risks. The risks are names by their associated number from Table 4.12.

- **R1:** The original PDR was accurate enough to scope appropriately the objectives, except for the change in the 3rd research questioned mentioned in 4.8.2. The risk

#	Title	Description	Risk	Week
R_1	Scoping	Project Description did not scope appropriately the project objectives.	4	40
R_2	Existing data quality	In the processing of the existing data of the project, the quality and extension of it may not be adequate. This risk is reduced performing a new experiment for new data acquisition	3	42
R_3	Experiment data acquisition	This step is aimed to increase the dataset and reduce risks associated to the previous point. However, a poor data acquisition may trigger wrong thesis conclusions. Hence, the process must be properly set up and documented.	5	44
R_4	Data merging issues	Existing and new data may not show similar patterns. This could trigger the doubt on which dataset must be used. This is why a proper data acquisition process must be carried out, thus assuming that new data will be the valid one.	5	46
R_5	Timeline issues at hand in	Some activities may carry more time than planned in the Gantt Chart. However, this risk is mitigated as the Thesis Hand In is scheduled 2 weeks before the actual limit for the hand in.	2	6

Table 4.12: Risks Analysis from the PDR. 1:low risk, 5:high risk of delays

was mitigated after agreeing with the thesis supervisor to rephrase the learning objective and stating it in the deviation chapter.

- **R2:** The risk of low amount of data is present along the whole project, specially for the Machine Learning applications. This risk is reduced generating more controlled date via the 3rd experiment.
- **R3:** The set-up of the experiment was intended to be as controlled and as standardized as possible to ensure a good data quality. The defective datasets generated during the experiment were discarded to ensure it would not interfere with the rest of datasets.
- **R4:** The different datasets arouse the questions of how similar the data would be. The validation of the similarity of the different datasets are evaluated in a visual and statistical way, providing reasoning on the possible explanations for the divergences. Risk was found to be lower than expected (it was marked as high risk of delay)
- **R5:** The Gantt chart was not followed strictly in terms of starting and ending week, however the order of activities was accurate and no big delays were experienced in the different tasks. The thesis writing tasks

5 Conclusion

5.1 Research goal

Can we identify synchrony in biosignals of people performing a collaborative task with a wearable device?

The overall research goal has been achieved, with positive results in the synchrony identification on participants of the experiment. The satisfactory experiment sessions run in the scope of the project and the data validation and acquisition, together with the dataset expansion, have enabled to understand how biosignals behave in the experiment. Most of the biosignals studies, such as HR and EDA show some synchrony insights, which are reflected in shared information and correlation between teammates. The formulation and development of the ML models have been fed by these datasets has been successful. These model have shown a reliable performance in the predictive classification tasks, with most of the scores over 70%.

More information on every individual research goal is found in the next sections.

5.2 Research question 1

Can we detect collaboration levels from the synchronization biosignals?

Defining collaboration levels has been a challenge in this work, as the only potential measure of active collaboration in the experiment is the number of Tangram puzzles finished in the 5 minutes period of the activity. This measure, however, is not exclusively representative of collaboration among participants, as it depends also on other factors. These factors include:

- the difficulty of the Tangram puzzle selected
- proper assignation of parent-child role
- prior knowledge of the other team member
- previous experience doing puzzles, tangrams or collaborative experiments
- Environment conditions of the experiment: number of people in the room, temperature, room layout and size.

These results go in line with a similar study studying the CSCL [18], where the lack of an adequate definition of collaboration metrics is stated. This same study also analyzed the difficulty perception as a team variable, while the experiment conducted as part of this work , analyzes it dependent exclusively on the individual answers.

A high variability of successfully done tangrams is present, with some teams not achieving any successful results, while other teams doing up to 6 tangrams in the experiment. Overall, the whole experiment process served as a learning collaborative experience for the participants. This collaborative experience led to synchrony in their biosignals in the experiments, concurring with the affective states' convergence stated by Duffy et al. [15].

Also, the current measure of collaboration can be assumed to be a performance metric rather than a collaboration metric. In previous studies, performance in collaboration tasks was not associated with physiological synchrony [50]. Hence, the number of tangrams

completed may not be an adequate metric, and consequently has not been actively applied in the research.

