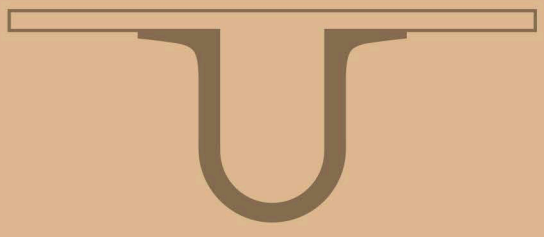
UNIVERSIDADE D
COIMBRA



Gonçalo Daniel Tavares Duarte

Cognitive and autonomic neural manifestations captured using wearable sensors for reliable software development

Thesis for the degree of Master in Biomedical Engineering,
Supervised by Prof. Dr. Henrique Madeira and Dr. Ricardo Couceiro,
and submitted to the Faculty of Sciencies and Technology, University of Coimbra

2019

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Gonçalo Daniel Tavares Duarte

Cognitive and autonomic neural manifestations captured using wearable sensors for reliable software development

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. Henrique Madeira (CISUC)
Dr. Ricardo Couceiro (CISUC)

Coimbra, 2019

This work was developed in collaboration with:

Centre for Informatics and Systems of the University of Coimbra

Coimbra Institute for Biomedical Imaging and Translational Research

Resumo

Neste trabalho são apresentados resultados experimentais que revelam que o esforço mental de programadores em tarefas de compreensão de código pode ser monitorizado através da variabilidade da frequência cardíaca usando dispositivos não invasivos. Os resultados sugerem que a variabilidade da frequência cardíaca é um bom preditor de carga cognitiva durante a análise de códigos e também que os resultados de variabilidade da frequência cardíaca são consistentes com o esforço mental percecionado pelos programadores. Além disso, as métricas de complexidade de código não se correlacionam devidamente com o esforço mental e não aparentam ser um bom indicador da perceção subjetiva de complexidade sentida pelos programadores. Este estudo é apresentado no contexto do projeto *BASE* - "Biofeedback Augmented Software Engineering" que propõe uma abordagem inovadora, introduzindo o "biofeedback" no desenvolvimento de software.

Palavras-chave: engenharia de software, variabilidade da frequência cardíaca, esforço mental, complexidade

Abstract

This work presents emergent experimental results showing that mental effort of programmers in code understanding tasks can be monitored through heart rate variability (HRV) using non-intrusive wearable devices. Results suggest that HRV is a good predictor for cognitive load when analysing code and HRV results are consistent with the mental effort perceived by programmers. Furthermore, code complexity metrics do not correlate entirely with mental effort and do not seem to be a good indicator of the subjective perception of the complexity felt by programmers. These first results are presented in the context of the project *BASE* - Biofeedback Augmented Software Engineering - that proposes a radical neuroscience enabled approach to introduce biofeedback in software development.

Keywords: software engineering, heart rate variability, mental effort, complexity

Contents

List of Figures

List of Tables

Glossary

ANS Autonomic nervous system.

CNS Central nervous system.

ECG Electrocardiogram.
EDA Electrodermal activity.
EEG Electroencephalogram.

FFT Fast Fourier transform.

HF Power in high frequency range.
HR Heart rate.
HRV Heart rate variablity.

ICG Impedance cardiography.

LDA Linear discriminant analysis.
LF Power in low frequency range.
LH Ratio of low to high frequency.

MMD Multiscale morphological derivative.

NN Interval between consecutive normal R peaks.
NN50 Number of consecutive NN intervals that differ more than 50 milliseconds.

PCA Principal component analysis.
pNN50 Ratio between NN50 and the total number of NN intervals.
PNS Peripheral nervous system.
PPG Photoplethysmogram.
PSD Power spectrum density.

rHF ratio of HF to the power in all frequency components.
rLF ratio of LF to the power in all frequency components.
RMSSD Square root of the mean squared differences between adjacent NN intervals.
RR Interval between consecutive R peaks.
RSME Rating scale of mental effort.

SD Standard deviation.
SDNN Standard deviation of the NN intervals.

SDSD Standard deviation of successive differences between adjacent NN intervals.
SVM Support-vector machine.

TVL power difference between VLF range and power in all frequency components.

VLF Power in very low frequency range.

1

Introduction

1.1 Context and motivation

The problem of residual software faults (bugs) has been heavily researched in the context of software engineering, but, unfortunately, a sound breakthrough on software reliability has not been reached. In fact, the software industry average for software bugs is about 15 to 50 errors per 1000 lines of delivered code [6]. There was even a report, in 2016, that claimed that 1.1 trillion dollars were lost worldwide on software bugs, glitches and security failures that year [7]. Since software bugs are human errors, it makes sense to investigate possible mechanisms that can provide feedback on programmers' mental stress and other conditions that may precipitate making bugs or bugs escaping programmers' attention. A solution based on physiologic responses could be a possible alternative considering that intellectual activities cause responses in the autonomic nervous system like heart rate variation [8] and pupil dilatation [9], for example. These responses could potentially be measured with non-invasive devices that are compatible with daily software development activities and be used to represent the complexity perceived by the programmer/tester on a certain task, something that the classic software complexity metrics fail to do. With this new approach, improvements in software development such as high-complexity code segment warnings and enhancement of the developing and testing experience with bug prediction models using biofeedback data would be possible. These are the main goals of the project *BASE*, *Biofeedback Augmented Software Engineering*, in which this thesis is inserted.

The main setback for the use of physiologic responses in software reliability is that their measurement is influenced by other physiological responses not associated with software development activities and can, therefore, be non-representative of the mental effort.

1.2 Objectives and planning

The objective of this thesis is to develop and evaluate a new algorithm for the assessment of autonomic nervous system activity manifestations in order to discriminate different levels of cognitive stress. This main goal can be subdivided into 6 main steps:

- Design of the experimental protocol for the collection of bio-signals;
- Development of the experimental set up;
- Recruitment of volunteers and conduction of the experiments;
- Creation of a database with the collected data;
- Development and evaluation of an algorithm for cognitive stress quantification based on heart rate variability analysis;
- Analyse the impact of participants' proficiency in programming and their perceived complexity of the different tasks on the results.

To achieve these an experimental protocol is defined where the recruited participants of medium to high proficiency programming skills use wearable sensors to monitor physiologic reactions while reading and trying to understand code of different complexity levels. Experiments take place in the *Institute for Biomedical Imaging and Life Sciences*, *IBILI*, a research Institution of the *Faculty of Medicine* from *University of Coimbra*. During the experiments, several bio-signals are collected regarding brain activity, electro-dermal activity, eye movement, skin blood flow and heart activity is recorded. Despite the diversity of the collected bio-signals, the current thesis only focuses on the analysis of the information regarding the heart rate variability. Different HRV features are used, both in time and frequency domain, for the study of the experienced cognitive effort of the subjects during the analysis of different code complexities.

This thesis was started in February 2018 and followed the following schedule displayed in Table 1.1.

Task	2018											2019	
	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb
State of the art addressing previous literature on software engineering and autonomic nervous system													
Experiment design and hypotheses definition													
Development of the experimental set up and preliminary experiments													
Recruiting and screening													
Experiments and data collection													
Dataset creation													
Feature extraction													
Definition and evaluation of models expressing code complexity and mental effort													
Study of the participants' programming proficiency and perceived complexity and their impact on the results													
Writing of MSc thesis document													

Table 1.1: Schedule of the milestones defined for this thesis.

1.3 Structure of the thesis

The document is structured as follows: Chapter 2 details the physiological context of this work related to autonomous nervous system and heart rate variability; Chapter 3 dissects the current state of art in software engineering regarding complexity metrics and cognitive effort assessment using HRV indexes; Chapter 4 describes how the participants were selected and the definition o the experimental set up and the protocol; Chapter 5 dissects the used methodology namely HRV calculation and feature extraction as well as the approaches made; Chapter 6 presents and discusses the results; Chapter 7 suggests possible applications of the findings and work that can be done in the future. Finally Chapter 8 outlines the main conclusions of the work and some problems that may justify certain results.

1.4 Contributions

A preliminary study related to the results that differentiated distinct levels of "mental effort" was submitted and accepted in the International Conference on Software Engineering, *ICSE 2019 Montreal*, Canada on May 25-31 2019. A oral presentation regarding these results will also be made in the IEEE 6th Portuguese Meeting in Bioengineering, *6th ENBENG*, organized by the Portuguese Chapter of IEEE Engineering in Medicine & Biology Society (*EMBS*), in Lisbon, on February 22-23, 2019.

Others studies regarding data collected in this work, namely EEG data and pupillography will also present their findings in the *6th ENBENG* meeting. The study related to pupillography was also submitted to *IEEE DSN 2019*.

2

Physiological context

2.1 The nervous system

2.1.1 Overview

The nervous system is one of the most complex body systems that has sensory, motor and integrative functions and is organized in two subdivisions: the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS is composed of the spinal cord and the brain and is responsible for processing different kinds of sensory information. All the nervous tissue outside the CNS, nerves, ganglia (small masses of brain tissue), enteric plexuses (networks of neurons located in the gastrointestinal tract) and sensory receptors, is part of the PNS. This system is also divided into the somatic, enteric and autonomic nervous system. The somatic nervous system mediates voluntary movement with motor neurons and contains sensory neurons that convey information from the head, body wall, limbs and from the 5 senses. The enteric nervous system controls the gastrointestinal system and its operation is involuntary. [3]

2.1.2 The Autonomic nervous system

The ANS consists of sensory neurons associated with autonomic sensory receptors located in visceral organs (large interior organ in any of the great body cavities) and motor neurons that send signals to smooth muscle, glands and cardiac muscle. The activity associated with the ANS is involuntary (not under conscious control). The motor component of the ANS is subdivided into the sympathetic and parasympathetic division, with mainly opposing actions. For example, heart rate increases with sympathetic activity while parasympathetic activity decreases it. In sum, the sympathetic component supports activity and emergency actions, also known as

"fight-or-flight" situations, and the parasympathetic component is in charge of the "rest-and-digest" activities. [3]

The ANS acts over many organs (Figure 2.1) and its activity can be seen in, for example, electrodermal activity influenced by the sweat glands, retinography and tracking of the eye movements, photoplethysmography (used to estimate the skin blood flow using infra-red light) due to the vasoconstrictions and vasodilatations as well skin temperature and different heart activity parameters measured through impedance cardiography (used to measure the electrical conductivity of the thorax) and electrocardiogram. [3]

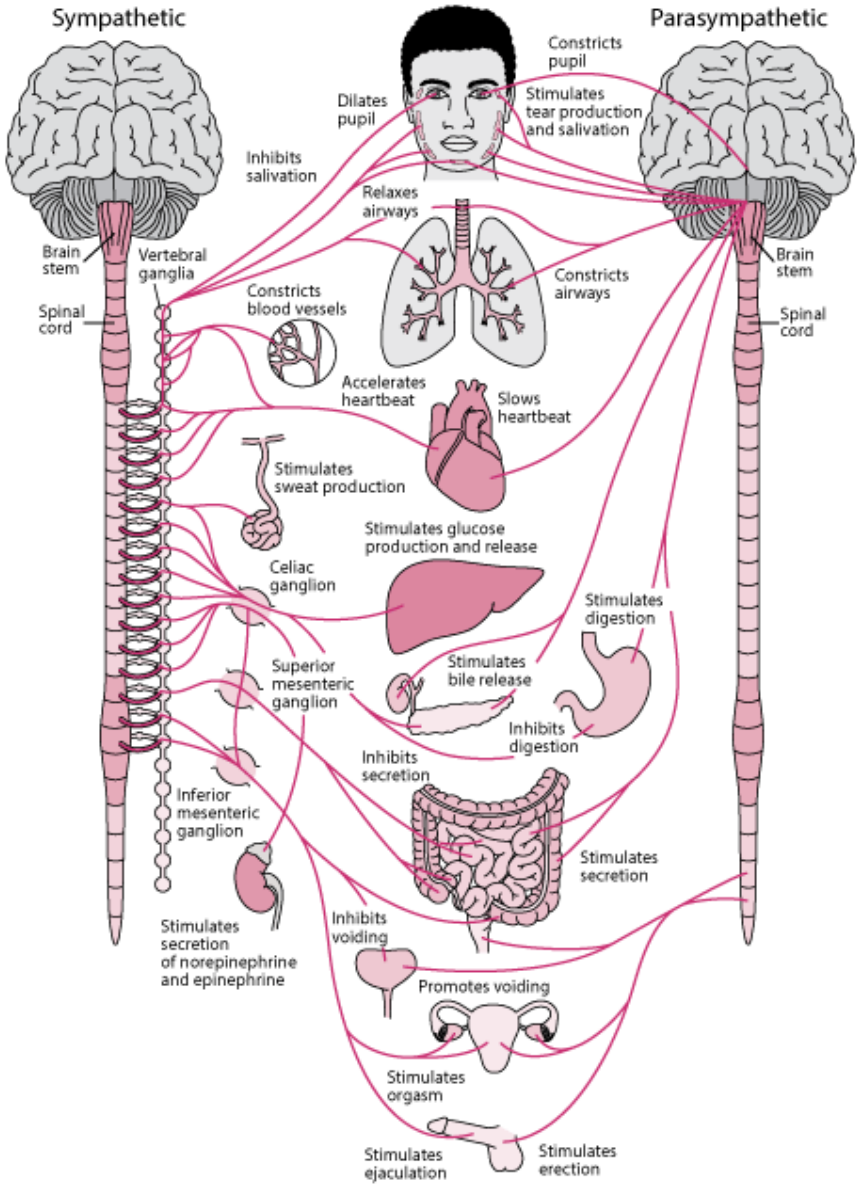


Figure 2.1: Representation of the ANS and different systems it operates on. Adapted from [1]

As previously mentioned, the main focus of this thesis is on the cardiovascular system, more specifically heart rate variability and its connection to mental effort and cognitive stress. This is explored later in this chapter and in chapter 3. The cardiovascular system supports a wide range of systems and, therefore, is regulated by a constant monitoring of the pressure in the arterial system. Tracking of this primarily mechanical information, relevant for the action of both parasympathetic and sympathetic systems, is conveyed by baroreceptors located in the heart and in major blood vessels. These are activated through their nerve endings with the expansion and contraction of the vessels walls. This afferent system sends its information via the vagus nerve to the solitary tract which communicates with the hypothalamus. With the raise of blood pressure, for example, baroreceptors are activated and sympathetic activity is inhibited while parasympathetic activity is stimulated mainly through the vagus nerve resulting in the reduction of innervation of the cardiac pacemaker and musculature and reduction in the heart rate. Associated to this phenomenon is the dilation of the peripheral arterioles that result in an overall decrease in blood pressure. In contrast to this sequence of events, a drop in blood pressure, has the opposite effect, inhibiting parasympathetic activity while increasing sympathetic activity. The baroreflex feedback system is represented in Figure 2.2. [10]

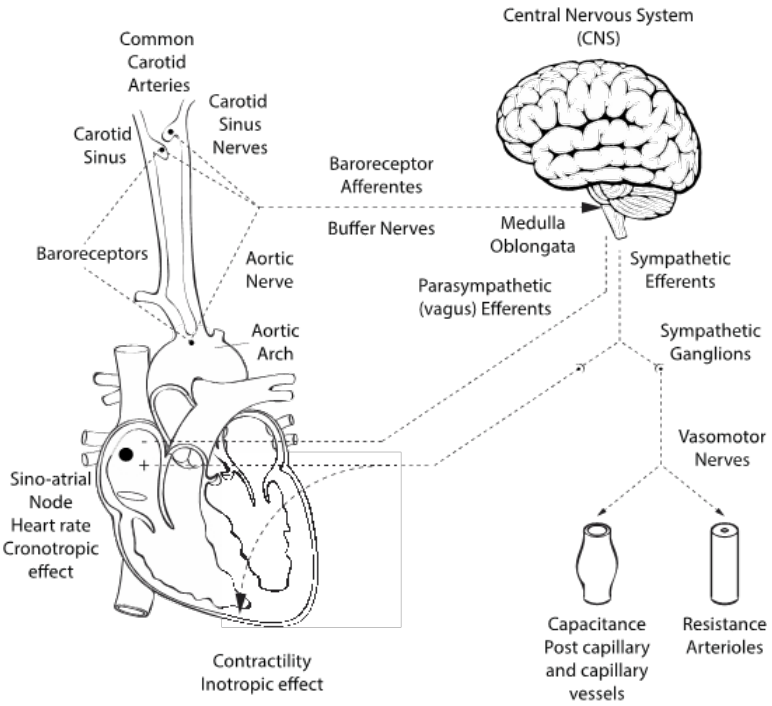


Figure 2.2: Representation of the baroreflex feedback system. Adapted from [2]

2.2 The cardiovascular system

The cardiovascular system is constituted by: the heart, the lymphatic system and blood vessels. The last two components serve the purpose of irrigating the whole body while serving as a exchange system. The blood vessels themselves can be subdivided in different categories: arteries, arterioles (that result from the division of arteries), capillaries (that are a smaller version of the arteries and therefore allow exchanges between blood and the surrounding systems); these blood vessels then see a increase in their volume and progressively become venules and finally veins. The heart is responsible for pumping the blood through the whole system with its ventricles while the atria are the cavities that receive the blood from vascular system. [4]

There are two main circulations in this system: the pulmonary and the systemic circulation, which are represented in Figure 2.3. The first one is the one responsible for the oxygenation of blood and starts at the right ventricle that injects the blood to the lungs where blood is oxygenated. The blood then returns to the heart through the left atrium. Regarding the systemic circulation, oxygenated blood is sent from the left ventricle through the aorta to the whole body in blood vessels while diffusing oxygen and receiving carbon dioxide from cells. Blood then returns to the right atrium through the vena cava. [4]

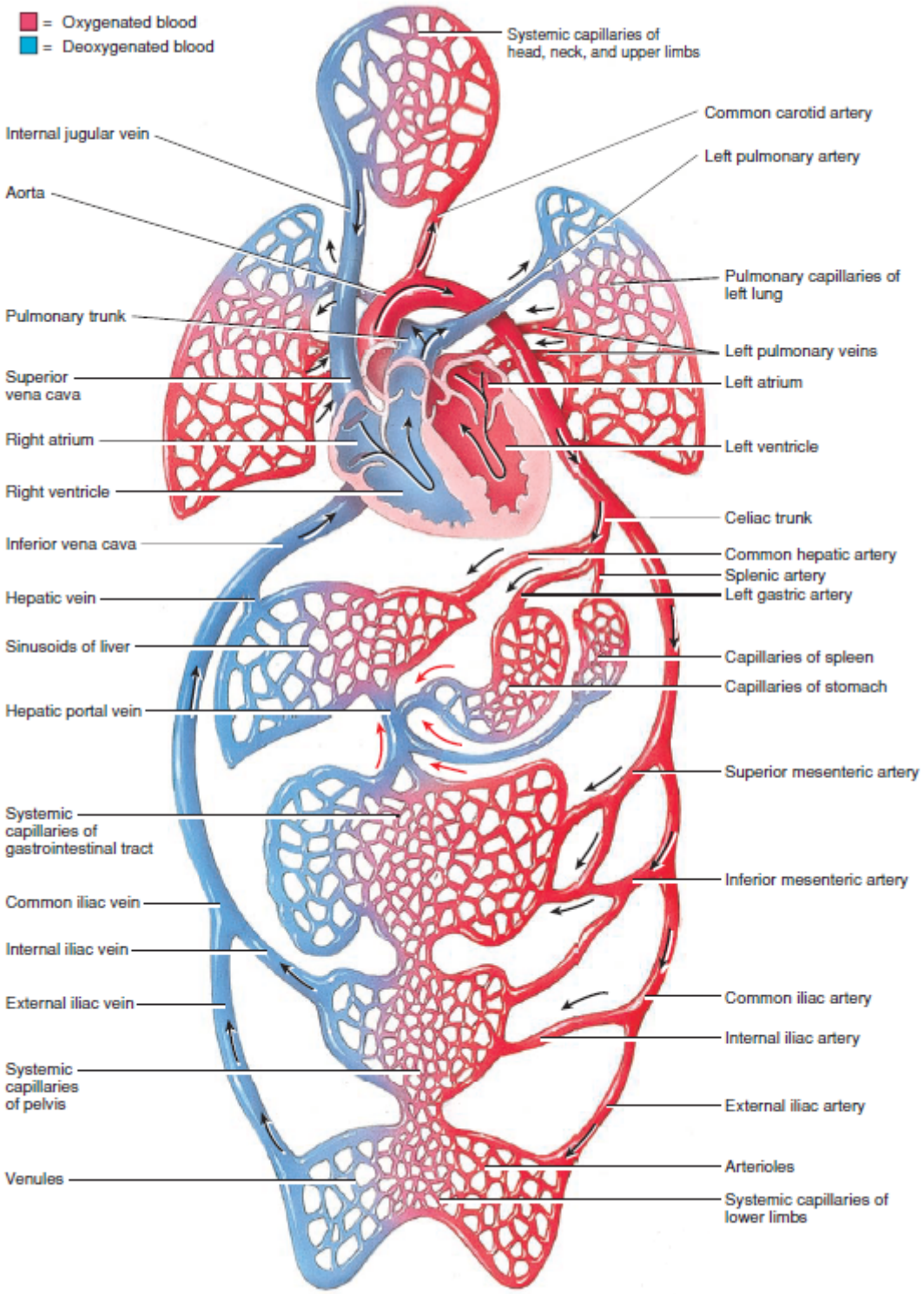


Figure 2.3: Representation of the cardiovascular system and circulatory routes. Adapted from [3].

