1 2 9 0

UNIVERSIDADE Ð COIMBRA

André Matias Bernardes

# Quality Assessment of Inspection and Code Development Using Non-Intrusive Physiological Indicators

## HRV and Pupillography Study

Dissertação no âmbito do Mestrado em Engenharia Biomédica, orientada pelo Professor Doutor Paulo Fernando Pereira de Carvalho, pelo Doutor Ricardo Jorge dos Santos Couceiro e pelo Engenheiro Júlio Cordeiro Medeiros e apresentada ao Departamento de Física da Universidade de Coimbra.

Setembro de 2022

This work was developped under project BASE, Biofeedback Augmented Software Engineering (POCI - 01-0145 - FEDER- 031581), in collaboration with:

**CISUC - Centre for Informatics and Systems of University of Coimbra**

# Agradecimentos

Gostaria de começar por agradecer aos meus orientadores Professor Doutor Paulo de Carvalho, Doutor Ricardo Couceiro e Engenheiro Júlio Medeiros. Muito obrigado por toda a disponibilidade, ajuda, espírito crítico e ensinamentos que proporcionaram ao longo deste último ano que permitiram não só a produção de um trabalho de maior qualidade, mas também contribuíram bastante para a minha formação enquanto engenheiro. Gostaria também de expressar a minha gratidão aos restantes Professores que foram coautores dos artigos desenvolvidos durante a realização do presente trabalho, que foram presença assídua nas reuniões do projeto e contribuíram também para a minha aprendizagem.

A presente tese foi financiada parcialmente pelo projeto BASE, Biofeedback Augmented Software Engineering (Grant POCI-01-0145-FEDER-031581) e pelo Centro de Informática e Sistemas da Universidade de Coimbra (CISUC – financiamento obtido pela FCT - Foundation for Science and Technology, I.P./MCTES através de fundo nacionais (PIDDAC), no âmbito do projeto CISUC-AC-Adaptive Computation (Referência do Projeto: UIDB/00326/2020). A estas entidades gostaria também de expressar a minha gratidão e agradecimento.

Gostaria ainda de agradecer ao Núcleo de Estudantes do Departamento de Física da Associação Académica de Coimbra (NEDF/AAC) e a todos os seus elementos. Obrigado por toda a vossa ajuda e disponibilidade durante este ano que permitiram conciliar o presente trabalho com a presidência do NEDF/AAC, e que fomentaram o crescimento do meu espírito crítico, trabalho de grupo e liderança.

Aos meus amigos, os que fiz durante os últimos 5 anos e aos que já trazia, um

muito obrigado por todas as experiências partilhadas, que tornaram a minha jornada enquanto estudante de Coimbra numa das minhas melhores memórias. Um agradecimento especial à Botinas e à Jacinta que foram duas "póneis" que encontrei no Departamento de Física e que me acompanharam em todos os momentos. Ao Fernando e ao Roger, um obrigado por todo o companheirismo, por todas as conversas e desabafos e por estarem sempre presentes. Aos "BCL" ficam as memórias de todos os jantares e momentos que partilhámos, foi mesmo BCL durante 5 anos! Um obrigado a todos. Aos "Kebabs", que surgiram no meu último ano de estudante de Coimbra, quando achava que mais nada me ia surpreender, um obrigado por mostrarem que estava errado, e por todos os "tachos" que partilhámos. Finalmente, e como não poderia deixar de ser, um enorme agradecimento à minha família de praxe "Os Borgas". Aos mais velhos, Dias, Artur e Jojo, os fundadores da família, foi com vocês que que aprendi o verdadeiro significado de ser estudante de Coimbra e que nunca vou esquecer todos os valores que me ensinaram. Ao Costini e ao Mendonça, os melhores caloiros que alguém pode ter, um obrigado pela vossa constante boa disposição que alegra toda a família. Aos mais novos Afonso, Edu, Herbert, Preces, Ivo, Morais, Arnaut e Jaime, obrigado por obrigarem a família a reinventar-se constantemente, são vocês que vão continuar o nosso legado. Costini e Afonso, muito obrigado por todo o companheirismo, que continuem as reuniões para comer marisco e passar fins-de-semana na "Republica of West Virginia".

Por último, gostaria de deixar o meu enorme agradecimento à minha família, em especial aos meus pais, Delfim e Filomena, e à minha irmã Carolina. Muito obrigado por me passarem todos os valores e princípios que uma pessoa deve ter e por sempre me incentivarem a querer aprender mais e a nunca baixar os braços, mesmo nas situações mais adversas. Sem vocês não seria quem sou hoje, e é sem dúvida a vocês que tenho mais a agradecer ter chegado aqui. Um enorme obrigado por tudo!