The design and execution of the experiment carried out lacks of a reliable way to measure collaboration level. This issue needs to be addressed to formulate an objective and unbiased collaboration measure. Some ideas that could be implemented in the future could be:

- the time extension of phase 2, so participants have longer time to complete the Tangram, allowing more data to be acquired in the process.
- assignation of points to the specific tangram according to the level of difficulty.
- standardizing the condition of participants, i.e: use of same language, similar level of acquaintance with the teammate.
- Adding a new response from the participants stating the collaboration level perceived.

5.3 Research question 2

Can we generate reliable and useful biosignal data for the synchrony analysis? Is data consistent along the different experiments?

The data generated in the experiments is reliable, except for the signals discarded (ACC and BVP). The issues with the data generation can be overcome with the new Empatica device Embrace Plus [1], which offers new sensors. These sensors include a clinically-validated optical PPG for the BVP values and a new accelerometer which include a Gyroscope. Not only that, but the rest of sensors are also improved.

The data is not fully consistent along experiments, as some variables seem to differ substantially from experiments. An example of this difference is the EDA Tonic values, which follow a different behavior in the 2nd experiment, maybe caused by the experiment conditions (atmospheric conditions influence the skin transpiration, which affects the EDA recordings). However, given the nature of the experiment, whose data come from human tasks and answers (big level of variability expected), the results are satisfactory according to the statistical tests shown in Section 4.4.2. The similarity of biosignal features between datasets ranges between 37% and 52% (T-test), and the similarity between feelings answered in the questionnaire ranges between 34 % and 83 % (Mann-Whitney U test), as seen in the discussion of Section 4.4.2.

However, as from Figures 4.16 and 4.17 it has been observed that the data generated and acquired in the scope of this report (experiment 3), is consistent and similar along sessions. This statement needs to be taken into account for the 2 sessions with more data (4 participants in the analyzed sessions vs 2 participants in the others). In consequence, the discussion must include that the data exclusively dependent on the author's report set up is consistent, establishing proof of similarity.

The set-up, execution, data acquisition and data processing of the experiment have been successful. The new data is added to the dataset of the previous experiments and will be used for further research.

Further experiments with more consistent environment conditions (like in the sessions of experiment 3) should be performed in the future for better results, as stated in the validation summary from Section 4.4.4.

[1] https://www.empatica.com/en-eu/embraceplus/

5.4 Research question 3

Which variables show the highest levels of synchrony among participants?

From the variables analyzed, certain synchrony patterns have appeared. We can find that variables of people doing an activity together tend to have higher level of synchrony (cross correlation) than those who are not paired together.

In the Heart Rate, the participants experienced marginally higher level of absolute cross correlation level when performing the task together than in rest phase. This phenomenon is inverted in the temperature analysis, where participants have a higher synchrony level in rest phase. The two signals with EDA as origin, phasic and tonic, provide useful information of synchrony among participants. More specifically, the SCR, or EDA phasic, is a good indicator of synchrony as the peaks in this signal appear when a person is subjected to a stimulus. Therefore, two subjects collaborating in a task are usually subjected to that stimuli, showing levels of synchrony.

There could exist a lag in correlation among participants, but usually the top values of correlation are identified with instantaneous cross correlation (no lag induced). This suggests that the highest synchrony level is in the moment that the participants are collaborating.

A last element to be mentioned is the subjective information that the participants provide. The subjective feelings answered in the questionnaires have a certain level of correlation with some of the biosignals, providing valuable information to understand the synchrony among these participants. Furthermore, these feelings answers show lower dispersion in pairs when they perform a task together rather than when they are in resting phase.

A limitation in this research goal is the lack of use of sparse domain feature extraction in the analysis. The use of SPCA may be a key driver for automatic event detection, such as synchrony, using long-term observational signals [51].

Nevertheless, this limitation is mainly understood as a possible future addition to the project, and does not tarnish the satisfactory results related to the synchrony identification with biosignals.

5.5 Research question 4

Can we generate predictive models with this data?

The data generated can be used to set up and develop ML algorithms. The classification models of the target variables (participant role and frustration level) showed a satisfactory performance, with accuracy and F1 score values above 70% for both the role prediction and the frustration level.

Due to a majority of participants answers being placed in the very low or low spectrum of frustration, the distribution of the variable is not balanced. Due to this unbalancing in frustration, we needed to modify the multi-category classification to a binary classification, improving the results. This enhancement of performance is seen for the scores change from Table 4.11 to the ones in Figure 4.37.