2.3 The heart

The heart is constituted by 4 main cavities, represented in Figure 2.4: 2 atria and 2 ventricles. The right compartments are connected through the tricuspid valve while the left side is linked through the bicuspid vale. These valves, also known as atriventricular valves force the blood from the atria to the ventricles and not the other way around. There are also the semilunar valves that control the flow of blood from the ventricles to the rest of the body known as the aortic valve that controls the exit of blood to the systemic circulation and the pulmonary valve that controls the blood flow to the pulmonary circulation. [4]

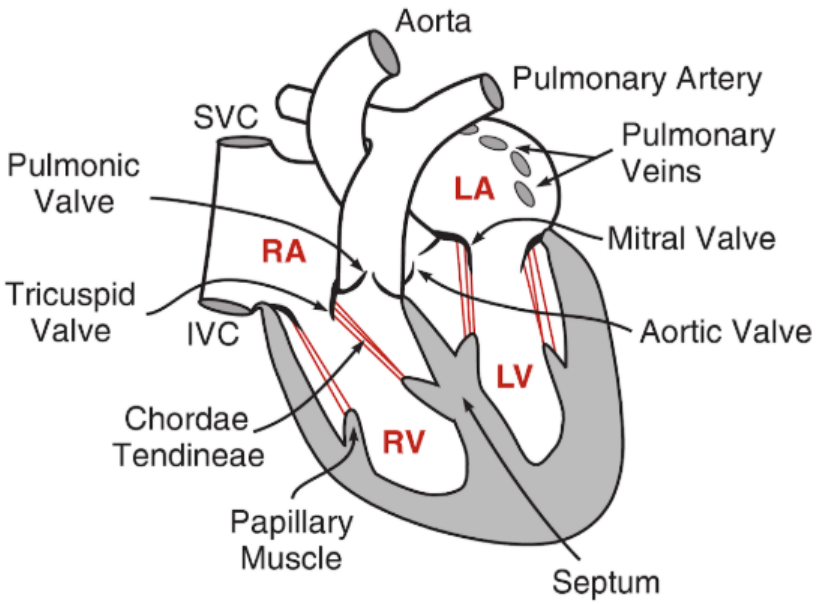


Figure 2.4: Representation of the heart. SVC, superior vena cava; IVC, inferior vena cava; RA, right atrium; RV, right ventricle; LA, left atrium; LV, left ventricle. Adapted from [4].

2.4 Electrocardiogram and the cardiac cycle

An electrocardiogram (ECG) is a signal recording that results from the electrical currents caused by action potentials propagated through the heart detected at the surface of the body. These electrical currents promote successive contractions and relaxations of the heart cavities known as systole and diastole, respectively, creating the cardiac cycle. With the depolarization of the atria, that spreads from the sinoatrial node (has the ability to spontaneously produce an electrical impulse), the P wave is formed and there is an atrial systole. Following this there is a rapid ventricular depolarization (ventricular systole) that is represented by the QRS complex. The T wave represents the ventricular repolarization (ventricular diastole), which is succeeded for the ending flat ECG segment. This process is represented in Figures 2.5 and 2.6. [4] [3]

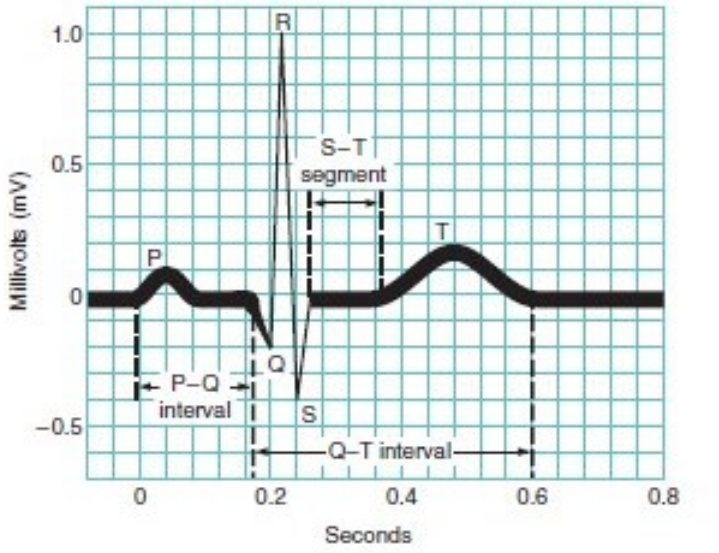


Figure 2.5: Representation of the ECG wave and its main components (P, QRS and T waves). Adapted from [3]

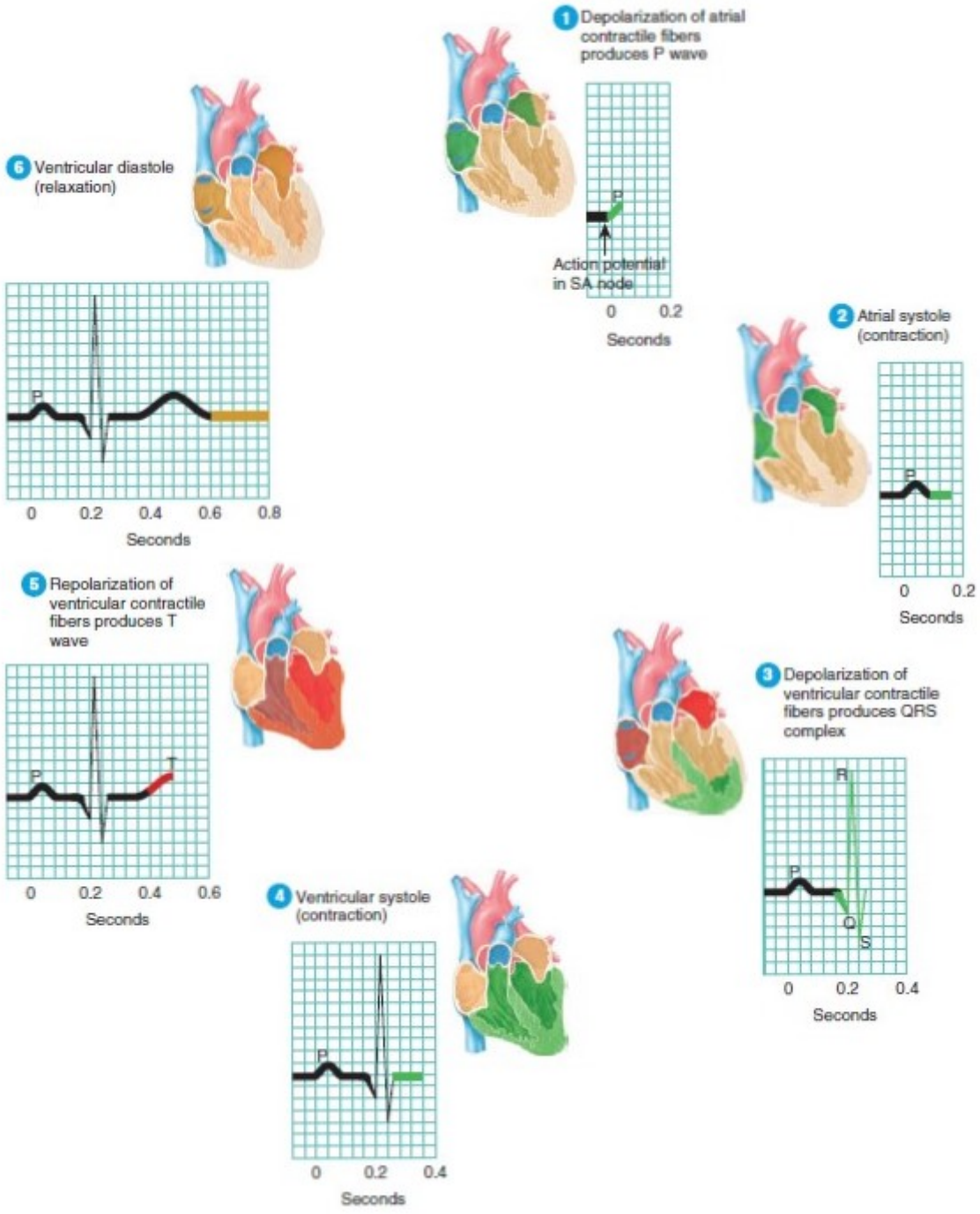


Figure 2.6: Representation of the cardiac cycle and respective depolarization (green) and repolarization (red) phenomenons. Adapted from [3].

2.5 Heart rate variability

With the R peak information it is possible to derive a RR interval time series, successive distances between two consecutive peaks, also known as tachogram. This variance in time is also known as heart rate variability [11]. HRV is regulated by both parasympathetic and sympathetic divisions of the ANS even though the former has a more predominant effect than the later [12]. The vagus nerve is the main nerve of the parasympathetic cardiac regulation (also controls lungs and digestive track), which makes the HRV an index for cardiac vagal tone, which represents the contribution of the parasympathetic division in cardiac regulation and is used in many researches in cognitive, health, emotional and social subjects [13]. The high frequency components of the HRV, between 0.15 and 0.4 Hz, are the ones accepted as a measure of the vagal action in the heart while low frequencies, around 0.1 Hz, are linked to vasomotion and sympathetic division influence on the heart [8].

3

State of Art

3.1 Overview

Software Engineering is an area that faces many hard decisions when developing new software. One of the most important is the choice of software development life cycle that involves important phases like planning, analysis, design, implementation and maintenance. Over time there were developed a lot of software development approaches namely the waterfall and agile models, which are two very different and popular methods [14]. The waterfall model, one of the oldest and best known approach, is a sequential development model where each activity is completed before moving on to the next without any overlap in each step. On the other hand, the agile model welcomes requirement changes even late in the development process and prioritizes customer satisfaction with rapid delivery of incremental software versions. The first approach excels when applied to a larger project and the requirements are clear from the start, while the second one is optimal when the requirements change quickly in smaller projects [15] [16].

During the life cycle of a software product all activities are prone to faults. To deal with these faults there are many techniques that improve software reliability and have been classified into 3 different categories: fault avoidance that aim to prevent the introduction of faults (process oriented), fault detection that aim to detect the fault (product oriented) and fault tolerance that give a controlled response to the uncovered faults [17]. Due to the aim of project *BASE*, the focus is the fault avoidance category, which includes verification and validation, that encourage careful examination of the program and testing, that tries to compensate human fallibility and proof methodology.[14][17]

Code complexity can also be used as part of a fault avoidance strategy, considering that the complexity of a software has a direct impact in the numbers of faults

associated with code. The higher the complexity the harder it is to understand, test, and maintain the software [17]. The level of complexity is difficult to access because it is not always clear what the measures are supposed to be measuring. There was a study in 1988 by Elaine J. Weyuker that formalized certain theoretical proprieties to evaluate measures like McCabe's cyclomatic metric (based in a program's decision structure) [18] and Halstead's effort metric (based in the number and frequency of operands and operators in a program) [19] in more specific settings. The results showed that the first metric was not sensitive enough to evaluate code complexity because of its inability to evaluate the computation performed, therefore giving different complexity programs the same score as long as their decision structure was similar. Both metrics also revealed a lack of responsiveness to the interactions that take place between the different program units [20].

Overall the methods and metrics used in software reliability are difficult to apply or have a limited applicability beyond the traditional use of complexity metrics in testing approaches. Considering this, new alternatives should be explored with the constant growth of software complexity [17]. The goal of the BASE project is to research possible ways to gather information on programmers' mental effort and cognitive load (among other metal aspects that can be related to human error) while programming and/or inspecting software code and introduce such information in the software development process with the goal of reducing the number of software faults.

Over the years many studies in brain activity and autonomous nervous system manifestations have been made regarding mental effort and cognitive stress, which may play an important role in the future of software reliability.

In 1998, *SAM Technology* and *EEG Systems Laboratory*, in San Francisco, California evaluated the working memory load of people performing computer-based tasks that involved memorizing letters and their respective position over time. Working memory can be defined as the capacity to retain information in mind for a short time in a cognitive activity context [21]. In this study spectral components of the measured electroencephalogram (EEG), known to be sensitive to cognitive activity, were used and allowed to conclude that theta activity (4-7 Hz) at frontal sites increased while alpha (8-12 Hz) and beta (13-30 Hz) activity decreased with the rise of complexity of the tasks (increase in required working memory load) [22].

Another approach to the study of mental effort, this time using eye activity, was made by the *University of New South Wales* in Australia in 2011. A group of basketball players was assigned to participate in a computer-based experiment where

the players had to watch a 15 second video of a basketball game and then identify attackers and defenders and recall their respective positions. Results showed that with the increase in task difficulty, participants increased their blink latency, pupil size and fixation duration and there was an overall decrease in blink rate, fixation rate and saccade speed and size (conjugate eye movements that abruptly change the point of fixation [23]) [9].

There are many fields yet to explore that could improve software reliability as showed in the last two examples however the focus of this thesis is on heart activity, more specifically, heart rate variability, which is outline in the next section.

3.2 Heart rate variability

Based on the information of a ECG many features can be calculated from the time, frequency and even non-linear domain. According to [24] the most prominent features in the time domain are the following:

- **Mean** mean of the time intervals within a fixed sliding window;
- **SDNN** SD of the NN intervals;
- **SDSD** SD of successive differences between adjacent NN intervals;
- **RMSSD** square root of the mean squared differences between adjacent NN intervals;
- **NN50** number of consecutive NN intervals that differ more than 50 milliseconds;
- **pNN50** ratio between NN50 and the total number of NN intervals.

The frequency domain features are normally computed through the power spectrum density (PSD) with the Fast Fourier Transform (FFT) where powers are calculated by integrating the whole spectrum or autoregressive models where components are divided and then the band powers are computed [25]. Some examples of frequency domain HRV variables are:

- **HF** power in high frequency range (0.15-0.4 Hz);
- **LF** power in low frequency range (0.04-0.15 Hz);
- **VLF** power in very low frequency range (0.003-0.04 Hz);
- **LH** ratio of low and high frequency.

Relative versions of these features are also commonly used in HRV studies [25].

Regarding the non-linear approaches, the *Poincaré* plot takes in consideration the dynamics of variations between RR intervals. This is a plot in which each RR interval is a function of its previous RR interval. It can be used to provide summary information, by calculating the standard deviations of the distances of *R-R(i)* to $y=x$ and $y=-x+2R-Rm$, where *R-Rm* is the mean of all *R-R(i)*. As shown in Figure 3.1 *SD1* is related to the fast beat-to-beat variability while *SD2* describes longer-term variations of *R-R(i)*. [5]

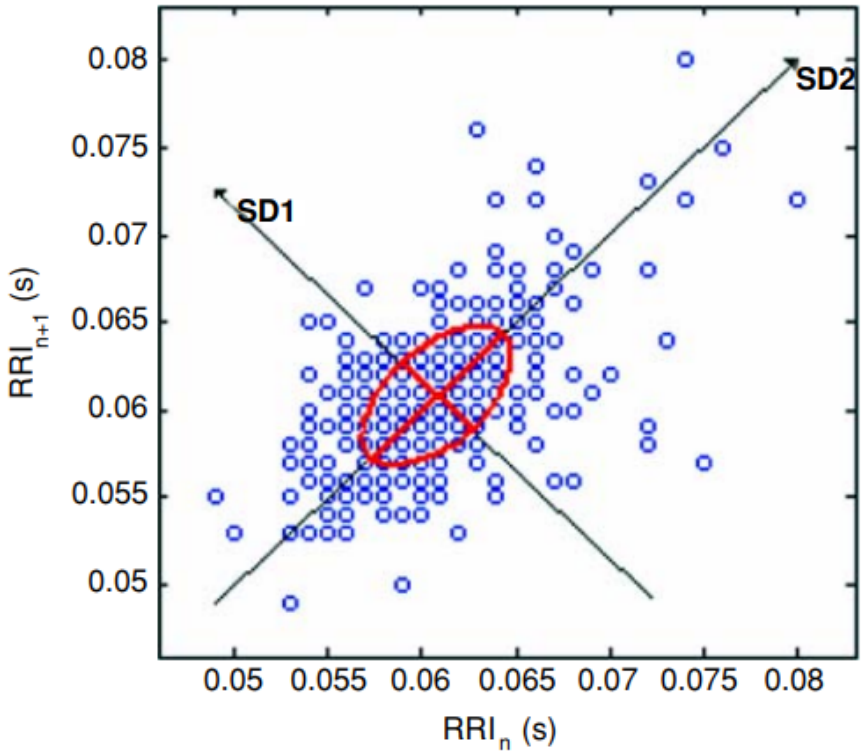


Figure 3.1: Example of a *Poincaré* plot. Adapted from [5].

In the *Poincaré* plot every RR interval is plotted against its previous RR interval. With this plot it is possible to fit an ellipse in which the two main axis (perpendicular to each other) can be used to calculate the SD of the distance from all points to each axis. The SD associated to the minor axis reflects short term variability while the SD associated with the major axis represents long-term variability [26].

3.3 HRV and cognitive stress quantification

In 2006, Sergio Cerutti et al. [8] studied a possible solution for a stress and sleep management biofeedback tool that aims to improve quality of life by empowering the users to have more control over their health. To achieve this the proposed solution would use HRV that reflects the status of the autonomic nervous system in controlling heart frequency and the sympatho-vagal balance. Considering this, registration of heart rate fluctuations is suggested to offer an insight into the autonomic cardiovascular regulation. This data was then analysed using power spectrum density that allows the separation of the vagal and sympathetic activity in the HRV signal using HF and LF. With this study it was concluded that heart rate fluctuations indexes, namely HF and LF, can be a measure of stress by exploring their impact on the autonomic cardiovascular control.

J. Taelman et al., in [11], studied the changes of heart rate (HR) and HRV in a group of 28 subjects performing an IQ test (mental task) and during rest. When exposed to a mental task and stress is induced, the parasympathetic nervous system is suppressed, and the sympathetic nervous system is activated [27] which lead to, for example, vasoconstriction of blood vessels and increased muscle tension and affects HR and HRV. To measure HRV the *Pan-Tompkins* algorithm was used to detect the QRS-complexes that allowed to determine the RR-intervals for each subject in the two situations [28]. The results showed that mean RR was significantly lower in the mental task compared to the rest condition, pNN50 was significantly higher in the rest condition and that the SD did not vary significantly in the two conditions. Regarding the frequency measures, calculated in this study using the Fourier transform, there weren't any significant differences. Despite this, a tendency was found for higher LF/HF ratio in the mental task condition.