The introduction of teammate data into the predictive model is beneficial, as observed from 2 different analysis:

- Higher score in metrics, shown in Table 4.9
- Feature importance showed teammate variables in the top variables with influence in the role prediction, as from Table 4.10.

A big limitation of the ML application in this project is the amount of data available. The scarcity of data affects the quality of results, as low amounts of data usually lead to overfitting, together with the fact that the data come from multiple observations from the same source. Participants carried out 12 phases, with their correspondent questionnaires and biosignals measures.

This problem was already known and stated in the Risk section of the project report. In the PDR, it was stated that the data acquisition is a time-consuming process due to all the requirements and assets necessary to perform the experiment. The risk was mitigated with the set-up of the new experiment, which duplicated the amount of existent data. The risk mitigation was successful with the dataset extension, however this process should not stop here, as the project benefits from larger amounts of data.

The intention of this research question was not the development of a mature ML model to predict the attitude or emotional arousal of the Empatica E4 holder, but to prove and show the potential applications of these biosignals. The positive results obtained are in line with other projects published in the last 2-3 years, such as: bike accident recognition through biosignals[52]; classification of physical activity level with a wrist device [53] via EDA, with similarities to the work in this M.Sc. thesis project; further understanding and interpretability of models fed with biosignals [54]; antiseizure medication effects assessment analyzing EDA and HR via an Empatica device [55]; or other already cited projects where an Empatica device has been used [10] [11] [12].

Therefore, this specific research goal has been successfully proven, leading to the possibility for advanced future models. These future models will contain, more data, bigger variety of signals and a higher level of complexity, such as DL.

5.6 Future work and perspectives

The reflection from section 5.4 suggests that this project may benefit from the addition of more advanced techniques in the feature extraction process. Future work may include SPCA techniques which provides more complex features, derived from combinations of already analyzed biosignals and their features.

One of the tasks to be worked in the future for this project is the addition of the BVP as a signal for the different analysis performed. Even though the HR signal is synthesized from the BVP in the Empatica E4 device, the inclusion of the new signal may improve the quality of the studies performed and may induce better predictive applications. One example of the full potential of this addition is the work on stress detection algorithm via an DL network with accuracy results over 92% [56].

A frequently mentioned limitation along the document is the lack of a well established collaboration measure. The proposed measure was the number of tangrams completed, however this measure may not take into account the actual collaboration level, as this may be influenced by other non-observed factors.

Improvements in the experiment set up and instructions may be added, removing excessive resting times and substituting them for task phases where the collaboration level can be measured, via unbiased metric (team success) or biased (answers in a questionnaire addressing specifically the feeling of collaboration). These improvements should also provide more specific information on the set-up, such as exact number of participants in a session, physical layout and environment conditions of the experiment, or a guideline on the teammate assignation and distribution.

Furthermore, a joint questionnaire of the team participants could be added to the experiment set up, as collaboration process is a common phenomenon and joint perception should be taken into account

Bibliography

[1] Leonard G Wilson. “ERASISTRATUS, GALEN, AND” THE PNEUMA””. In: *Bulletin of the History of Medicine* 33.4 (1959), pp. 293–314.

[2] Dr. Marc Barton. *WILLEM EINTHOVEN AND THE ELECTROCARDIOGRAM*. 2017. URL: https://www.pastmedicalhistory.co.uk/willem-einthoven-and-the-electrocardiogram/ (visited on 09/30/2022).

[3] BW Johansson. “A history of the electrocardiogram”. In: *Dansk Medicinhistorisk Arbog* (2001), pp. 163–176.

[4] S Serge Barold. “Willem Einthoven and the birth of clinical electrocardiography a hundred years ago”. In: *Cardiac electrophysiology review* 7.1 (2003), pp. 99–104.

[5] Fifth Edition et al. “Diagnostic and statistical manual of mental disorders”. In: *Am Psychiatric Assoc* 21.21 (2013), pp. 591–643.

[6] Alexandra I Korda et al. “Recognition of blinks activity patterns during stress conditions using cnn and markovian analysis”. In: *Signals* 2.1 (2021), pp. 55–71.

[7] Md Fahim Rizwan et al. “Design of a biosignal based stress detection system using machine learning techniques”. In: *2019 international conference on robotics, electrical and signal processing techniques (ICREST)*. IEEE. 2019, pp. 364–368.

[8] Dukyong Yoon et al. “Discovering hidden information in biosignals from patients using artificial intelligence”. In: *Korean Journal of Anesthesiology* 73.4 (2020), pp. 275–284.

[9] Sridhar Krishnan and Yashodhan Athavale. “Trends in biomedical signal feature extraction”. In: *Biomedical Signal Processing and Control* 43 (2018), pp. 41–63.

[10] Emilia Grzesiak et al. “Assessment of the feasibility of using noninvasive wearable biometric monitoring sensors to detect influenza and the common cold before symptom onset”. In: *JAMA network open* 4.9 (2021), e2128534–e2128534.

[11] Gregory P Strauss et al. “Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia”. In: *Schizophrenia* 8.1 (2022), p. 37.

[12] Roberto Sánchez-Reolid et al. “One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity”. In: *Biomedical Signal Processing and Control* 71 (2022), p. 103203.

[13] Lauri Ahonen et al. “Biosignals reflect pair-dynamics in collaborative work: EDA and ECG study of pair-programming in a classroom environment”. In: *Scientific reports* 8.1 (2018), pp. 1–16.

[14] Cathy Mengying Fang et al. “Cardiac Arrest: Evaluating the Role of Biosignals in Gameplay Strategies and Players’ Physiological Synchrony in Social Deception Games”. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–7.

[15] Melissa C Duffy et al. “Team regulation in a simulated medical emergency: An in-depth analysis of cognitive, metacognitive, and affective processes”. In: *Instructional Science* 43 (2015), pp. 401–426.

[16] Richard V Palumbo et al. “Interpersonal autonomic physiology: A systematic review of the literature”. In: *Personality and Social Psychology Review* 21.2 (2017), pp. 99–141.

[17] Carl Marci and Helen Riess. “The clinical relevance of psychophysiology: Support for the psychobiology of empathy and psychodynamic process”. In: *American journal of psychotherapy* 59.3 (2005), pp. 213–226.

[18] Jonna Malmberg et al. “Are we together or not? The temporal interplay of monitoring, physiological arousal and physiological synchrony during a collaborative exam”. In: *International Journal of Computer-Supported Collaborative Learning* 14 (2019), pp. 467–490.
[19] Bradley M Appelhans and Linda J Luecken. “Heart rate variability as an index of regulated emotional responding”. In: *Review of general psychology* 10.3 (2006), pp. 229–240.
[20] Yan Wu et al. “How do amusement, anger and fear influence heart rate and heart rate variability?” In: *Frontiers in neuroscience* 13 (2019), p. 1131.
[21] Sylvia D Kreibig et al. “Cardiovascular, electrodermal, and respiratory response patterns to fear-and sadness-inducing films”. In: *Psychophysiology* 44.5 (2007), pp. 787–806.
[22] Gabriele Calabrò et al. “M-MS: A Multi-Modal Synchrony Dataset to Explore Dyadic Interaction in ASD”. In: *Progresses in Artificial Intelligence and Neural Systems*. Springer, 2021, pp. 543–553.
[23] F Behrens et al. “Physiological synchrony is associated with cooperative success in real-life interactions”. In: *Scientific reports* 10.1 (2020), pp. 1–9.
[24] Hugo Plácido Da Silva, Ana Fred, and Raúl Martins. “Biosignals for everyone”. In: *IEEE Pervasive Computing* 13.4 (2014), pp. 64–71.
[25] Akara Supratak et al. “Survey on feature extraction and applications of biosignals”. In: *Machine learning for health informatics*. Springer, 2016, pp. 161–182.
[26] Ozgur Dedehayir and Martin Steinert. “The hype cycle model: A review and future directions”. In: *Technological Forecasting and Social Change* 108 (2016), pp. 28–41.
[27] World Health Organization et al. *Global status report on noncommunicable diseases 2014*. WHO/NMH/NVI/15.1. World Health Organization, 2014.
[28] Mohamed Elgendi et al. “A six-step framework on biomedical signal analysis for tackling noncommunicable diseases: Current and future perspectives”. In: *JMIR Biomedical Engineering* 1.1 (2016), e6401.
[29] David E Bloom et al. *The global economic burden of noncommunicable diseases*. Tech. rep. Program on the Global Demography of Aging, 2012.
[30] MESUT Çiçek. “Wearable technologies and its future applications”. In: *International Journal of Electrical, Electronics and Data Communication* 3.4 (2015), pp. 45–50.
[31] Empatica. *E4 wristband*. 2020. URL: https://www.empatica.com/en-eu/research/e4/ (visited on 09/30/2022).
[32] Jason J Braithwaite et al. “A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments”. In: *Psychophysiology* 49.1 (2013), pp. 1017–1034.
[33] Erik Vavrinsky et al. “The Concept of Advanced Multi-Sensor Monitoring of Human Stress”. In: *Sensors* 21.10 (2021), p. 3499.
[34] Joon-Ho Choi and Vivian Loftness. “Investigation of human body skin temperatures as a bio-signal to indicate overall thermal sensations”. In: *Building and Environment* 58 (2012), pp. 258–269.
[35] Dominique Makowski. *Neurophysiological Data Analysis with NeuroKit2*. 2022. URL: https://neuropsychology.github.io/NeuroKit/ (visited on 09/30/2022).
[36] Tony Fischer-Cripps. *Newnes interfacing companion: computers, transducers, instrumentation and signal processing*. Elsevier, 2002.
[37] Hugo F Posada-Quintero. “Electrodermal Activity: What it can Contribute to the Assessment of the Autonomic Nervous System”. In: (2016).