Another study that was published in 2008 by M. De Rivecourt et al. [29], in which different levels of mental effort were present instead of the typical rest vs mental task conditions exemplified before. The experiment involved 19 male individuals participating in simulated flight tasks with varying numbers of maneuvers with different complexities. These tasks were divided into 6 groups and were classified by each subject according to the rating scale of mental effort (RSME [30]) in order to provide a subjective measurement of each task perceived by the participant. This metric allowed to conclude that the complexity was increasing as expected throughout the tasks. Results showed that HF spectral power calculated using direct Fourier transform computed from the time points of R peaks from ECG were

able to distinguish between 3 levels of mental effort.

Contrarily to all the previous studies focusing on linear techniques to analyse HRV, in 2011, Shalini Mukherjee et al. [31] proposed an approach focused on the *Poincaré* method to estimate the HRV function in two visual working memory tasks carried out by healthy seniors. The results showed that linear metrics, like the mean RR, were more sensitive to varying mental effort loads. Despite this, *Poincaré* measures also displayed sensitivity to mental effort changes and proved to be a useful marker for cognitive based tasks.

There were many studies over the years, like the ones mentioned before, that proved that HRV can be used as an index for cardiac vagal tone. As a result of the prominence of this signal many considerations regarding experience set up and overall methodologies have been used as a reference for these studies over the last two decades: within-subject studies are advised due to the high variability in inter-individual HRV measures [32]; comparations between studies also aren't advised considering that methodological aspects such as the body position during HRV acquisition can affect HRV; sampling rate of the data acquisition system should be at least 200 Hz [24] (conservative guidelines suggest sampling frequency between 500 and 1000 Hz) [33]; 5-minute data acquisition is the gold standard proposed by the Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology to optimize HRV parameters for short-term studies. This value comes from the fact that recordings should last for at least 10 times the lowest frequency component present in the study for analysis [24]. Despite this, depending on the objectives of the experiment, 1-minute recordings can be used considering that these are sufficient to study HF from HRV [13].

4

Data collection

4.1 Overview

The experiment included 30 participants and all the subjects had experience in the *Java* programming language. There was a screening process that included an interview to assign a programming level skill in Java to each participant as "Intermediate", "Advanced" or "Expert".

The experimental set up covered a comprehensive set of sensors including EEG with a 64-channel cap, ECG, Electrodermal activity (EDA), Impedance cardiography (ICG), Photoplethysmogram (PPG) and eye tracking with pupillography. The signals from all these sources were synchronized in a common time base to allow consistent cross analysis.

Experiments involved the compression of 3 small *Java* programs that had different complexities according to software complexity metrics like lines of code and cyclomatic complexity. The programs were carefully designed to keep consistency in the programming style and to avoid that math or algorithm pose extra difficulties to the participants, not directly related to the code complexity. All the participants performed the experiments in the same room, without distractions, noise or presence of people unrelated to the experiments.

The study was approved by the Ethics Committee of the *Faculty of Medicine* of the *University of Coimbra*, in accordance with the Declaration of Helsinki and all experiments were performed in accordance with relevant guidelines and regulations. Informed consent (Appendix A) was signed by all participants.

4.2 Volunteers screening and recruiting

4.2.1 Announcement

Experiments were mainly promoted through e-mail and face to face talks with students and other potential participants including teachers and professional software developers. During this phase, potential participants were made aware of the overall goal of the experiment: "Assess whether different levels of cognitive stress induced during software inspection/reading can be captured and quantified using autonomic nervous system activity manifestations captured using wearable sensors", the reward associated with the participation (vouchers to use in *SONAE* stores) and the start date of the experiments: May 2018. Due to a delay in the vouchers' purchase process the experiments were delayed to June.

4.2.2 Screening interview

People who show interest in participating in the research were interviewed. Throughout this interview subjects were asked some questions regarding demographics, contacts, profession, programming experience and availability. Interviewers did also take notes related to the overall motivation of the person, regarding the experiment, that could impact the results of the research. All information regarding the participants has been managed according to the new legislation of data protection, *GDPR*.

During the interview information regarding the location: "IBILI Pólo3 – lab 58" and estimated duration: "2 hours 30 minutes" was disclosed. It was also important that the subject was aware that the experiments involved product application on hair for the EEG set up and that they are were not eligible for the experiment if they had any kind of implanted cardio device.

Subjects didn't need any type of preparation before the experiment.

4.2.3 Screening

The goal was to have around 30 subjects, balanced in gender and proficiency in *Java* programming. To distinguish the proficiency of the candidates, the following criteria was applied:

- **Expert** Professional software developers that worked in the field for more than 3 year or had more than 15000 lines of *Java* language programmed in the last 3 months or more than 10000 *Java* lines written in their biggest program or teachers that had more than 20000 lines of *Java* language programmed in the last 3 months or more than 15000 *Java* lines written in their biggest program;
- **Advanced** Professional software developers that worked in the field for more than 1 year or had more than 10000 lines of *Java* language programmed in the last 3 months or more than 5000 *Java* lines written in their biggest program or students and teachers that had more than 10000 lines of *Java* language programmed in the last 3 months or more than 5000 *Java* lines written in their biggest program;
- **Intermediate** Students or teachers that had more than 5000 lines (and less than 10000 lines) of *Java* language programmed in the last 3 months or more than 3000 *Java* lines (and less than 5000 lines) written in their biggest program;

Subjects that did not fulfil these requirements or had a history of heart diseases, mental health issues or implanted cardio devices were not eligible for the experiments.

4.2.4 Recruitment

From the 47 interviewed candidates 30 participants were selected to participate in the experiments. Although gender and proficiency balance was established as an primary objective due to the insufficient number of "expert" and female candidates it wasn't possible to achieve it. The group of participants included 24 men and 6 women. Regarding the proficiency in *Java* programming 13 participants were classified as "intermediate", 12 as "advanced" and 5 as "expert" according to the responses given during the screening interview.

4.3 Experimental set up

In this experiment four different data recording systems were used to record different signals:

- **Neuroscan's equipment** composed by a 64-channel cap that connects to an amplifier headbox that is attached to a system unit that communicates through USB with a computer that controls the equipment with Neuroscan's *Scan 4.4* software. This system records EEG and ECG signals at a 1000 Hz sample frequency;
- **BiosignalsPlux** research kit with a Wireless 4-channel hub with a variety of sensors that is controlled through bluetooth with *Opensignals*. ECG and EDA sensors were connected to the hub to record these signals at 500 Hz sample frequency;
- **SMI eye tracker** equipment that was positioned under the monitor that displayed the stimulus and allowed to collect pupillography and eye movement tracking at 500 Hz and was monitored by the *Iviewx* software through USB;
- **Sensatron** equipment developed by *Philips* that was used to measure ECG at 250 Hz, ICG at 125 Hz, PPG at 62.5 Hz. This equipment has a memory card and a button that starts and ends the recording therefore no software was used to control the acquisition of these signals.

The sensors positioning of the different data recording equipment is displayed in Figures 4.1, 4.2, 4.3 and 4.4. In figure 4.4 the 64-channel cap used to record EEG is placed on a participant. This process requires special preparation in which the scalp of the participant is cleaned with abrasive gel and alcohol to improve the conductivity of a different gel applied to each sensor to capture the signal with reduced noise. The eye tracker doesn't require any sensors due to the fact that data is recorded directly from the eye movements and changes compared to a reference state recorded right before each trial. The room's light in which the experiments were conducted was controlled and constant to reduce inter-subject variations.

Since all the participants answered surveys after each trial the removal of the sensors from the hand was allowed (e.g., the PPG sensor) to facilitate writing and were then reapplied before the start of the following trial.

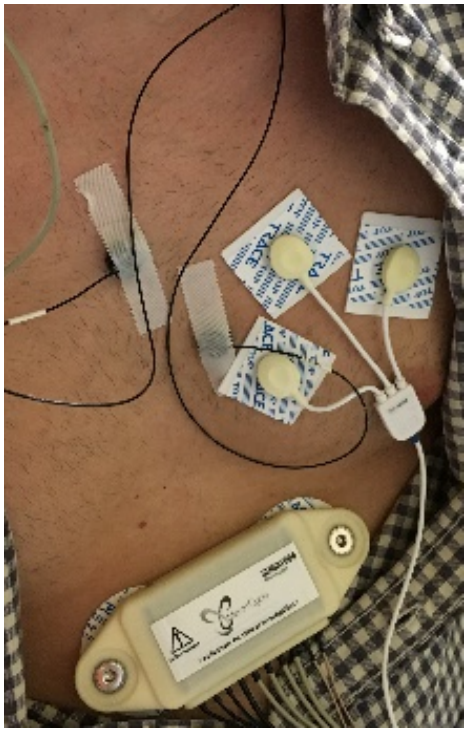


Figure 4.1: Neuroscan's (black) and BiosignalsPlux's sensors (white) positioned for ECG data recording. In this image it is also possible to see the Sensatron equipment on the diaphragm.

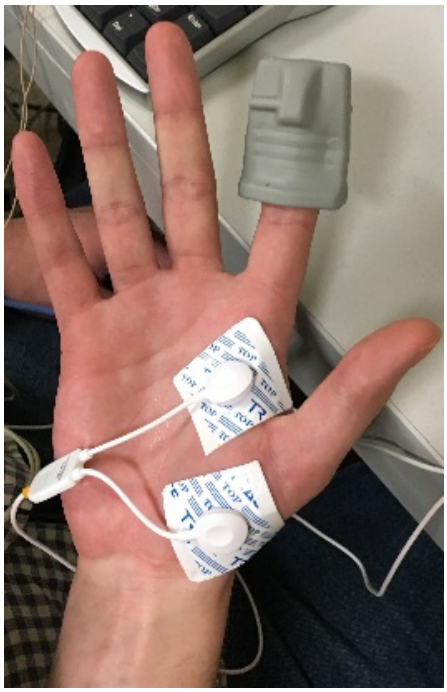


Figure 4.2: BiosignalsPlux's sensors (white) positioned for EDA measuring. The index finger's sensor measures PPG and is connected to the Sensatron equipment.

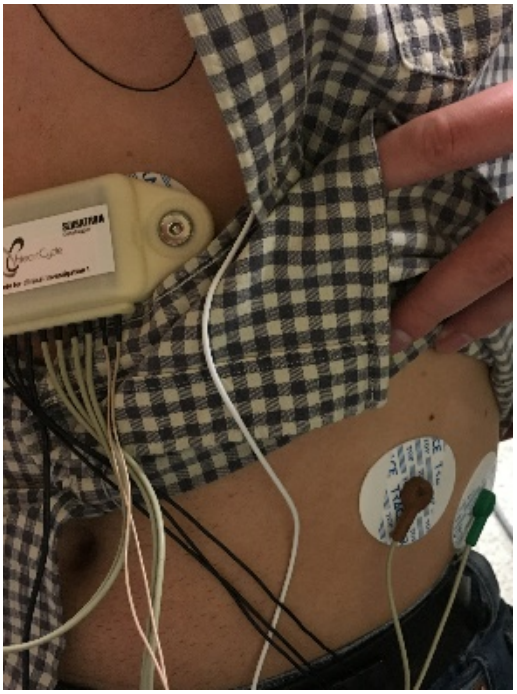


Figure 4.3: Sensatron's sensors positioned to measure ECG and ICG with the aid of two sensors displayed in Figure 4.4.

Figure 4.4: Sensatron's sensors positioned in the neck to measure ECG and ICG with the aid of the sensors from Figure 4.3. In this figure the participant is using a cap with sensors that is part of the Neuroscan's equipment for EEG measuring.

The overall experimental set up is displayed in figure 4.5. Neuroscan's equipment was connected to a pc mainly responsible for the EEG acquisition which was connected to the computer that controls the overall experiment set up and sent its triggers through a parallel port. The eye tracker and the BiosignalsPlux hub were also connected to their own computers responsible for the data acquisitions which communicated with the computer that controlled the set up through a TCP/IP connection. Finally, the Sensatron wasn't directly controlled by any computer. Instead, the synchronization of this device with the remaining systems was performed in two phases:

1. A pre-synchronization step was performed to align the "bip" sounds (that act like triggers) produced by the computer responsible for the stimulus, which were captured by the microphones of this device with the remaining triggers;

2. The signals captured by this device were then synchronized using the RR intervals of the collected ECG and the RR intervals of the ECG collected by the other devices.

The synchronization procedure using the ECG signals is explained in detail in [34].

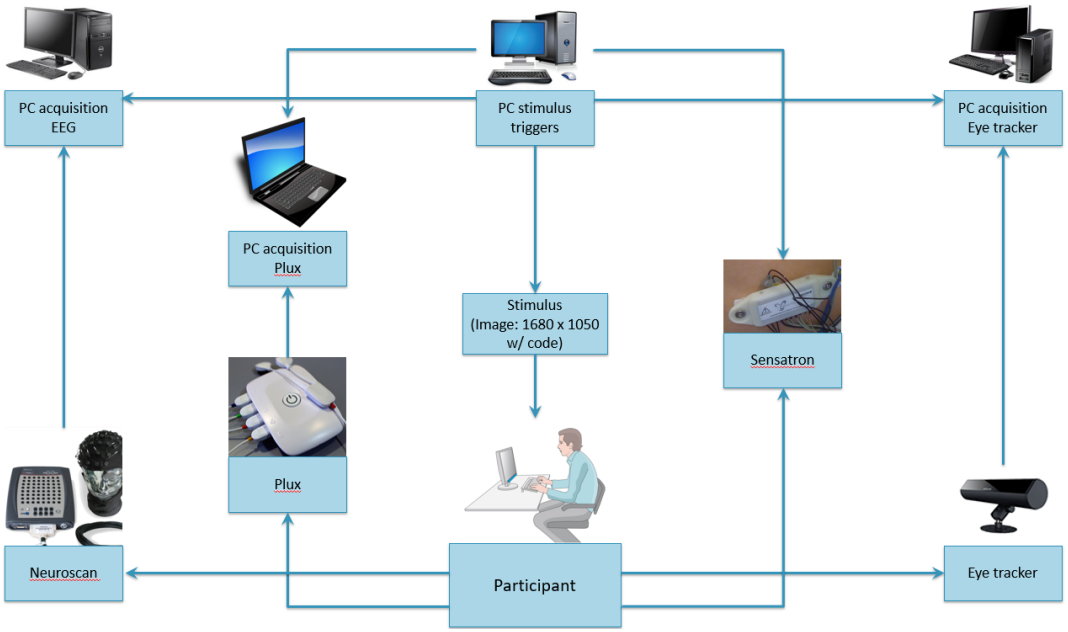


Figure 4.5: Schematic representation of the experimental set up with its components and respective relations.

The stimulus of the experiment was created with *Psychtoolbox-3* software in *Matlab* and displayed by the computer that controls the experimental set up directly to the subject. This computer had 2 monitors, one that displayed the stimulus and another that was used to control the experiment, and a keyboard that allowed the

user to skip to the next phase of the protocol if necessary (as explained in the next section).

4.4 Experimental protocol

The experiment started with an empty grey screen with a black cross in its center. During this phase, the participant focused on the cross in order to abstract himself from the environment and to adjust to the experimental environment. This phase lasted for 30 seconds and is considered the baseline phase of the experiment where the participant didn't perform any activity.

In the second phase, a text in natural language was displayed to the subject. In order to minimize the effort of the participant related to the reading task, it was defined that the text should be written in natural language.For this purpose, the selected text was based on Portuguese histories in a tentative to avoid fluctuations in measurements due to emotions triggered by the narrative. The duration of this step was 60 seconds and to ensure that the participant was reading during the whole time of the current step, the text had a higher estimated reading time and the participants were informed that, during this phase, they should read the text until it stopped displaying on the screen (60 seconds) and should not be concerned with reading the whole story. This activity serves as a reference phase for data analysis.

The third phase was equal to the first one. It had the purpose of abstracting the participant from the previous step and avoid interfering with the next one.

During the fourth phase, a code snippet in *Java* programming language was displayed to the participants in order to be interpreted and analyzed. The duration of this phase was set to a maximum of 10 minutes, but it could be interrupted at any time, in case the participant concluded its analysis earlier. This possibility was included in the experiment in order to ensure that during this step the participant was always focused on the performed task and didn't tamper the experiment objectives.

The last phase was equal to first and third and served the purpose of allowing data recording after the main stimulus for residual information and avoided ending the experiment abruptly.

This whole process, represented in Figure 4.6, was repeated 3 times.

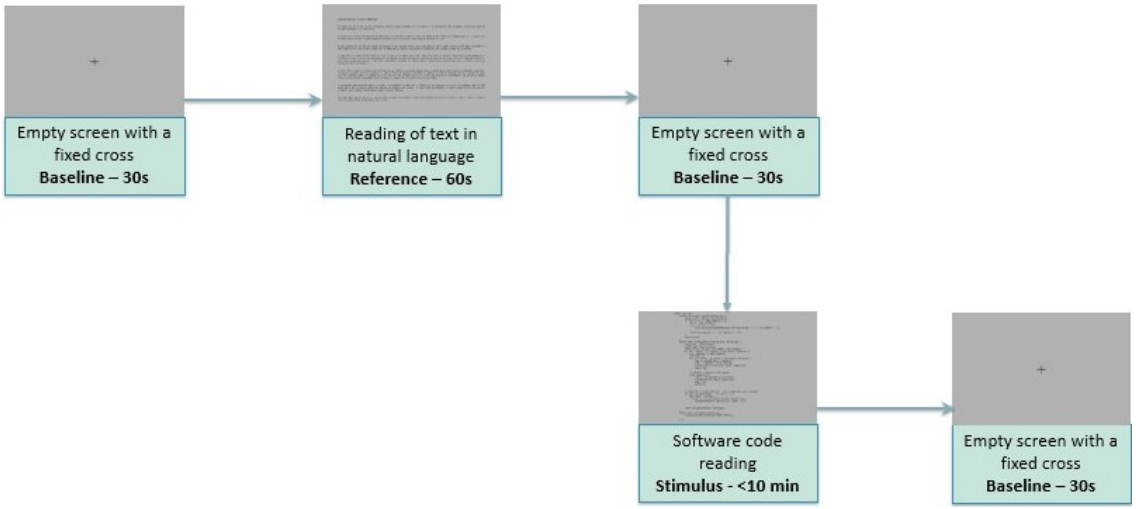


Figure 4.6: Schematic representation of the sequence of the main phases followed in the experimental protocol.

4.4.1 Code snippets and texts

In each trial the code snippet displayed to the subject was different from the remaining trials, as well as the code snippet complexity. A short description of the code snippets is provided below:

1. The first program counts the number of values existing in a given array that fall within a given interval using a straightforward loop;
2. The second one multiplies two numbers using the basic algorithm where every digit from one number is multiplied by every digit from the other number, from right to left. The numbers are given as strings and the algorithm includes converting the strings to byte arrays;
3. The last program seeks occurrences of an integer cubic array inside a larger cubic array, trying to find the larger occurrence of the smaller one inside the larger one.

The complexities of the 3 programs, according to different metrics, are displayed in Table 4.1. The first code, due to its overall inferior complexity, had a maximum display time during phase four of 5 minutes while the other two codes had a maximum time of 10 minutes.

Code snippet	Number of lines	Number of parameters	Cyclomatic complexity
1	13	3	3
2	47	3	4
3	49	4	15

Table 4.1: Complexity of the 3 programs displayed according to 3 distinct software engineering metrics.

The displayed texts were also different in each trial to keep the reader engaged throughout the experiment. The texts (extracted from [35], [36] and [37]) and the codes displayed to the subjects during the experiments are presented in the Appendix C.

4.4.2 Cautions

During the trials participants tried to understand the code without the help of the supervisors in the room and without using any auxiliary tools, like pen and paper, for example. Participants also tried not move and talk during the trials to avoid interfering in the data recording.

Before the start of the experiment a set of sensors had to be applied to the participant. These sensors were verified before each trial in order to ensure they were correctly positioned and acquiring the signals correctly.

At the beginning of the experiment, that is, before the positioning of the sensors and preparation of the participant, all the participants were informed with a summary of the informed consent (Appendix A) and were asked to read the consent and sign it. In this consent, the experiment details and the participants' rights were fully explained. After the end of the experiment the participants were rewarded with a voucher from the *SONAE* group, valued at 50 euros, and were asked to sign a declaration stating they received the reward (Appendix B).

4.4.3 Performance evaluation

After every trial the participants answered two questionnaires. In the first questionnaire the participants were asked to explain the general goal of the algorithm and what were the steps taken in the algorithm to accomplish its objective. The purpose of these questions was to ensure that the participants were focused on the activity

and that there was mental effort involved during the trial. Participants didn't have to completely understand the algorithms considering that the effort displayed during the process was the relevant part of the data recording. This assumption is based on the fact that the participant can make an effort to comprehend the algorithm and still don't understand it by the end of the trial. Nevertheless, the responses were analysed to make sure that there was a minimum comprehension of the displayed codes. This document, as well as the document with the responses used as reference for evaluation, are in Appendix D.

The second questionnaire was based on the *NASA's* task load index [38] and had 4 questions in which participants had to rate the levels of mental effort, task completion, temporal demand and frustration, (opposed to the 6 parameters presented in [38] that also regarded physical effort) from 1 to 6 (in *NASA's* version participants have a scale from 1 to 20 that was too large for this experiment) regarding the trial. In this document the participant chose which parameter was thought to be more relevant during the trial from groups of two. These responses could be readjusted by the participants after the completion of all trials in case the perceived complexity of the first two programs changed at the end of the experiment. The purpose of this survey was to access the complexity of the 3 algorithms in the perspective of each participant and its most prominent factors in the subjective workload. These results are later compared with the algorithms actual complexity according to the complexity measures mentioned before while the perceived relevancy of the factors will be used in future studies. The document used for this purpose can be found in Appendix D.3.

Between each trial, it was not defined a limited amount of time to complete the survey. Since in these phases the participant was not performing any specific task of code inspection and no signals were being recorded, it can be considered as a moment of relaxation and rest, which can provide a clearer separation between trials (avoiding them influencing each other) and strengthen the conclusions of the study.

The estimated duration of the experiment was 2 hours.

4.5 Data structure

For every participant's trial there is a *Matlab* file containing its associated information regarding the collected signals making a total of 90 *matlab* files (3 trials for 30 participants). Each file contains a cell array that is organized in 4 columns with a varying number of rows that contains a structure array in each cell. The first column has information about the BiosignalsPlux system in the third row and the EDA and ECG signals recorded from this equipment in the first and second rows, respectively. The second column with 25 rows contains information related to the Sensatron system. The recorded microphone data is in row 1 and 2, ECG in row 3, ICG in row 18 and PPG in row 21, 22 and 23. This system has other sensors implemented that weren't used in this study that have their information in the remaining rows. In the third column, the first row contains the EEG data and the second one the information from the Neuroscan's equipment. Finally, the fourth and last column retains the SMI eye tracker recorded information in the first row and equipment's specific settings in the second one. Information related to the synchronization triggers are saved in the last row of each column associated with the respective system. A simplified representation of this structure is displayed in Table 4.2.

Row	BiosignalsPlux	Sensatron	Neuroscan	SMI
1	EDA	Sound (1)	EEG	Eye tracker and pupillography
2	ECG	Sound (2)	-	-
3	-	ECG	-	-
18	-	ICG	-	-
21	-	PPG (1)	-	-
22	-	PPG (2)	-	-
23	-	PPG (3)	-	-

Table 4.2: Representation of the collected signals' data in the cell array.

5

Methods

5.1 Preprocessing

Before data analysis there was a visual inspection of the electrocardiograms to check for segments of data with noise that affected R peak detection in order to remove them.

5.2 Data segmentation

Two different segments of the recorded data were used. The first segment started at beginning of the display of the text and ended at the start of the display of the code. This segment was defined to represent a resting phase (only containing the second cross display). Due to the fact that a 60-second window was used to calculate HRV features (considering the low amount of data and the fact that 1-minute window is known to be enough to analyse some features like HF [32]) and the display of the text only lasted for 60 seconds there was a need to add the second cross display phase to have enough data to analyse HRV. This way, the baseline data comprehends both minimal mental effort activity (reading a text) and a resting state (staring at a cross). The second segment, corresponding to the phase where the code snippet is displayed to the subject, represents a mental effort segment. The data used for both segments didn't start immediately after their starting triggers to avoid interferences characteristic of protocol transitions and instead started 5 seconds after each trigger that marks the transaction of the different protocol phases. For every participant there were 3 pairs of segments (rest and effort) correspondent to the 3 trials.

Considering that 4 of the participants read and understood the code of the first trial in less than 1 minute there wouldn't be enough data to analyse HRV, so their data

wasn't used in this study making a total of 26 participants' data used in the current analysis.

5.3 Heart rate variability

5.3.1 R peak detection

For the calculation of HRV, first, the R peaks were detected with the ECG processing toolbox provided by *LiNK* project based on the algorithm developed by Yan Sun et al. [39]. This toolbox uses a multiscale morphological derivative (MMD) transform-based singularity detector that identifies fiducial points in the ECG signal. A singular point is a point where the derivative on the right and left have different signs. These points' derivatives on the left and right are defined using MMD, respectively, as following:

$$M_f^-(x) = \lim_{s \to 0} \frac{f(x) - (f \ominus g_s)(x)}{s}$$

$$M_f^+(x) = \lim_{s \to 0} \frac{(f \oplus g_s)(x) - f(x)}{s}$$

where $\oplus$ (dilation) and $\ominus$ (erosion) are morphological operators defined as:

$$(f \oplus g_s)(x) = sup\left\{f(x-t) + g_s(t)\right\}, t\epsilon(G_S \cap D_x)$$

$$(f \ominus g_s)(x) = inf\left\{f(x+t) - g_s(t)\right\}, t\epsilon(G_S \cap D_x)$$

respectively and $f : D \subset \mathbb{R}^n \to \mathbb{R}$, g_s is the scaled structuring function $g_s : G_s \subset \mathbb{R}^n \to R, s > 0$, D_x is the translation of D, $D_x = x + t : t\epsilon D$, $sup(f)$ and $inf(f)$ refer to the least upper bound and greatest lower bound of f and s is the scale.

In this algorithm, a MMD difference is defined as:

$$M_f^d(x) = M_f^+(x) - M_f^-(x)$$

$$M_f^d(x) = \lim_{s \to 0} \frac{(f \oplus g_s)(x) + (f \ominus g_s)(x) - 2f(x)}{s}$$

with the following expression

$$M_f^d(x) = \frac{(f \oplus g_s)(x) + (f \ominus g_s)(x) - 2f(x)}{s}$$

being its scaled version. The previous expression can be simplified with the choice of a flat structuring function to

$$M_f^{d_s}(x) = \frac{maxf(t) + minf(t) - 2f(x)}{s}, t\epsilon\,[x - s,x + s]$$

By choosing a moving window with a length of (2s+1) samples, it is possible to calculate MMD transform in a point with the maximum and minimum in the window and the value of the function at said point. Considering this, to detect R peaks the following steps were taken:

1. Noise reduction and baseline correction of the ECG with morphological filtering using morphological operators (for more details see [40]);
2. Calculation of the multiscale morphological transform for the processed signal;
3. Determination of the local minima that had its absolute amplitude larger than a threshold. This adaptive threshold was calculated with the transformed data's histogram.

In this study a 10-second window size was used with 1-sample increments and a scale value of 8.75 (0.035 $\times$ 250 Hz).

5.3.2 RR intervals

After the identification of the R peaks, the RR intervals, in which each point is given by

$$RR(i) = R_{peak}(i + 1) - R_{peak}(i), i\epsilon[1,n - 1]$$

where n is the number of R peaks, were calculated.

Considering that only R peak information is relevant for HRV, all data was downsampled from 250 Hz to 4 Hz [8].

5.4 Feature extraction

5.4.1 Extraction

Features were calculated using both time and frequency domain HRV analysis in a 60 sec sliding window (see section 3.3), shifted by 5 second increments. The 6 following time domain features were calculated:

- **Mean** mean of RR intervals;
- **SDNN** standard deviation of RR intervals;
- **SDSD** standard deviation of successive RR intervals differences;
- **RMSSD** square root of the mean squared differences between adjacent RR intervals;
- **NN50** number of consecutive RR intervals that differ more than 50 milliseconds;
- **pNN50** ratio between NN50 and the total number of NN intervals.

Regarding the frequency domain, 6 features were determined through an autoregressive PSD estimate using Burg's method of order 9 with a blackman window.

- **HF** power in high frequency range (0.15-0.4 Hz);
- **rHF** ratio of HF to the power in all frequency components (<= 0.4 Hz);
- **LF** power in low frequency range (0.04-0.15 Hz);
- **rLF** ratio of LF to the power in all frequency components (<= 0.4 Hz);
- **TVL** power difference between very low frequency range (<= 0.04 Hz) and power in all frequency components (<= 0.4 Hz);
- **LH** ratio of LF to HF.

5.4.2 Normalization

Considering the high inter-subject variance which is characteristic of these studies, as mentioned in [32], features calculated from the 2 types of segments ("resting" and "mental effort") were divided by the value of the mean or median for all parameters, depending on the normality of the data. The normality of the data was accessed using a one-sample Kolmogorov-Smirnov test (see [41]), of the respective feature in

the “resting” segment of the corresponding trial. This way, for every participant's trial all the features in both phases were normalized according to their mean/median value in the corresponding “resting” phase. Considering this, not only did this approach allowed inter-subject but also intra-subject analysis due to the fact that there were 3 ”resting” segments associated to the 3 ”mental effort” segments for each participant giving the possibility to compare data across trials (the same participant can have distinct baseline values in different circumstances therefore the need for a different reference for each task).

5.4.3 Transformation

The features extracted in the previous step were transformed using the following operations in order to increase the amount of information extracted from each segment: mean, median, standard deviation, maximum and minimum. Considering this, the final dataset has 156 instances (26 for every segment in every trial ($26 \times 2 \times 3$)) and 60 features (12 for every transform (12×5)).

5.5 Feature selection

To increase the interpretability of the classification models and improve the efficiency of the training processes and their generalization capability, the most important feature transformations were selected using 2 methods: principal components analysis (PCA) [42] and the normalized mutual information feature selection algorithm [43]. Principal component analysis focuses on reducing data dimensionality while preserving its "variability". To achieve this it finds new variables, named principal components, that are linear functions of the ones in the original data that maximize variance while being uncorrelated to each other [42]. This method had worse performances than the second one, therefore, the results and discussion are focused on the mutual information algorithm. The normalized mutual information feature selection algorithm aims to create a subset containing the most relevant features for the problem and simultaneously the least redundant ones. This algorithm improves on its non-normalized versions by introducing a entropy normalization factor that prevents the overshadowing of the redundancy aspect of the algorithm when compared to the relevance of the features when the group of selected features increases (see [43]).

Given two continuous random variables X and Y, mutual information can be defined as

$$I(X;Y) = \int \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$$

If the variables X and Y are discrete and random with alphabets χ and φ, respectively, mutual information can be rewritten as

$$I(X;Y) = \sum_{x \epsilon \chi} \sum_{y \epsilon \varphi} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

The normalized mutual information algorithm aims to maximize mutual information $I(C;S)$ between class variable C and the subset of selected features S. The normalized mutual information proposed in [43] is calculated as follows considering a initial set F with n features and set $S \subset F$ with k features:

$$NI(f_i;f_s) = \frac{I(f_i;f_s)}{\min\{H(f_i),H(f_s)\}}$$

where H is entropy defined as

$$H(f_s) = -\sum_{f_s \epsilon S} P(f_s) \log P(f_s)$$

This method also proposes the use of an average normalized mutual information as a measure of redundancy between the ith feature and the subset of selected features S:

$$\frac{1}{S} \sum_{f_s \epsilon S} NI(f_i; f_s)$$

that leads to a selection criteria that maximizes:

$$G = I(C; f_i) - \frac{1}{S} \sum_{f_s \epsilon S} NI(f_i; f_s)$$

The k values chosen in this work for the different study cases that are explained further were selected based on the features' normalized mutual information score (G). The number of selected features is based on the stabilization of these scores, therefore, when the speed of the decrease of these values decreases abruptly no more features are selected.

5.6 Case studies

In this thesis 3 case studies were explored:

- **Global** case study where the data from all 26 participants was used and the 4 different mental states, resting ("R") and "mental effort" during trial 1 ("C1"), 2 ("C2") and 3 ("C3"). One-against-one classification scheme ("R vs C1", "R vs C2", "R vs C3", "C1 vs C2", "C1 vs C3" and C2 vs C3") and one-against-all;
- **Proficiency** case study where the proficiency in programming of the participants was taken into account by dividing the 26 participants in 3 groups, "intermediate", "advanced" and "expert" composed by 11 participants in each of the first 2 groups and 4 in the last. One-against-one classification scheme and one-against-all multi-class classification scheme for all groups.;
- **Perception** case study where the information collected from the NASA's adapted questionnaire (see Appendix D.3) is used (in particular, the first question related to the mental effort quantification of the tasks) to relabel the extracted segments. With this information all "mental effort" data segments were relabelled according to the quantification given by the participants as "C1" if 1 or 2, "C2" if it was 3 or 4 and "C3" if it was 5 or 6, instead of the labels being defined by complexity metrics like in the previous case studies. Once again the one-against-all approach was used.

5.7 Classification and performance evaluation

The discrimination of the various predefined classes was performed using 2 methods: linear discriminant analysis (LDA) and support-vector machine (SVM). Once again, considering that SVM outperformed LDA, only the second one is presented on the results and discussion. Support-vector machine (SVM) models with a radial basis function kernel was used both in the binomial classifiers in the proof of concept study and in the one-against-all multi-class classification. The multi-class classification of the cognitive effort experienced during the analysis of each code snippet was performed as follows: data was relabelled for each of the classes to 1 (if it made part of the class) or to -1 (if it didn't); a SVM classifier was trained for each of these classes ("R","C1","C2","C3") according to the relabelling; performance was evaluated by giving all samples a score (predicted class posterior probability according to the trained classifiers) for each of the 4 classifiers; a sample's predicted class was determined by its highest score in the 4 classifiers.

The proposed methodologies were validated using a 10-fold cross-validation scheme. Here, the datasets were randomly partitioned in 10 equal size subsets and 9 subsets were used for training and the remaining was used for testing the classification model. This process was repeated 10 times corresponding to each of the 10 subsets.

In situations of data imbalance in the number of samples of each class, particularly relevant in the "Perception" case study where there was the relabelling of the "mental effort" data segments, the datasets were balanced by randomly selecting the same number of samples for all classes equal to the one with the least samples. This process was then repeated 20 times given this situation.

The performance of the models was evaluated by the average and standard deviation ($mean \pm SD$) of the following metrics: recall and precision. The recall defined by

$$\frac{TP}{(TP + FN)}$$

where TP are the true positives and FN the false negatives. The precision is defined by

$$\frac{TP}{(TP + FP)}$$

where FP are the false positives.

5.8 Correlation

To further study the impact of the "mental effort" perceived by participants on the results, the Spearman correlation test (see [41]) was used to test if the responses given from 1 to 6 in the questionnaires had a monotonous relationship with the selected features from the data of all participants in the "Global" case study. The same correlation test was done between the McCabe's cyclomatic metric (see [18]) values, displayed in Table 4.1, and the same features. To allow both these studies the selected features data was reduced to one dimension using principal component analysis (PCA).

The Spearman correlation test was used instead of Pearson correlation test due to the fact that both the responses given in the "Perception" questionnaire and the cyclomatic metric values weren't continuous variables therefore the need for a non-parametric test [41].

6

Results and discussion

As presented in section 4.3, several ECG signals were collected in the context of the presented study. For the present analysis, the selected ECG was recorded with Sensatron.

Through visual inspection of the electrocardiograms, it was verified that even though there were regions with small movement artefacts (shown in Figure 6.1 - fourth beat) none interfered with the R peaks (used to calculate HRV). Considering this, no pre-processing was made, rather then the usual baseline and noise removal, implemented in the algorithm provided by *LiNK* project.

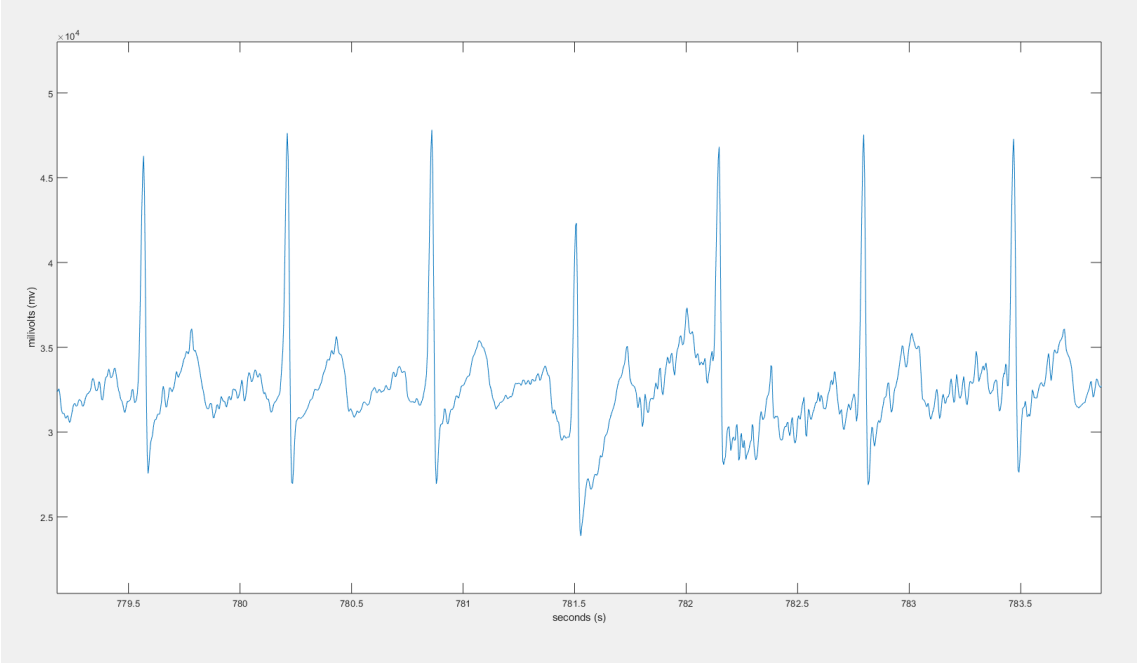


Figure 6.1: Segment of the ECG recorded from the third trial of participant 16 with noise.

After the segmentation of data, explained in section 5.2, R peak detection was achieved through MMD. In Figure 6.2 it's possible to see the ECG of the "mental

effort" segment of a participant's first trial with the respective R peaks in orange. The corresponding tachogram that resulted from extraction of the R peaks of the same signal can be seen in Figure 6.3.

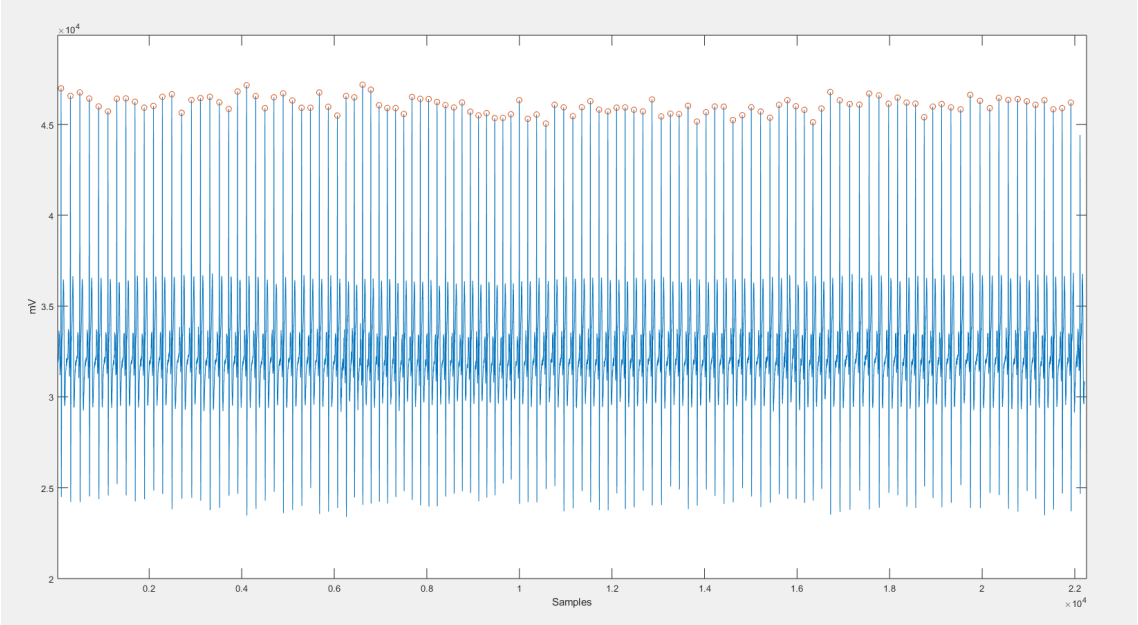


Figure 6.2: ECG segment correspondent to the "mental effort" phase of the first trial of participant 16. R peaks are identified in orange.