[38] Bo Hjorth. “EEG analysis based on time domain properties”. In: *Electroencephalography and clinical neurophysiology* 29.3 (1970), pp. 306–310.
[39] D Devi, S Sophia, and SR Boselin Prabhu. “Deep learning-based cognitive state prediction analysis using brain wave signal”. In: *Cognitive Computing for Human-Robot Interaction*. Elsevier, 2021, pp. 69–84.
[40] Toni Giorgino. “Computing and visualizing dynamic time warping alignments in R: the dtw package”. In: *Journal of statistical Software* 31 (2009), pp. 1–24.
[41] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and statistical modeling with python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 57. 61. Austin, TX. 2010, pp. 10–25080.
[42] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* 17.3 (2020), pp. 261–272.
[43] T Van Hecke. “Power study of anova versus Kruskal-Wallis test”. In: *Journal of Statistics and Management Systems* 15.2-3 (2012), pp. 241–247.
[44] Nadim Nachar et al. “The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution”. In: *Tutorials in quantitative Methods for Psychology* 4.1 (2008), pp. 13–20.
[45] Hugo F Posada-Quintero et al. “Time-varying analysis of electrodermal activity during exercise”. In: *PloS one* 13.6 (2018), e0198328.
[46] Claude Breault and Raymond Ducharme. “Effect of intertrial intervals on recovery and amplitude of electrodermal reactions”. In: *International journal of psychophysiology* 14.1 (1993), pp. 75–80.
[47] Mathias Benedek and Christian Kaernbach. “Decomposition of skin conductance data by means of nonnegative deconvolution”. In: *psychophysiology* 47.4 (2010), pp. 647–658.
[48] Don C Fowles et al. “Publication recommendations for electrodermal measurements”. In: *Psychophysiology* 18.3 (1981), pp. 232–239.
[49] Derek W Johnston and Pavlos Anastasiades. “The relationship between heart rate and mood in real life”. In: *Journal of psychosomatic research* 34.1 (1990), pp. 21–27.
[50] Muhterem Dindar, Sanna Järvelä, and Eetu Haataja. “What does physiological synchrony reveal about metacognitive experiences and group performance?” In: *British Journal of Educational Technology* 51.5 (2020), pp. 1577–1596.
[51] Shengkun Xie and Sridhar Krishnan. “Model based sparse feature extraction for biomedical signal classification”. In: *International Journal* 6.1 (2017), p. 11.
[52] Joo Woo et al. “Wearable Airbag System for Real-Time Bicycle Rider Accident Recognition by Orthogonal Convolutional Neural Network (O-CNN) Model”. In: *Electronics* 10.12 (2021), p. 1423.
[53] Angelica Poli et al. “Cross-Domain Classification of Physical Activity Intensity: An EDA-Based Approach Validated by Wrist-Measured Acceleration and Physiological Data”. In: *Electronics* 10.17 (2021), p. 2159.
[54] Marília Barandas et al. “Uncertainty-based rejection in machine learning: Implications for model development and interpretability”. In: *Electronics* 11.3 (2022), p. 396.
[55] Mustafa Halimeh et al. “Wearable device assessments of antiseizure medication effects on diurnal patterns of electrodermal activity, heart rate, and heart rate variability”. In: *Epilepsy & Behavior* 129 (2022), p. 108635.
[56] Eda Eren and Tuğba Selcen Navruz. “Stress Detection with Deep Learning Using BVP and EDA Signals”. In: *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE. 2022, pp. 1–7.