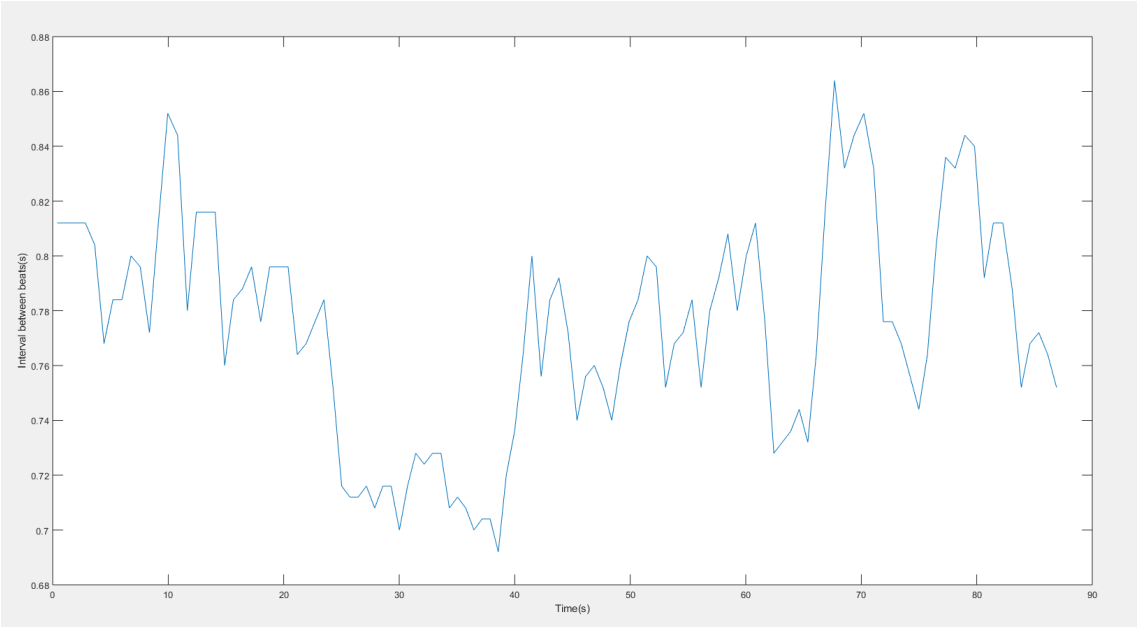


Figure 6.3: Variability of the heart rate over the "mental effort" phase of the first trial of participant 16.

Through visual exploration of the data some features displayed evidences of different cognitive stress levels across trials in the ”mental effort” segments as shown in Figure 6.4 where participant 16 demonstrated an overall increase in the mean of the RR intervals from trial 1 to 3. It should be noticed that trial 1 had minimal difficulty which made the participants have a much faster understanding of the code snippet. Considering that they had the option to end the analysis before the end of the predefined maximum task duration the signals acquired during this task are generally much shorter in duration, as explained in section 4.4.

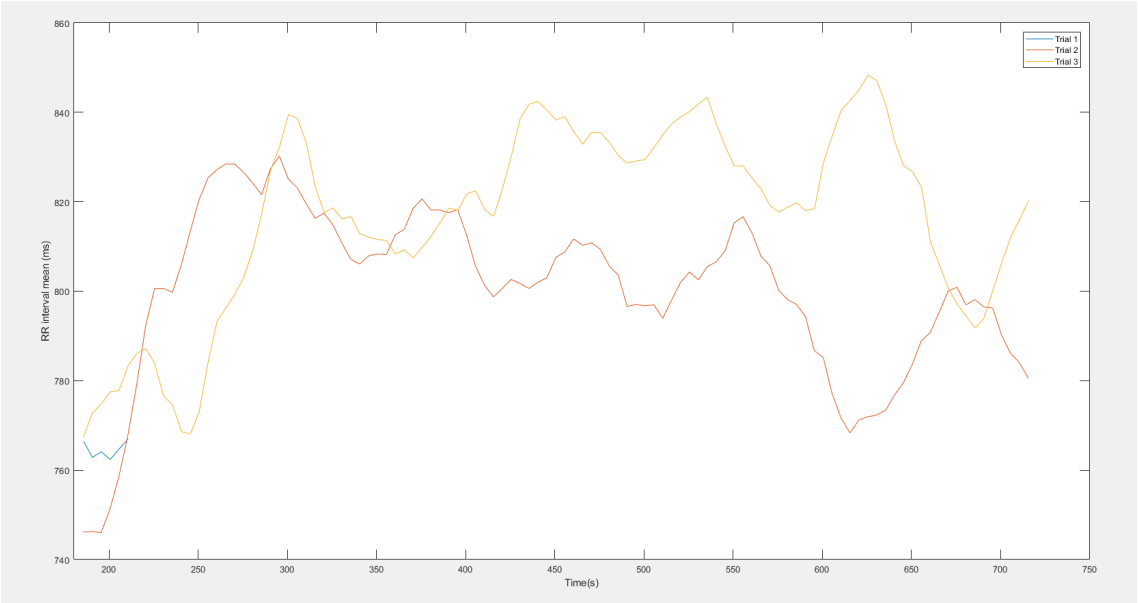


Figure 6.4: Mean of the RR intervals calculated in 60-second windows, every 5 seconds, over time for participant 16, during the “mental effort” segment in trial 1 (blue), 2 (orange) and 3 (yellow).

As explained in subsection 5.4.2 there is inter and intra-subject variance associated with HRV studies that is noticeable in Figure 6.5. It is possible to see different resting baselines for the same person across trials and also overall distinct ”resting” values for participant 1 (650-750 ms) and participant 16 (800-900 ms).

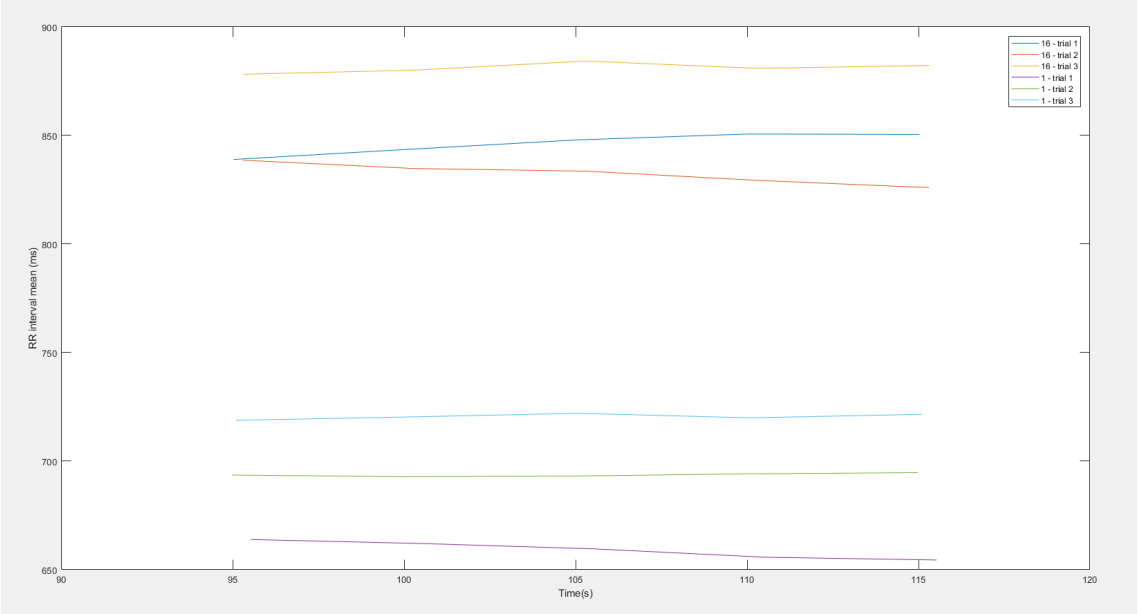


Figure 6.5: Mean of the RR intervals calculated in 60-second windows, every 5 seconds, over time for participant 16, during the "resting" segment in trial 1 (dark blue), 2 (orange) and 3 (yellow) and for participant 1 in trial 1 (magenta), 2 (green) and 3 (light blue).

In Table 6.1 the features selected through the normalized mutual information selection algorithm for the 3 case studies with the respective scores are presented. It is possible to see that overall the "Global" (first selected feature score = 0.82) and "Perception" (first selected feature score = 0.81) case studies and also the "Advanced" category from the "Proficiency" case study (first selected feature score = 0.52) have selected features with good scores. The "Intermediate" (first selected feature score = 0.27) and specially the "Expert" (first selected feature score = 0.07) categories from the "Proficiency" case studies produced bad scores, which translate in relatively bad performances shown in Table 6.3 that are discussed further.

Global		Proficiency						Perception	
		Intermediate		Advanced		Expert			
Feature name	Score	Feature name	Score	Feature name	Score	Feature name	Score	Feature name	Score
min Mean	0,82	mean Mean	0,27	min Mean	0,49	mean TVL	0,07	min Mean	0,81
mean Mean	0,73	max LH	0,27	mean Mean	0,37	mean HF	0,04	mean Mean	0,73
median Mean	0,69	min Mean	0,26	median Mean	0,36	mean RMSSD	0,04	median Mean	0,72
median SDNN	0,64	median Mean	0,26	max Mean	0,31	sd rLF	0,04	median pNN50	0,69
max pNN50	0,61	min rHF	0,21	median NN50	0,25	mean SDSD	0,04	median NN50	0,68
sd Mean	0,61	max TVL	0,18	min pNN50	0,17	mean SDNN	0,03	median SDNN	0,64
max NN50	0,57	-	-	-	-	mean rLF	0,03	max pNN50	0,63
median pNN50	0,55	-	-	-	-	sd SDSD	0,02	max NN50	0,58
-	-	-	-	-	-	-	-	sd Mean	0,56
-	-	-		-	-	-	-	min SDNN	0,52

Table 6.1: Scores of the selected transformed features through the normalized mutual information feature selection algorithm for the 3 case studies.

In Table 6.2 the initial results show that the "resting" state is easily distinguishable from the 3 states of "mental effort" (overall performance results close to 1) and that the "mental effort" produced in trial 1 can be differentiated from the "mental effort" produced in trial 2 (recall = 0.95 ± 0.16, precision = 0.93 ± 0.17) and 3 (recall = 0.85 ± 0.25, precision = 0.98 ± 0.08). "Mental effort" can't be distinguished between trial 2 and 3 (recall = 0.40 ± 0.40, precision = 0.42 ± 0.32). These results are concordant with the ones obtained with the one-against-all approach in Table 6.3 where "Mental effort" in trial 2 (recall = 0.26 ± 0.35, precision = 0.41 ± 0.31) and 3 (recall = 0.50 ± 0.40, precision = 0.46 ± 0.21) are difficult to identify correctly while the "mental effort" in trial 1 (recall = 0.79 ± 0.26, precision = 0.82 ± 0.22) and "resting" state across all trials (recall = 0.99 ± 0.06, precision = 0.85 ± 0.21) can be identified with high accuracy. Previous studies on the area also achieved good results when trying to distinguishing a mental task from a "resting" state ([8]) and even different mental tasks ([11]).

	R vs C1	R vs C2	R vs C3	C1 vs C2	C1 vs C3	C2 vs C3
Recall	0.98 ± 0.10	0.99 ± 0.04	0.99 ± 0.06	0.95 ± 0.16	0.85 ± 0.25	0.40 ± 0.40
Precision	0.94 ± 0.12	0.99 ± 0.04	1.00 ± 0.00	0.93 ± 0.17	0.98 ± 0.08	0.42 ± 0.32

Table 6.2: Performance of the binomial classifiers for the "Global" case study as $mean \pm SD$.

Results from the "Proficiency" case study in Table 6.3 show that while the "Advanced" group of participants had similar results to the previous study where all participants data was used, the "Intermediate" and "Expert" groups didn't. The "Expert" group results show, once again difficulty in classifying "Mental effort" in trial 2 (recall = 0.23 ± 0.42, precision = 0.29 ± 0.31) and 3 (recall = 0.19 ± 0.39, precision = 0.22 ± 0.28) but also in trial 1 (recall = 0.33 ± 0.47, precision = 0.40 ± 0.35). Overall the results for this group have low recall, precision and big standard deviations (sometimes bigger then the mean values themselves) and even the "resting" state classification had relativity low performance when compared with the previous results. One of the main reasons of this poor performance can be the fact that there were only 4 individuals in this category which made the classifier not have sufficient data for a proper training. This assumption is supported by the scores from the "Expert" category in Table 6.1 where all values are low (lower order of magnitude)(e.g., first selected feature score = 0.07), especially when compared with the scores of the other groups and studies (that range from 0.27 in the "Intermediate" category to 0.82 in the "Global" category for the first select feature score), and show that the selected features weren't relevant and/or non-redundant. Regarding the "Intermediate" group, the "mental effort" in trial 2 (recall = 0.19 ± 0.39,

precision = 0.35 ± 0.35) and 3 (recall = 0.25 ± 0.43, precision = 0.33 ± 0.32) still showed poor results and the "mental effort" in trial 1 (recall = 0.57 ± 0.49, precision = 0.60 ± 0.38), while better classified, still had worse results relative to the "Advanced" group (recall = 0.91 ± 0.28, precision = 0.92 ± 0.20). The "resting" state results (recall = 0.99 ± 0.10, precision = 0.72 ± 0.29) were also more in line with the previous study.

With the increase in a participant's programming proficiency it was expected that results were progressively worse assuming that less "mental effort" would be required to understand the different codes. It was expected that the similarity of the extracted feature transform values across the four states would be greater and therefore, making it harder to distinguish the 3 "mental effort" states. This assumption was not supported by the results presented in Table 6.3 that show clearly better performance for the "Advanced" group when compared to the "Intermediate". One aspect that could have contributed to this was the screening of the participants that may have included participants with "low" or less than "intermediate" proficiency in programming making the "Intermediate" group less concise in the programmers' skill than it should. Another aspect that can be conjectured about this unexpected results is that proficiency isn't necessarily related to the "mental effort" exerted during this tasks and maybe speed, for example, could be a better indicator.

		Global	Proficiency			Perception
			Intermediate	Advanced	Expert	
R	Recall	0.99 ± 0.06	0.99 ± 0.10	0.99 ± 0.07	0.75 ± 0.44	0.98 ± 0.11
	Precision	0.85 ± 0.21	0.72 ± 0.29	0.80 ± 0.28	0.65 ± 0.37	0.87 ± 0.20
C1	Recall	0.79 ± 0.26	0.57 ± 0.49	0.91 ± 0.28	0.33 ± 0.47	0.73 ± 0.31
	Precision	0.82 ± 0.22	0.60 ± 0.38	0.92 ± 0.20	0.40 ± 0.35	0.82 ± 0.25
C2	Recall	0.26 ± 0.35	0.19 ± 0.39	0.39 ± 0.49	0.23 ± 0.42	0.33 ± 0.38
	Precision	0.41 ± 0.31	0.35 ± 0.35	0.44 ± 0.29	0.29 ± 0.31	0.50 ± 0.32
C3	Recall	0.50 ± 0.40	0.25 ± 0.43	0.26 ± 0.43	0.19 ± 0.39	0.57 ± 0.41
	Precision	0.46 ± 0.21	0.33 ± 0.32	0.47 ± 0.35	0.22 ± 0.28	0.54 ± 0.26

Table 6.3: Performance of the one-against-all classifiers for the 3 case studies as $mean \pm SD$.

When it comes to the "Perception" case study the responses given by the participants in D.3 pointed once again the similarity of "mental effort" states between trial 2 and 3 evident in the poor classification of "mental effort" in those trials in the previous results: 16 out of 26 participants considered their "mental effort" in trial 2 high and 9 in 26 considered their "mental effort" in trial 3 medium. The new sample distribution according this data is shown in table 6.4.

	Global/ Proficiency	Perception
C1	26	24
C2	26	21
C3	26	33

Table 6.4: Number of samples in the "mental effort" classes with the class labels given by the codes (software engineering complexity metrics) in the "Global" and "Proficiency" case studies and given by the perceived "mental effort" from the participants in the "Perception" case study.

Regarding the "Perception" results showed in Table 6.3 the classification of the "mental effort" in trial 2 (recall = 0.33±0.38, precision = 0.50±0.32) and 3 (recall = 0.57±0.41, precision = 0.54±0.26) show clear improvements relative to the "Global" case study (where all participants' data was also used) which is in line with what was expected considering that those were the most relabelled instances. On the other hand, the classification of the trial 1 (recall = 0.73 ± 0.31, precision = 0.82 ± 0.25) had a worse performance. One possible reason to this performance decrease in trial 1 is the fact that 2 participants classified the "mental effort" displayed in trial 1 as "medium" when the code was clearly less complex than the other 2. This example shows another possible scenario that must be taken into account, that is, the subject perceived complexity (answered in the questionnaire), can differ from the physiological response of the subject, captured in the signals collected during the study. Considering this, with the shown results it can be said that the complexity of the codes was better described by the "mental effort" displayed in the trials through the data and even through the perception of the participants in most of the cases then through software engineering complexity metrics.

The selected features across the 3 studies (see Table 6.1) were mainly time domain features with more emphasis on the RR interval mean, NN50 and pNN50. Mean and pNN50 also displayed significant differences in differentiating "rest" and "mental task" conditions in [11]. On the other hand, frequency domain features didn't prove themselves to be relevant in this analysis given that they were mainly selected in the "Expert" group of the "Proficiency" study where the classifier performance was worst. Despite this, LF is selected in some cases in Table 6.5 which were also used in [8] for a stress related study. This type of features could maybe play a more important role if a 5-minute window (as discussed in 3.3) was used for HRV analysis.

Considering the similarity between the "mental effort" exerted in trial 2 and 3 suggested from the previous results, the methodology was repeated but with only 3 classes: "resting" state (R), "mental effort" in trial 1 (C1) and "mental effort" in trial

2 and 3 (C4). Results are shown in Tables 6.5 and 6.6. With this class change some frequency domain features were selected in the better performing classifiers which can be due to the fact that they weren't relevant when trying to distinguish higher complexity codes. Overall, given the supposed similarity between the "mental effort" done in trial 2 and 3, the performances of the classifiers were better as expected.

Global		Proficiency						Perception	
		Intermediate		Advanced		Expert			
Feature name	Score	Feature name	Score	Feature name	Score	Feature name	Score	Feature name	Score
median NN50	0,87	max TVL	0,52	min Mean	0,85	mean TVL	0,07	median NN50	0,91
min Mean	0,76	min Mean	0,52	mean Mean	0,62	mean SDSD	0,04	median pNN50	0,78
median pNN50	0,77	max LH	0,49	median Mean	0,61	mean SDNN	0,03	min Mean	0,76
mean Mean	0,68	min HF	0,47	max Mean	0,52	mean RMSSD	0,03	median Mean	0,73
median Mean	0,68	mean Mean	0,45	min SDSD	0,48	mean HF	0,03	mean Mean	0,70
median SDNN	0,63	sd Mean	0,45	-	-	sd rLF	0,03	median LF	0,64
median LF	0,60	-	-	-	-	mean rLF	0,02	median SDNN	0,63
-	-	-	-	-	-	-	-	max pNN50	0,61
-	-	-	-	-	-	-	-	max NN50	0,58
-	-	-	-	-	-	-	-	median TVL	0,56

Table 6.5: Scores of the selected transformed features through the normalized mutual information feature selection algorithm for the 3 case studies with the new class C4 (conjunction of C2 and C3).

		Global	Proficiency			Perception
			Intermediate	Advanced	Expert	
R	Recall	0.97 ± 0.11	0.98 ± 0.14	1.00 ± 0.00	0.73 ± 0.45	0.99 ± 0.07
	Precision	0.93 ± 0.13	0.86 ± 0.23	0.98 ± 0.07	0.61 ± 0.31	0.94 ± 0.12
C1	Recall	0.79 ± 0.27	0.50 ± 0.49	0.91 ± 0.28	0.40 ± 0.49	0.76 ± 0.29
	Precision	0.95 ± 0.12	0.82 ± 0.34	0.98 ± 0.09	0.75 ± 0.32	0.89 ± 0.19
C4	Recall	0.94 ± 0.15	0.80 ± 0.39	0.96 ± 0.19	0.53 ± 0.50	0.87 ± 0.23
	Precision	0.91 ± 0.15	0.78 ± 0.29	0.96 ± 0.14	0.61 ± 0.32	0.87 ± 0.17

Table 6.6: Performance of the one-against-all classifiers for the 3 case studies as $mean \pm SD$ with the new class C4 (conjunction of C2 and C3).

Aiming to access how HRV correlates with the perceived "mental effort" and software engineering metrics, the feature space containing the best features was submitted to a PCA and the principal component was extracted. The agreement of the extracted principal component was accessed by calculating its correlation with the perception classes defined a posteriori and also with the cyclomatic metrics of the code snippets, which are presented in Table 6.7.

Both p values are lower than the significance level of 0.05 so the null hypothesis that there is no correlation is rejected. The rho coefficients show that while not strong, there is a negative monotonous relation between both variables and the principal component of the features' space (also visible in Figures 6.6 and 6.7). Both values were higher in the correlation between the perceived "mental effort" data and

	rho	p value
Perception	-0,3877	0,0005
Cyclomatic metric	-0,3334	0,0029

Table 6.7: Results of the Spearman correlation tests for the study of the correlation between the "mental effort" participants' perception and the selected features in the "Global" case study in Table 6.1 and between the cyclomatic metric values of the 3 codes (Table 4.1) and the same features. *rho* - Spearman's coeficient and p value testing the hypothesis of no correlation against the alternative hypothesis of a nonzero correlation.

the features, which show a higher confidence (p value = 0.0005) and a stronger correlation ($rho = -0.3877$) when compared to the cyclomatic metric. This was expected considering that the cyclomatic metric had a big difference between the code 2 and 3 which displayed similarity according to the previous "mental effort" results. Nonetheless, the results from the perceived "mental effort" correlation were expected to be better. This could be due to the fact that some participants can't accurately classify their exerted "mental effort" across the trials and instead answered to the questionnaire in terms of the expected "mental effort" necessary to complete the task instead of the actual "mental effort" made.

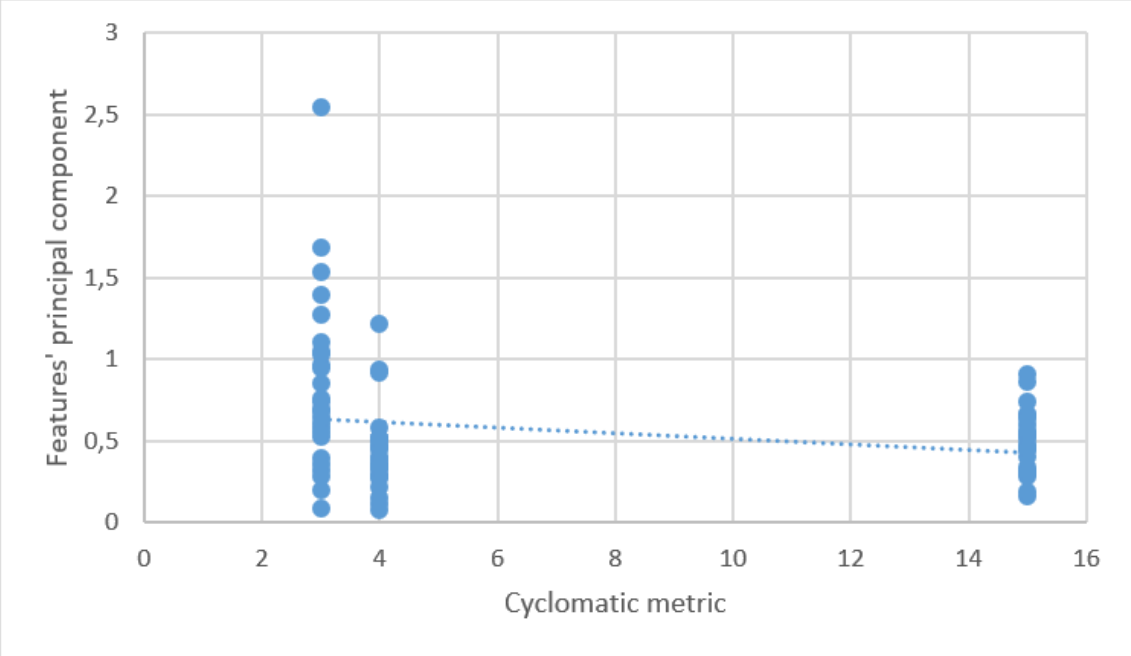


Figure 6.6: Plot of the principal component extracted from the selected data through PCA (77.7% of the space explained) against the values of the cyclomatic metric across the 3 codes with the linear tendency.

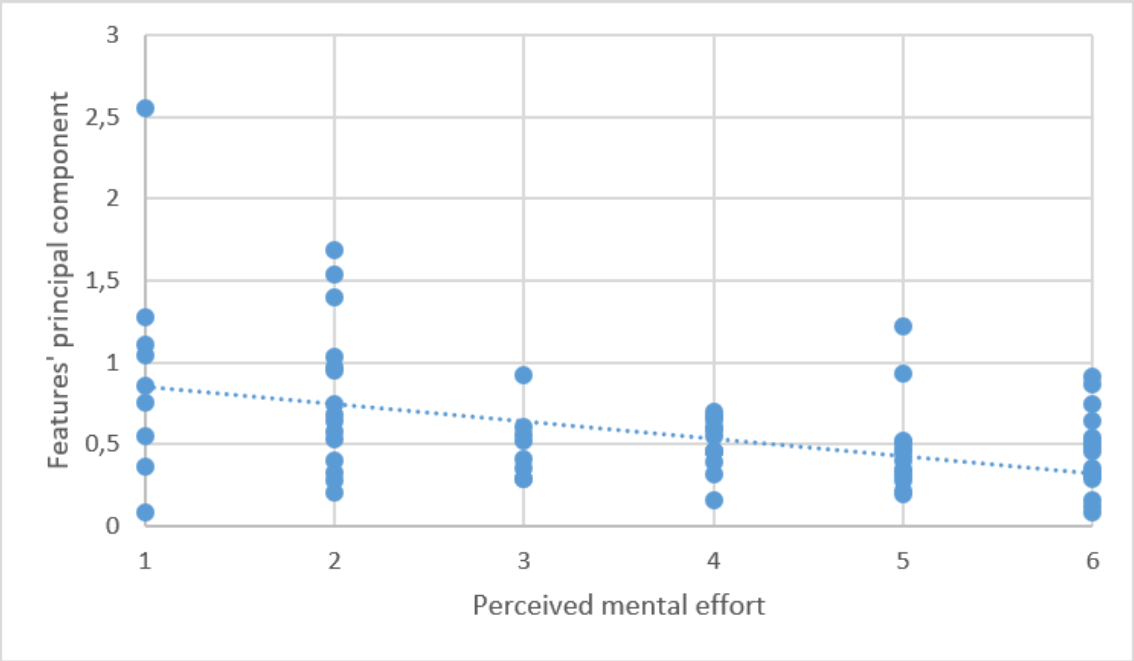


Figure 6.7: Plot of the principal component extracted from the selected data through PCA (77.7% of the space explained) against the values of the "mental effort" perceived by participants for all 3 codes with the linear tendency.

Once again the perceived "mental effort" showed itself to be more representative of the actual "mental effort" when compared to a very well known software engineering complexity metric.

7

Future work and applications

The main results show that it is possible to distinguish the 3 different mental states with the use of non-invasive sensors, which opens a range of possibilities for future studies and applications in real life settings. Programmers can use, for example, a *smartwatch* that measures heart rate while they are programming and with the collected data and a system with optimized algorithms for detection or maybe even prediction, high "mental effort" states could be detected (this process would involve the measurement of a baseline reference similar to the procedure done in this thesis) and warnings could be displayed to the programmer highlighting the need to pay a closer attention to a higher complexity task at hand or even make a pause to avoid mistakes. This type of information, in association with the software engineering complexity metrics could be used to create biofeedback improved models of bug density estimation and software risk analysis.

Another aspect that could be studied but would require sensors less suited for a real life setting but still non-invasive would be the analysis of the brain activity involved in bug creation and discovery and its connection to physiological manifestations driven by the ANS. These types of studies could be useful for the better understanding of conditions associated with making bugs or bugs escaping human attention. Functional magnetic resonance imaging could play an important role in this area.

In this thesis the "mental effort" was studied as a whole along the segments from different participant's trial. With the association of the eye tracking information that was also recorded in this study, the "mental effort" in different code sections among the trial could be explored based on the the analysis of the participants' eyes movements or even the time they spent staring at a specific code region. Specific HRV features could also be tracked and studied over time which was something that was also left for futures analysis.

Lastly, the information gathered in Nasa-TLX adapted questionnaire could also

be better explored regarding different indicators (like pressure with time) and the weight of each component according to the different participants instead of only focusing on the perceived "mental effort".

8

Conclusions

Software development is essentially a human activity, therefore, most of the software produced today is still the result of an intensive human made process. In this work it was shown that it is possible to distinguish 3 main mental states: "resting" and 2 different "mental effort" phases (trial 1 and trials 2/3) using HRV features captured from physiological manifestations using non-invasive wearable devices compatible with typical software development environments. Despite this, according to one of the most well known software engineering complexity metrics - McCabe's cyclomatic metric - 3 "mental effort" states were defined and the complexity of the code snippet in trial 3 was supposed to be considerably higher than the others, which isn't consistent with the HRV results. According to the information gathered regarding participants perception of "mental effort" during the 3 trials it was also given the idea that globally the participants found the complexity of the code snippet displayed during trials 2 and 3 to be similar. These results, associated with the fact that the HRV data correlated better with the participants' perceived "mental effort" than with the cyclomatic metric, show that software engineering complexity metrics have flaws and that biofeedback is a very promising research avenue that can enhance software development paradigms.

Regardless of the overall conclusive results, the study of the impact of proficiency in "mental effort" didn't produce the expected results. "Mental effort" was supposed to be harder to distinguish with the increase of participants' programming skills considering that it would come closer to the "resting" state values which wasn't verified by the present analysis.Considering this, participants' proficiency category should have been attributed with a higher level of confidence, therefore, more indicators should have been used to differentiate the proficiency levels. The recruitment of more "expert" level programmers should also be taken into account considering the mismatch between the numbers of participants in the first 2 categories when compared to the last.

The results regarding the perceived "mental effort" were also worse than expected specially the agreement given by the Spearman test between participants' perceived "mental effort" and the principal component of the features' space. This can be due to the fact that some of the participants were not able to translate their exerted "mental effort" into a classification in a scale from 1 to 6 or, on the other hand, the given answers were related with "mental effort" expected by the subject to be required to perform the task instead of the actual "mental effort" exerted by the subject.

Lastly, the duration of the experience, specially regarding the duration of certain segments, should have been longer. Due to this, features regarding frequency information were extracted with higher uncertainty and 4 participants had to be excluded from the study due to their understanding of the first trial's code in less than 1 minute (size of the window used for HRV analysis).

Bibliography

[1] P. Low, "Overview of the autonomic nervous system." https://www.msdmanuals.com/home/brain,-spinal-cord,-and-nerve-disorders/autonomic-nervous-system-disorders/overview-of-the-autonomic-nervous-system, 2018. Accessed: 2018-03-12.

[2] J. T. Ottesen, "Modelling of the baroreflex-feedback mechanism with time-delay," *Journal of mathematical biology*, vol. 36, no. 1, pp. 41–63, 1997.

[3] G. J. Tortora and B. H. Derrickson, *Principles of anatomy and physiology*. John Wiley & Sons, 2008.

[4] R. Klabunde, *Cardiovascular physiology concepts*. Lippincott Williams & Wilkins, 2011.

[5] U. R. Acharya, K. P. Joseph, N. Kannathal, L. C. Min, and J. S. Suri, "Heart rate variability," in *Advances in cardiac signal processing*, pp. 121–165, Springer, 2007.

[6] S. McConnell, "Code complete: A practical handbook of software construction (2nd edition)," 2004.

[7] K. Makarla, "Report: Software failures cost 1.1 trillion in 2016." http://servicevirtualization.com/report-software-failures-cost-1-1-trillion-2016/, 2017. Accessed: 2018-04-22.

[8] S. Cerutti, A. M. Bianchi, and H. Reiter, "Analysis of sleep and stress profiles from biomedical signal processing in wearable devices," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 6530–6532, IEEE, 2006.

[9] S. Chen, J. Epps, N. Ruiz, and F. Chen, "Eye activity as a measure of human mental effort in hci," in *Proceedings of the 16th international conference on Intelligent user interfaces*, pp. 315–318, ACM, 2011.

[10] F. D. Purves D, Augustine GJ, "Autonomic regulation of cardiovascular function," 2001.

[11] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel, "Influence of mental stress on heart rate and heart rate variability," in *4th European conference of the international federation for medical and biological engineering*, pp. 1366–1369, Springer, 2009.

[12] G. E. Prinsloo, H. L. Rauch, M. I. Lambert, F. Muench, T. D. Noakes, and W. E. Derman, "The effect of short duration heart rate variability (hrv) biofeedback on cognitive performance during laboratory induced cognitive stress," *Applied Cognitive Psychology*, vol. 25, no. 5, pp. 792–801, 2011.

[13] S. Laborde, E. Mosley, and J. F. Thayer, "Heart rate variability and cardiac vagal tone in psychophysiological research–recommendations for experiment planning, data analysis, and data reporting," *Frontiers in psychology*, vol. 8, p. 213, 2017.

[14] G. K. Saha, "Software fault avoidance issues," *Ubiquity*, vol. 2006, no. November, p. 5, 2006.

[15] D. Huizinga and A. Kolawa, *Automated defect prevention: best practices in software management*. John Wiley & Sons, 2007.

[16] S. Balaji and M. S. Murugaiyan, "Waterfall vs. v-model vs. agile: A comparative study on sdlc," *International Journal of Information Technology and Business Management*, vol. 2, no. 1, pp. 26–30, 2012.

[17] E. Valido-Cabrera, "Software reliability methods," *Report, Technical University of Madrid, Agustus*, 2006.

[18] T. J. McCabe, "A complexity measure," *IEEE Transactions on software Engineering*, no. 4, pp. 308–320, 1976.

[19] B. Curtis, S. B. Sheppard, P. Milliman, M. Borst, and T. Love, "Measuring the psychological complexity of software maintenance tasks with the halstead and mccabe metrics," *IEEE Transactions on software engineering*, no. 2, pp. 96–104, 1979.

[20] E. J. Weyuker, "Evaluating software complexity measures," *IEEE transactions on Software Engineering*, vol. 14, no. 9, pp. 1357–1365, 1988.

[21] A. Baddeley, "Working memory," *Science*, vol. 255, no. 5044, pp. 556–559, 1992.

[22] A. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with eeg pattern recognition methods," *Human factors*, vol. 40, no. 1, pp. 79–91, 1998.

[23] D. Purves, G. Augustine, D. Fitzpatrick, L. Katz, A. LaMantia, J. McNamara, and S. Williams, "Neuroscience 2nd edition. sunderland (ma) sinauer associates," 2001.

[24] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *European heart journal*, vol. 17, no. 3, pp. 354–381, 1996.

[25] D. Bansal, M. Khan, and A. Salhan, "A review of measurement and analysis of heart rate variability," in *Computer and Automation Engineering, 2009. ICCAE'09. International Conference on*, pp. 243–246, IEEE, 2009.

[26] M. Brennan, M. Palaniswami, and P. Kamen, "Do existing measures of poincare plot geometry reflect nonlinear features of heart rate variability?," *IEEE transactions on biomedical engineering*, vol. 48, no. 11, pp. 1342–1347, 2001.

[27] S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. Berger, and R. J. Cohen, "Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control," *science*, vol. 213, no. 4504, pp. 220–222, 1981.

[28] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Trans. Biomed. Eng*, vol. 32, no. 3, pp. 230–236, 1985.

[29] M. De Rivecourt, M. Kuperus, W. Post, and L. Mulder, "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight," *Ergonomics*, vol. 51, no. 9, pp. 1295–1319, 2008.

[30] A. Johnson and A. Widyanti, "Cultural influences on the measurement of subjective mental workload," *Ergonomics*, vol. 54, no. 6, pp. 509–518, 2011.

[31] S. Mukherjee, R. Yadav, I. Yung, D. P. Zajdel, and B. S. Oken, "Sensitivity to mental effort and test–retest reliability of heart rate variability measures

in healthy seniors," *Clinical Neurophysiology*, vol. 122, no. 10, pp. 2059–2066, 2011.

[32] D. S. Quintana and J. A. Heathers, "Considerations in the assessment of heart rate variability in biobehavioral research," *Frontiers in psychology*, vol. 5, p. 805, 2014.

[33] T. Riniolo and S. W. Porges, "Inferential and descriptive influences on measures of respiratory sinus arrhythmia: sampling rate, r-wave trigger accuracy, and variance estimates," *Psychophysiology*, vol. 34, no. 5, pp. 613–621, 1997.

[34] R. Couceiro, P. Carvalho, R. Paiva, J. Henriques, I. Quintal, M. Antunes, J. Muehlsteff, C. Eickholt, C. Brinkmeyer, M. Kelm, *et al.*, "Assessment of cardiovascular function from multi-gaussian fitting of a finger photoplethysmogram," *Physiological measurement*, vol. 36, no. 9, p. 1801, 2015.

[35] L. . Calendas, "Lenda de beatriz e o mouro (almourol)." `http://lendasecalendas.forumeiros.com/t121-lenda-de-beatriz-e-o-mouro-almourol`, 2009. Accessed: 2018-03-29.

[36] L. . Calendas, "Lenda da caninha verde (vouzela)." `http://lendasecalendas.forumeiros.com/t138-lenda-da-caninha-verde-vouzela?highlight=lenda+da+caninha+verde`, 2009. Accessed: 2018-03-29.

[37] L. . Calendas, "Lenda da serra do nó (viana do castelo." `http://lendasecalendas.forumeiros.com/t129-lenda-da-serra-do-no-viana-do-castelo?highlight=lenda+da+caninha+verde`, 2009. Accessed: 2018-03-29.

[38] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*, vol. 52, pp. 139–183, Elsevier, 1988.

[39] Y. Sun, K. L. Chan, and S. M. Krishnan, "Characteristic wave detection in ecg signal using morphological transform," *BMC cardiovascular disorders*, vol. 5, no. 1, p. 28, 2005.

[40] Y. Sun, K. L. Chan, and S. M. Krishnan, "Ecg signal conditioning by morphological filtering," *Computers in biology and medicine*, vol. 32, no. 6, pp. 465–479, 2002.

[41] J.-B. Du Prel, B. Röhrig, G. Hommel, and M. Blettner, "Choosing statistical tests: part 12 of a series on evaluation of scientific publications," *Deutsches Ärzteblatt International*, vol. 107, no. 19, p. 343, 2010.