A Appendix 1: Participant Questionnaire

Participant ID: _____ Team ID:_____ Biosensor ID.:_____ Round:_____

Pre-task (Phase 1)

Below are a list of different feelings and emotions. Please read each feeling and circle the number that best matches how much you feel each feeling right now. Circle 1 if you feel the feeling, ”very slightly or not at all”. Choose 5 if you feel the feeling ”extremely”.

	Very slightly or not at all	A little	Moderately	Quite a bit	Extremely
1. Upset	1	2	3	4	5
2. Hostile	1	2	3	4	5
3. Alert	1	2	3	4	5
4. Ashamed	1	2	3	4	5
5. Inspired	1	2	3	4	5
6. Nervous	1	2	3	4	5
7. Determined	1	2	3	4	5
8. Attentive	1	2	3	4	5
9. Afraid	1	2	3	4	5
10. Active	1	2	3	4	5

1. How frustrated are you feeling right now? (Circle the number on the scale below that best describes how you are feeling right now.)

Participant ID: _____ Team ID:_____ Biosensor ID.:_____ Round:_____

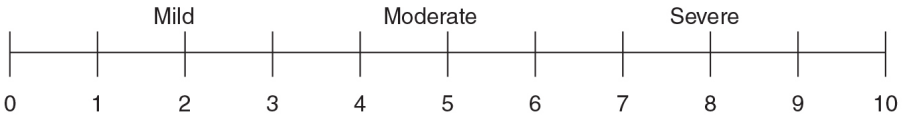
In-task (Phase 2)

Below are a list of different feelings and emotions. Please read each feeling and circle the number that best matches how much you feel each feeling right now. Circle 1 if you feel the feeling, ”very slightly or not at all”. Circle 5 if you feel the feeling ”extremely”.

	Very slightly or not at all	A little	Moderately	Quite a bit	Extremely
1. Upset	1	2	3	4	5
2. Hostile	1	2	3	4	5
3. Alert	1	2	3	4	5
4. Ashamed	1	2	3	4	5
5. Inspired	1	2	3	4	5
6. Nervous	1	2	3	4	5
7. Determined	1	2	3	4	5
8. Attentive	1	2	3	4	5
9. Afraid	1	2	3	4	5
10. Active	1	2	3	4	5

1. How frustrated are you feeling right now? (Circle the number on the scale below that best describes how you are feeling right now.)

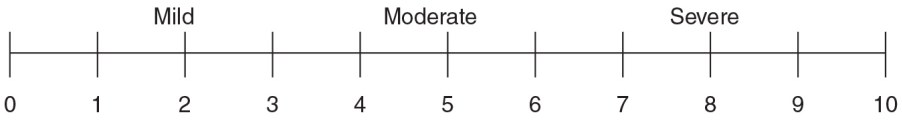
Not at all frustrated.



I could not be more frustrated.

2. How difficult did you find the task? (Circle the number on the scale below that best describes your experience.)

Not at all difficult.



Extremely difficult.

Participant ID: _____ Team ID:_____ Biosensor ID.:_____ Round:_____

Post-task (Phase 3)

Below are a list of different feelings and emotions. Please read each feeling and circle the number that best matches how much you feel each feeling right now. Circle 1 if you feel the feeling, ”very slightly or not at all”. Choose 5 if you feel the feeling ”extremely”.

	Very slightly or not at all	A little	Moderately	Quite a bit	Extremely
1. Upset	1	2	3	4	5
2. Hostile	1	2	3	4	5
3. Alert	1	2	3	4	5
4. Ashamed	1	2	3	4	5
5. Inspired	1	2	3	4	5
6. Nervous	1	2	3	4	5
7. Determined	1	2	3	4	5
8. Attentive	1	2	3	4	5
9. Afraid	1	2	3	4	5
10. Active	1	2	3	4	5

1. How frustrated are you feeling right now? (Circle the number on the scale below that best describes how you are feeling right now.)

Not at all frustrated.

I could not be more frustrated.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Technical
University of
Denmark

Building 101
2800 Kgs. Lyngby
Tlf. 4525 1700

www.dtu.dk