[42] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[43] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

Appendices

A

Informed consent

FORMULÁRIO DE INFORMAÇÃO E CONSENTIMENTO INFORMADO

TÍTULO DO PROJECTO DE INVESTIGAÇÃO:

Engenharia de Software Potenciada com Bio-informação

PROTOCOLO	EEG e sensores fisiológicas (ECG, EDA e PPG)
PROMOTOR	Instituto de Ciências Nucleares Aplicadas à Saúde (ICNAS) e Departamento de Informática (DEI) da Universidade de Coimbra (UC)
INVESTIGADOR COORDENADOR	Professor Doutor Miguel Castelo-Branco e Professor Doutor Henrique Madeira
CENTRO DE ESTUDO	ICNAS, UC
INVESTIGADOR PRINCIPAL	Professor Doutor Miguel Castelo-Branco
MORADA	Azinhaga Santa Comba, Celas - 3000-548 Coimbra
CONTACTO TELEFÓNICO	239 480 200

NOME DO PARTICIPANTE
(LETRA DE IMPRENSA) _____

É convidado a participar neste estudo porque tem experiência como programador de *software*. Independentemente do tempo dedicado a esta prática e do seu grau de experiência, essa é a característica de interesse naqueles(as) que, voluntariamente, participarão neste estudo.

Este procedimento é chamado **Consentimento Informado** e descreve a finalidade do estudo, os procedimentos e os possíveis benefícios e riscos. A sua participação poderá contribuir para melhorar o conhecimento de processos cognitivos envolvidos na tarefa de compreensão de programas de *software* (como exemplo de tarefas complexas e abstratas) e as manifestações fisiológicas correlacionadas com esses processos.

Receberá uma cópia deste Consentimento Informado para rever e/ou solicitar aconselhamento de familiares e amigos. Os investigadores do estudo irão esclarecer qualquer dúvida que tenha sobre o termo de consentimento e também alguma palavra ou informação que possa não entender.

Deve tomar a decisão de participar ou não no estudo depois de entendê-lo e de não ter qualquer dúvida acerca do mesmo. Caso queira participar, ser-lhe-á solicitado que assine e date este formulário. Após a sua assinatura e a dos investigadores, ser-lhe-á entregue uma cópia. Caso não queira participar, não haverá qualquer penalização para o voluntário.

1. INFORMAÇÃO GERAL E OBJECTIVOS DO ESTUDO

Este estudo irá decorrer no Instituto Biomédico de Investigação da Luz e da Imagem (IBILI) da Faculdade de Medicina da Universidade de Coimbra, e tem como objetivo estudar os circuitos neuronais envolvidos na tarefa de compreensão de programas de *software* (como exemplo de tarefas complexas e abstratas) e as manifestações fisiológicas correlacionadas com esses processos. Em particular, pretende-se investigar se é possível determinar através de sinais que caracterizam manifestações fisiológicas, capturados por sensores não intrusivos e compatíveis com os ambientes de desenvolvimento de *software*, se o programador está perante um excerto de software pouco complexo ou muito complexo, considerando métricas de complexidade de software tradicionais e aspetos como a natureza do código (iterativo ou recursivo).

Trata-se de um estudo observacional, pelo que não será feita nenhuma alteração nas suas rotinas diárias ou tratamentos habituais.

Este estudo foi aprovado pela Comissão de Ética da FMUC de modo a garantir a proteção dos direitos, segurança e bem-estar de todos os participantes e garantir prova pública dessa proteção.

Como participante neste estudo, beneficiará da vigilância institucional e encaminhamento adequado garantindo assim a sua segurança.

Este estudo é constituído por 1 visita que se estima que tenha uma duração máxima de 1 hora e 30 minutos. Serão incluídos cerca de 40 participantes considerados indivíduos com mais de 18 anos, com experiência de programação em Java e de qualquer sexo.

2. PROCEDIMENTOS E CONDUÇÃO DO ESTUDO

2.1. Procedimentos

Se fizer parte deste estudo, ser-lhe-á solicitado que colabore na realização de um conjunto de testes não invasivos. Estes procedimentos serão realizados por técnicos especializados que integram a equipa de investigação deste estudo e serão realizados no IBILI.

O estudo inclui os seguintes dispositivos/técnicas:

- Gravação de EEG (eletroencefalograma) e ECG (eletrocardiograma)
- Gravação de EDA (atividade eletrocutânea)
- Gravação de PPG (fotopletismografia)
- Gravação do movimento do olhar e pupilografia.

2.2. Calendário das visitas/ Duração

Este estudo é constituído por 1 visita com duração não superior a 1 hora e 30 minutos. Segue-se uma descrição do estudo:

Descrição dos Procedimentos

Serão realizados os seguintes procedimentos/exames realizados no Instituto de Ciências Nucleares Aplicadas à Saúde, ICNAS, da Universidade de Coimbra:

Eletroencefalograma

A eletroencefalografia é uma técnica que permite gravar em tempo real a atividade elétrica cerebral, através da colocação de elétrodos na cabeça.

Eletrocardiograma

O eletrocardiograma permite que se estude as variações do ritmo cardíaco. Para isso, serão colocados elétrodos no peito.

Atividade eletrocutânea

Para gravar a atividade eletrocutânea serão colocados dois sensores nos seus dedos ou mãos. Esta atividade pode estar relacionada com o esforço mental e é por isso que gravamos ao longo da tarefa.

Gravação do movimento do olhar e pupilografia

A informação do movimento do olhar permite saber para que zona do ecrã estava a olhar em cada

momento. É fundamental sabermos que parte do programa está a analisar enquanto fazemos os registos dos sinais fisiológicos acima descritos.

A pupilografia permite gravar o aumento e a diminuição da pupila ao longo do tempo. Esta medida pode refletir a atenção e o esforço mental.

Estas duas medições são feitas através de uma câmara colocada junto ao ecrã do computador e que regista apenas os movimentos do olho e o tamanho da pupila.

Tarefa

Vamos-lhe pedir que analise três pequenos excertos de programas (um de cada vez), com complexidades bastante diferentes, e que os tente compreender. Enquanto o faz, são gravados no computador os sinais acima descritos e os passos da tarefa que está a realizar. Isto permitirá posteriormente estudar que ritmos cerebrais estão relacionados com a complexidade de cada um dos programas que analisou, elucidando-nos sobre o funcionamento do cérebro e o reflexo disso nestes sinais fisiológicos.

No final do tempo de observação de cada programa, o voluntário responde a um pequeno conjunto de perguntas para aferir se se ele compreendeu efetivamente cada programa.

2.3. Tratamento de dados/Randomização

Os participantes serão todos alocados ao mesmo grupo. Serão realizados tratamentos estatísticos sobre as variáveis medidas.

3. RISCOS E POTENCIAIS INCONVENIENTES PARA O DOENTE

Todos os testes indicados, que são parte do estudo, são considerados de rotina e são utilizados no trabalho experimental e/ou prática clínica. Todos os equipamentos têm marcação UE e são adequados para o uso em seres humanos. Os testes efetuados não são invasivos, não comportando qualquer perigo para a saúde dos participantes. O investigador responsável pela realização do estudo estará em contacto permanente consigo.

4. POTENCIAIS BENEFÍCIOS

A sua participação permite-nos estudar os circuitos neuronais envolvidos na tarefa de interpretação de um algoritmo, e perceber de que forma estes circuitos se envolvem na tarefa, dependendo se está

perante um algoritmo iterativo ou perante um algoritmo recursivo. A informação recolhida irá contribuir para uma melhor clareza na informação dos investigadores sobre o processamento neuronal durante a interpretação de algoritmos por aqueles que desenvolveram essa capacidade.

5. NOVAS INFORMAÇÕES

Será informado de qualquer informação que possa ser relevante para a sua condição ou que possa influenciar a sua vontade de continuar a participar no estudo.

6. TRATAMENTOS ALTERNATIVOS

Não receberá qualquer tratamento no contexto do estudo em que irá participar.

7. SEGURANÇA

Embora não se espere que venha a sofrer problemas de saúde devido à sua participação, se sofrer alguma lesão física como resultado de quaisquer procedimentos do estudo, realizados de acordo com o protocolo, será reembolsado(a) pelas despesas médicas necessárias para a tratar. Todos os procedimentos são revistos pelo corpo técnico e todos os incidentes reportados às pessoas e/ou entidades competentes.

8. PARTICIPAÇÃO/ABANDONO VOLUNTÁRIO

É importante que saiba que a decisão de participar neste estudo de investigação é inteiramente voluntária. Pode livremente aceitar ou recusar participar, ou ainda retirar o seu consentimento em qualquer altura sem qualquer consequência para si, sem precisar de explicar as razões, sem qualquer penalidade ou perda de benefícios e sem comprometer a sua relação com os investigadores que lhe propõem a participação neste estudo. Ser-lhe-á pedido para informar os investigadores se decidir retirar o seu consentimento.

Os investigadores do estudo podem decidir terminar a sua participação se entenderem que não é do melhor interesse para o seu bem-estar ou para a sua saúde continuar nele. A sua participação pode

ser também terminada se não estiver a seguir o plano do estudo, por decisão administrativa ou decisão da Comissão de Ética. Os investigadores do estudo notificá-lo-ão se surgir uma dessas circunstâncias e falarão consigo a respeito da mesma.

9. CONFIDENCIALIDADE

Sem violar as normas de Confidencialidade, serão atribuídos a Auditores e Autoridades Reguladoras acesso aos registos médicos para verificação dos procedimentos realizados e informação obtida no estudo, de acordo com as leis e regulamentos aplicáveis. Os seus registos manter-se-ão confidenciais e anonimizados de acordo com os regulamentos aplicáveis. Se os resultados deste estudo forem publicados, a sua identidade manter-se-á confidencial.

Ao assinar este consentimento informado autoriza este acesso condicionado e restrito.

Pode ainda, em qualquer altura, exercer o seu direito de acesso à informação. Pode ter também acesso à sua informação médica/biomédica/resultados dos testes através dos investigadores deste estudo. Tem também o direito de se opor à transmissão de dados que sejam cobertos pela confidencialidade profissional.

Os registos médicos que o identificarem e o formulário de Consentimento Informado que assinar serão verificados para fins do estudo pelo Promotor e/ou por representantes do Promotor, e para fins regulamentares pelo Promotor e/ou pelos representantes do Promotor e Agências Reguladoras noutros países. A Comissão de Ética responsável pelo estudo pode solicitar o acesso aos seus registos médicos para assegurar-se que o estudo está a ser realizado de acordo com o Protocolo. Não pode ser garantida confidencialidade absoluta devido à necessidade de passar a informação a essas partes.

Ao assinar este termo de Consentimento Informado, permite que as suas informações médicas neste estudo sejam verificadas, processadas e relatadas conforme for necessário para finalidades científicas legítimas.

Ao assinar este termo de Consentimento Informado, compromete-se a não revelar o conteúdo e a finalidade dos programas de *software* usados no procedimento, pois isso inviabilizaria as experiências para outros voluntários, que não podem conhecer previamente os programas que vão analisar no procedimento.

Confidencialidade e tratamento de dados pessoais

Os dados pessoais dos participantes no estudo, incluindo a informação médica ou de saúde recolhida ou criada como parte do estudo (tais como registos médicos ou resultados de testes), serão utilizados para condução do estudo, ou seja, para fins de investigação científica.

Ao dar o seu consentimento à participação no estudo, a informação a si respeitante, designadamente a informação experimental/clínica, será utilizada da seguinte forma:

1. O Promotor, os investigadores e as outras pessoas envolvidas no estudo recolherão e utilizarão os seus dados pessoais para as finalidades acima descritas;
2. Os dados do estudo, associados a um código que não o identifica diretamente (e não ao seu nome), serão comunicados pelos investigadores e/ou outras pessoas envolvidas no estudo ao Promotor do estudo, que os utilizará para as finalidades acima descritas;
3. Os dados do estudo, associados a um código que não permita identificá-lo diretamente, poderão ser comunicados a Autoridades de Saúde nacionais e internacionais;
4. A sua identidade não será revelada em quaisquer relatórios ou publicações resultantes deste estudo;
5. Todas as pessoas ou entidades com acesso aos seus dados pessoais estão sujeitas a sigilo profissional;
6. Ao dar o seu consentimento para participar no estudo autoriza o Promotor ou empresas de monitorização de estudos, especificamente contratadas para o efeito, e seus colaboradores e/ou Autoridades de Saúde a aceder aos dados constantes do seu processo clínico, para conferir a informação recolhida e registada pelos investigadores, designadamente para assegurar o rigor dos dados que lhe dizem respeito e para garantir que o estudo se encontra a ser desenvolvido corretamente e que os dados obtidos são fiáveis;
7. Nos termos da lei, tem o direito de, através de um dos médicos envolvidos no estudo, solicitar o acesso aos dados que lhe digam respeito, bem como de solicitar a retificação dos seus dados de identificação;
8. Tem ainda o direito de retirar este consentimento em qualquer altura através da notificação aos investigadores, o que implicará que deixe de participar no estudo. No entanto, os dados recolhidos ou criados como parte do estudo até essa altura que não o identifique poderão continuar a ser utilizados para o propósito do estudo, nomeadamente para manter a integridade científica do estudo, e a sua informação médica não será removida do arquivo do estudo;

9. Se não der o seu consentimento, assinando este documento, não poderá participar neste estudo. Se o consentimento agora prestado não for retirado e até que o faça, este será válido e manter-se-á em vigor.

10. COMPENSAÇÃO

Haverá lugar uma compensação financeira sob a forma de um voucher "Cartão Dá", no valor de 40 euros, válido nas lojas do grupo Sonae.

11. CONTACTOS

Se tiver perguntas relativas aos seus direitos como participante deste estudo, deve contactar:
Presidente da Comissão de Ética da FMUC
Azinhaga de Santa Comba, Celas – 3000-548 Coimbra
Telefone: 239 857 707
E-mail: comissaoetica@fmed.uc.pt

Se tiver questões sobre este estudo, deve contactar:
Direção do IBILI - Investigador Coordenador: Professor Doutor Miguel Castelo-Branco
IBILI, FMUC
Azinhaga de Santa Comba, Celas, 3000-548 Coimbra
Contacto telefónico: 239 480 200
E-mail: direccao@ibili.uc.pt

NÃO ASSINE ESTE FORMULÁRIO DE CONSENTIMENTO INFORMADO A MENOS QUE TENHA TIDO A OPORTUNIDADE DE PERGUNTAR E TER RECEBIDO RESPOSTAS SATISFATÓRIAS A TODAS AS SUAS PERGUNTAS.

CONSENTIMENTO INFORMADO

De acordo com a Declaração de Helsínquia da Associação Médica Mundial e suas atualizações:

1. Declaro ter lido este formulário e aceito de forma voluntária participar neste estudo;
2. Fui devidamente informado da natureza, objetivos, riscos e duração provável do estudo, bem como do que é esperado da minha parte;
3. Tive a oportunidade de fazer perguntas sobre o estudo e percebi as respostas e as informações que me foram dadas. A qualquer momento posso fazer mais perguntas ao médico responsável do estudo. Durante o estudo e sempre que quiser, posso receber informação sobre o seu desenvolvimento. O médico responsável dará toda a informação importante que surja durante o estudo que possa alterar a minha vontade de continuar a participar;
4. Aceito que utilizem a informação relativa à minha história clínica e os meus tratamentos no estrito respeito do segredo médico e anonimato. Os meus dados serão mantidos estritamente confidenciais. Autorizo a consulta dos meus dados apenas por pessoas designadas pelo Promotor e por representantes das autoridades reguladoras;
5. Aceito seguir todas as instruções que me forem dadas durante o estudo. Aceito em colaborar com o médico e informá-lo imediatamente das alterações do meu estado de saúde e bem-estar e de todos os sintomas inesperados e não usuais que ocorram;
6. Autorizo o uso dos resultados do estudo para fins exclusivamente científicos e, em particular, aceito que esses resultados sejam divulgados às autoridades sanitárias competentes;
7. Aceito que os dados gerados durante o estudo sejam informatizados pelo Promotor ou outrem por si designado. Eu posso exercer o meu direito de retificação e/ ou oposição;
8. Tenho conhecimento que sou livre de desistir do estudo a qualquer momento, sem ter de justificar a minha decisão e sem comprometer a qualidade dos meus cuidados médicos. Eu tenho conhecimento que o médico tem o direito de decidir sobre a minha saída prematura do estudo e que me informará da causa da mesma;
9. Fui informado que o estudo pode ser interrompido por decisão dos investigadores, do Promotor ou das autoridades reguladoras.

Nome do Participante ou Representante Legal

_____

Assinatura: _____

Data:_____/_____ / ____

Confirmo que expliquei ao participante acima mencionado a natureza, os objetivos e os potenciais riscos do estudo acima mencionado.

Nome do Investigador _____

Assinatura: _____

Data:_____/_____ / ____

B

Voucher declaration

Declaração

_____(nome), declaro que recebi 1 (um) voucher válido para as lojas do grupo Sonae, no valor de 40 euros, como compensação pela participação num estudo experimental do projeto BASE, Projeto de IC&DT – AAC n.º 02/SAICT/2017, projeto n.º 31581 (BASE - Biofeedback Augmented Software Engineering).

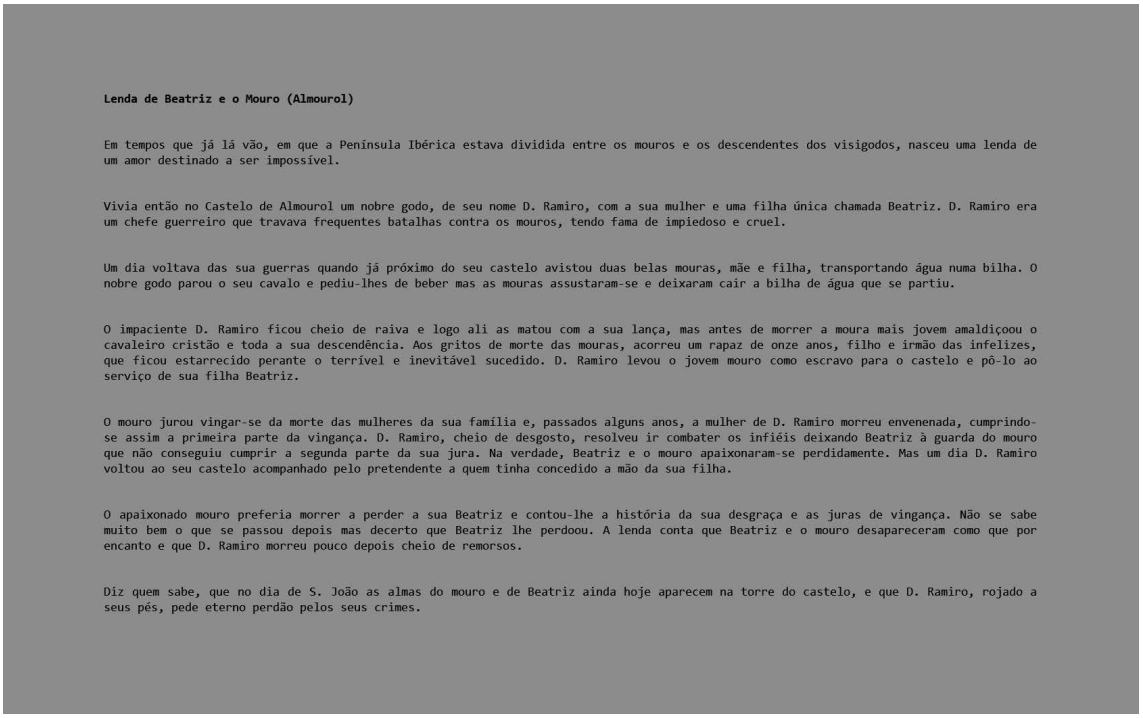
Data:_____

Assinatura: _____

C

Experimental protocol

C.1 Text 1



Lenda de Beatriz e o Mouro (Almourol)

Em tempos que já lá vão, em que a Península Ibérica estava dividida entre os mouros e os descendentes dos visigodos, nasceu uma lenda de um amor destinado a ser impossível.

Vivia então no Castelo de Almourol um nobre godo, de seu nome D. Ramiro, com a sua mulher e uma filha única chamada Beatriz. D. Ramiro era um chefe guerreiro que travava frequentes batalhas contra os mouros, tendo fama de impiedoso e cruel.

Um dia voltava das sua guerras quando já próximo do seu castelo avistou duas belas mouras, mãe e filha, transportando água numa bilha. O nobre godo parou o seu cavalo e pediu-lhes de beber mas as mouras assustaram-se e deixaram cair a bilha de água que se partiu.

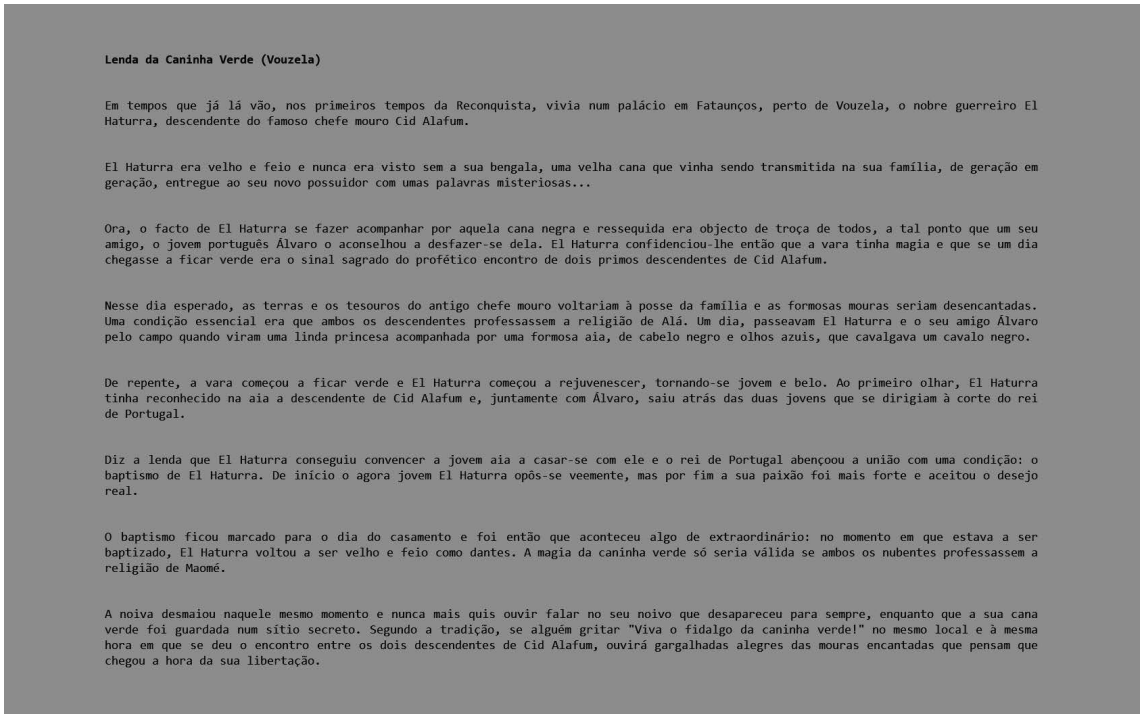
O impaciente D. Ramiro ficou cheio de raiva e logo ali as matou com a sua lança, mas antes de morrer a moura mais jovem amaldiçoou o cavaleiro cristão e toda a sua descendência. Aos gritos de morte das mouras, acorreu um rapaz de onze anos, filho e irmão das infelizes, que ficou estarrecido perante o terrível e inevitável sucedido. D. Ramiro levou o jovem mouro como escravo para o castelo e pô-lo ao serviço de sua filha Beatriz.

O mouro jurou vingar-se da morte das mulheres da sua família e, passados alguns anos, a mulher de D. Ramiro morreu envenenada, cumprindo-se assim a primeira parte da vingança. D. Ramiro, cheio de desgosto, resolveu ir combater os infiéis deixando Beatriz à guarda do mouro que não conseguiu cumprir a segunda parte da sua jura. Na verdade, Beatriz e o mouro apaixonaram-se perdidamente. Mas um dia D. Ramiro voltou ao seu castelo acompanhado pelo pretendente a quem tinha concedido a mão da sua filha.

O apaixonado mouro preferia morrer a perder a sua Beatriz e contou-lhe a história da sua desgraça e as juras de vingança. Não se sabe muito bem o que se passou depois mas decerto que Beatriz lhe perdoou. A lenda conta que Beatriz e o mouro desapareceram como que por encanto e que D. Ramiro morreu pouco depois cheio de remorsos.

Diz quem sabe, que no dia de S. João as almas do mouro e de Beatriz ainda hoje aparecem na torre do castelo, e que D. Ramiro, rojado a seus pés, pede eterno perdão pelos seus crimes.

C.2 Text 2



Lenda da Caninha Verde (Vouzela)

Em tempos que já lá vão, nos primeiros tempos da Reconquista, vivia num palácio em Fataunços, perto de Vouzela, o nobre guerreiro El Haturra, descendente do famoso chefe mouro Cid Alafum.

El Haturra era velho e feio e nunca era visto sem a sua bengala, uma velha cana que vinha sendo transmitida na sua família, de geração em geração, entregue ao seu novo possuidor com umas palavras misteriosas...

Ora, o facto de El Haturra se fazer acompanhar por aquela cana negra e ressequida era objecto de troça de todos, a tal ponto que um seu amigo, o jovem português Álvaro o aconselhou a desfazer-se dela. El Haturra confidenciou-lhe então que a vara tinha magia e que se um dia chegasse a ficar verde era o sinal sagrado do profético encontro de dois primos descendentes de Cid Alafum.

Nesse dia esperado, as terras e os tesouros do antigo chefe mouro voltariam à posse da família e as formosas mouras seriam desencantadas. Uma condição essencial era que ambos os descendentes professassem a religião de Alá. Um dia, passeavam El Haturra e o seu amigo Álvaro pelo campo quando viram uma linda princesa acompanhada por uma formosa aia, de cabelo negro e olhos azuis, que cavalgava um cavalo negro.

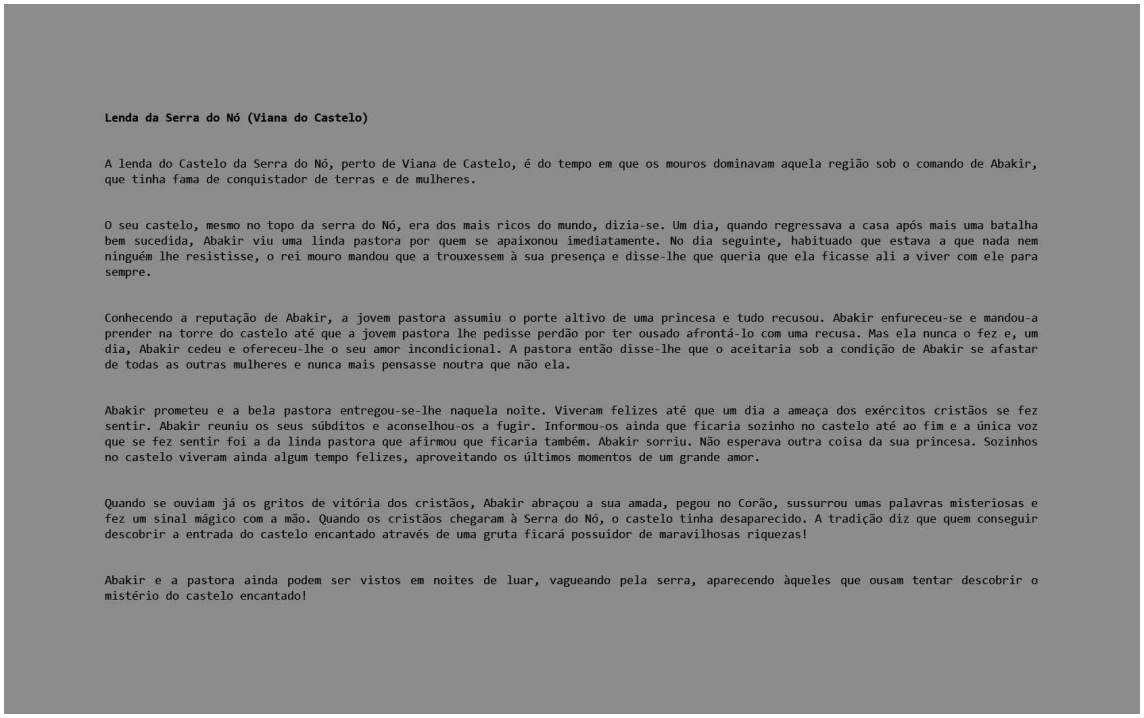
De repente, a vara começou a ficar verde e El Haturra começou a rejuvenescer, tornando-se jovem e belo. Ao primeiro olhar, El Haturra tinha reconhecido na aia a descendente de Cid Alafum e, juntamente com Álvaro, saiu atrás das duas jovens que se dirigiam à corte do rei de Portugal.

Diz a lenda que El Haturra conseguiu convencer a jovem aia a casar-se com ele e o rei de Portugal abençoou a união com uma condição: o baptismo de El Haturra. De início o agora jovem El Haturra opôs-se veemente, mas por fim a sua paixão foi mais forte e aceitou o desejo real.

O baptismo ficou marcado para o dia do casamento e foi então que aconteceu algo de extraordinário: no momento em que estava a ser baptizado, El Haturra voltou a ser velho e feio como dantes. A magia da caninha verde só seria válida se ambos os nubentes professassem a religião de Maomé.

A noiva desmaiou naquele mesmo momento e nunca mais quis ouvir falar no seu noivo que desapareceu para sempre, enquanto que a sua cana verde foi guardada num sítio secreto. Segundo a tradição, se alguém gritar "Viva o fidalgo da caninha verde!" no mesmo local e à mesma hora em que se deu o encontro entre os dois descendentes de Cid Alafum, ouvirá gargalhadas alegres das mouras encantadas que pensam que chegou a hora da sua libertação.

C.3 Text 3



Lenda da Serra do Nó (Viana do Castelo)

A lenda do Castelo da Serra do Nó, perto de Viana de Castelo, é do tempo em que os mouros dominavam aquela região sob o comando de Abakir, que tinha fama de conquistador de terras e de mulheres.

O seu castelo, mesmo no topo da serra do Nó, era dos mais ricos do mundo, dizia-se. Um dia, quando regressava a casa após mais uma batalha bem sucedida, Abakir viu uma linda pastora por quem se apaixonou imediatamente. No dia seguinte, habituado que estava a que nada nem ninguém lhe resistisse, o rei mouro mandou que a trouxessem à sua presença e disse-lhe que queria que ela ficasse ali a viver com ele para sempre.

Conhecendo a reputação de Abakir, a jovem pastora assumiu o porte altivo de uma princesa e tudo recusou. Abakir enfureceu-se e mandou-a prender na torre do castelo até que a jovem pastora lhe pedisse perdão por ter ousado afrontá-lo com uma recusa. Mas ela nunca o fez e, um dia, Abakir cedeu e ofereceu-lhe o seu amor incondicional. A pastora então disse-lhe que o aceitaria sob a condição de Abakir se afastar de todas as outras mulheres e nunca mais pensasse noutra que não ela.

Abakir prometeu e a bela pastora entregou-se-lhe naquela noite. Viveram felizes até que um dia a ameaça dos exércitos cristãos se fez sentir. Abakir reuniu os seus súbditos e aconselhou-os a fugir. Informou-os ainda que ficaria sozinho no castelo até ao fim e a única voz que se fez sentir foi a da linda pastora que afirmou que ficaria também. Abakir sorriu. Não esperava outra coisa da sua princesa. Sozinhos no castelo viveram ainda algum tempo felizes, aproveitando os últimos momentos de um grande amor.

Quando se ouviam já os gritos de vitória dos cristãos, Abakir abraçou a sua amada, pegou no Corão, sussurrou umas palavras misteriosas e fez um sinal mágico com a mão. Quando os cristãos chegaram à Serra do Nó, o castelo tinha desaparecido. A tradição diz que quem conseguir descobrir a entrada do castelo encantado através de uma gruta ficará possuidor de maravilhosas riquezas!

Abakir e a pastora ainda podem ser vistos em noites de luar, vagueando pela serra, aparecendo àqueles que ousam tentar descobrir o mistério do castelo encantado!

C.4 Code 1

```
public class ProcessNums {
     public static int getResult(int[] sequence, int lower, int upper) {
          int result = 0;
          if (sequence == null)
               return result;
          for (int n : sequence) {
               if (n>=lower && n<=upper)
                    result++;
          }
          return result;
     }
     public static void main(String[] args) {
          int[] sequence = {-7, 1, 5, 2, -4, 3, 0};
          int result = getResult(sequence,2,4);
          System.out.println("Result = " + result);
     }
}
```

C.5 Code 2

```
public class MA {
     private static byte[] getInts(String digs) {
          byte[] result = new byte[digs.length()];
          for (int i = 0; i < digs.length(); i++) {
               char c = digs.charAt(i);
               if (c < '0' || c > '9') {
                    throw new IllegalArgumentException("Invalid string " + c + " at position " + i);
               }
               result[digs.length() - 1 - i] = (byte) (c - '0');
          }
          return result;
     }
     public static String getResult(String num1, String num2) {
          byte[] left = getInts(num1);
          byte[] right = getInts(num2);
          byte[] result = new byte[left.length + right.length];
          for (int rightPos = 0; rightPos < right.length; rightPos++) {
               byte rightDigit = right[rightPos];
               byte temp = 0;
               for (int leftPos = 0; leftPos < left.length; leftPos++) {
                    temp += result[leftPos + rightPos];
                    temp += rightDigit * left[leftPos];
                    result[leftPos + rightPos] = (byte) (temp % 10);
                    temp /= 10;
               }
               int destPos = rightPos + left.length;
               while (temp != 0) {
                    temp += result[destPos] & 0xFFFFFFFFL;
                    result[destPos] = (byte) (temp % 10);
                    temp /= 10;
                    destPos++;
               }
          }
          StringBuilder stringResultBuilder = new StringBuilder(result.length);
          for (int i = result.length - 1; i >= 0; i--) {
               byte digit = result[i];
               if (digit != 0 || stringResultBuilder.length() > 0) {
                    stringResultBuilder.append((char) (digit + '0'));
               }
          }
          return stringResultBuilder.toString();
     }
     public static void main(String[] args) {
          System.out.println(getResult("1234","56789"));

     } }
```

C.6 Code 3

```
static boolean verif(int co[][][], int ci[][][], int p[], int d[]) {
     if (co == null || ci == null || p == null || d == null)
          return false;
     if (p.length <3 || d.length <3)
          return false;
     if (co.length < ci.length || co[0].length < ci[0].length || co[0][0].length < ci[0][0].length)
          return false;
     int x,y,z;
     int a,b,c;
     int ma,mb,mc;
     p[0] = p[1] = p[2] = d[0] = d[1] = d[2] = 0;
     for (x=0; x<co.length - ci.length; x++)
          for (y=0; y<co[0].length - ci[0].length;y++)
               for (z=0; z<co[0][0].length - ci[0][0].length; z++) {
                    ma = ci.length;
                    mb = ci[0].length;
                    mc = ci[0][0].length;
                    for (a = 0; a<ma; a++) {
                         b = 0;
                         for (; b<mb; b++) {
                              c = 0;
                              for (; c<mc && co[x+a][y+b][z+c] == ci[a][b][c]; c++);
                              if (c == 0) {
                                   mb = b;
                                   break;
                              }
                              if (c < mc)
                                   mc = c;
                         }
                         if (b==0) {
                              ma = a;
                              break;
                         }
                         if (b<mb)
                              mb = b;
                    }
                    if (ma*mb*mc > d[0]*d[1]*d[2]) {
                         p[0] = x; p[1] = y; p[2] = z;
                         d[0] = ma; d[1] = mb; d[2] = mc;
                    }
               }
          return true;
     }
```

D

Performance evaluation

D.1 Questionnaire 1

Nome:

Atividade nº:

Data:

1. De uma forma clara e muito directa, descreva o que faz o algoritmo – qual a tarefa que desempenha e o que representa o seu resultado.

2. Descreva de forma tão clara quanto possível como funciona o algoritmo – qual os seus passos e o significado deles.

D.2 Response model

Algoritmo 1

1) Obtém a quantidade de números numa matriz/sequência que estão entre dois limites a e b (inclusive)
2) A função percorre todos os números na sequência. Para cada um compara com os limites a e b. Se o número estiver entre a e b, incrementa um contador (inicializado a zero). No final retorna o valor do contador

Algoritmo 2

1) Obtém o produto de dois valores inteiros. Todos os valores indicados são expressos como strings.
2) (Nota: Algoritmo/ideia simples mas difícil de explicar)
   É usado o algoritmo "manual" que se aprende na escola. Os números e o resultado são expressos como strings. Inicialmente são obtidas as suas matrizes de bytes correspondentes em que cada byte corresponde ao valor de cada dígito da string. As matrizes de bytes têm os seus valores pela ordem inversa às das strings. Cada dígito (byte) do segundo número é multiplicado por cada dígito do primeiro número, do menos significativo para o mais significativo. O resultado da multiplicação de cada combinação dig/ num1 x dig/ num2 é colocado na posição que lhe corresponde na matriz de bytes para o resultado. Se o resultado da multiplicação de dois dígitos for superior a 10, na posição correspondente ao digito do resultado fica apenas o valor das unidades (módulo 10) e o dígitos restantes são somados ao dígitos sucessivamente mais significativos no resultado.
   No final as matrizes de bytes são convertidas para strings, colocando-se os dígitos pela ordem habitual.

Algoritmo 3

1) Verifica se uma matriz co contém outra ci, sendo ambas de 3 dimensões, obtendo a primeira ocorrência e indicando a posição dentro de co onde encontrou ci, e quanto de ci é que se encontra em co
2) (Nota: o algoritmo não é difícil de explicar, pode é ser difícil perceber a intenção dele pelo código)
   O algoritmo percorre através de um ciclo dentro doutro dentro de outro todas as posições da matriz co (a que pode conter). Cada elemento assim percorrido pode ser potencialmente o canto inicial da matriz ci (a que é contida). Para verificar se a matriz co realmente está nessa posição e quanto dela é que lá está, percorre-se através de um ciclo dentro de outro dentro de outro todas as posições de ambas as matrizes, a ci a partir do ponto actual dela, e co a partir do seu canto inicial. À medida que vai percorrendo co, o algoritmo vai mantendo as dimensões de co em que os elementos de co coincidem com ci – não é necessário que toda a matriz co esteja dentro de ci.

D.3 Questionnaire 2

Nome:

Atividade nº:

Data:

Em cada um dos indicadores seguintes, assinale com uma cruz a posição que melhor reflecte a sua percepção acerca do esforço sentido.

1. ***Esforço mental***. A tarefa foi mentalmente esgotante?

← Nada Muito →

2. ***Pressão com o tempo***. Durante a tarefa sentiu-se pressionado com o tempo?

← Nada Muito →

3. ***Cumprimento da tarefa***. Qual a sua percepção acerca de ter conseguido cumprir a tarefa?

← Nada cumprida Totalmente cumprida→

4. ***Incómodo***. Qual o seu grau de incómodo (frustração, irritação, stress) durante a tarefa?

← Nada Muito →

Na sua opinião, como indicador de esforço, e quando comparados dois a dois, quais dos indicadores acima é o mais relevante? Assinale com uma cruz <u>o mais importante</u> em cada uma das linhas abaixo.

Esforço mental	☐	*vs.*	Pressão com o tempo	☐
Esforço mental	☐	*vs.*	Cumprimento da tarefa	☐
Esforço mental	☐	*vs.*	Incómodo	☐
Pressão com o tempo	☐	*vs.*	Cumprimento da tarefa	☐
Pressão com o tempo	☐	*vs.*	Incómodo	☐
Cumprimento da tarefa	☐	*vs.*	Incómodo	☐