# Resumo

Esta tese foi desenvolvida no âmbito do projeto “Biofeedback Augmented Software Engineering” (BASE; Grant POCI - 01-0145 - FEDER- 031581), o qual tem o objetivo de desenvolver uma solução capaz de detetar as zonas de código com maior probabilidade de ocorrência de erros, baseado nos sinais vitais do programador. O objetivo desta tese é avaliar a qualidade e fiabilidade das características dos sinais temporais da variabilidade cardíaca (HRV) e da variação da dilatação da pupila (pupilografia) para a discriminação de diferentes níveis de stress cognitivo em ambientes de inspeção de código, que podem ser adquiridos usando métodos não invasivos.

De modo alcançar a solução proposta pelo projeto BASE, é necessário começar por descobrir a resolução temporal ideal que otimiza a deteção de variações no stress cognitivo para cada característica do sinal HRV, sem comprometer a sua fiabilidade no contexto de inspeção de código. No entanto, os estudos existentes relacionados com este tópico foram desenvolvidos com os sujeitos em repouso ou realizando tarefas básicas em ambientes muito controlados. De modo a descobrir quais as características do HRV mais adequadas para serem utilizadas em aplicações reais, como o contexto de inspeção de código mencionado, e para perceber as suas limitações temporais, foram realizadas abordagens de estudo de análise estatística e de classificação. Um total de 31 características do sinal HRV extraídas utilizando janelas temporais de diferentes tamanhos (entre 3 minutos e 10 segundos) foram analisadas em contexto de inspeção de código.

Seguindo a abordagem da análise estatística, foi possível identificar um conjunto de cinco características consideradas as mais fiáveis em janelas temporais curtas no presente contexto: mNN, HF, LF, LFpeak e totPow. Desta abordagem, determinou-se ainda que 30 segundos foi a duração mais curta contendo características consideradas fiáveis. A abordagem da classificação utilizou classificadores SVM (Support Vector Machine) para analisar o impacto da janela de extração nos resultados da classificação da complexidade de secções de código de software. As características do sinal HRV foram associadas às secções de código observadas pelo programador e transformadas estatísticas das mesmas foram calculadas. Os F1-Scores obtidos para os diferentes classificadores variaram entre 0.62 e 0.75, sendo que se desconsiderarmos os resultados da janela temporal de 10 segundos, que mostrou ser demasiado curta para o contexto atual, os F1-scores variaram entre 0.66 e 0.75. Estes resultados indicam que é possível obter performances de classificação semelhantes utilizando janelas mais curtas comparativamente com as mais longas.

Relativamente às características do sinal da pupilografia, verifica-se a falta de consenso nas linhas de orientação relativas às bandas de frequência deste sinal, com diversos autores a utilizarem diferentes bandas de frequência na sua análise. Com isto em mente, procurámos de entre várias hipóteses a combinação de limites de bandas que maximiza a correlação entre a banda das baixas frequências (LF) e a das altas frequências (HF) da pupilografia com estas mesmas bandas do sinal HRV. Seguindo este procedimento fomos capazes de selecionar os limites de banda adequados para as bandas LF e HF para a extração de características. Os nossos resultados indicam que a banda mais adequada para as LF vai desde 0.13Hz a 0.28Hz e para as HF desde 0.28Hz a 0.35Hz. Destas bandas foram extraídas características que foram associadas à respetiva secção observada pelo participante no respetivo momento de extração e foram calculadas transformadas estatísticas destas características. Recorrendo a classificadores SVM, treinados utilizando estas transformadas, alcançou-se um F1-Score médio de 0.76 com um desvio padrão de 0.07, o melhor resultado em todo o estudo, atingindo o maior F1-score médio com a menor variabilidade. Estes resultados indicam que poderá ser possível alcançar um método totalmente não invasivo baseado em características da pupilografia para classificação de complexidade de secções de código.

**Palavras-Chave**: engenharia de software; erro humano; biofeedback; tarefas cognitivamente exigentes; compreensão de código; processamento de biosinais; variabilidade cardíaca (HRV); características ultra-curtas da HRV; pupilografia; limite das bandas da frequência

# Abstract

This thesis was developed under the Biofeedback Augmented Software Engineering (BASE) project (Grant POCI - 01-0145 - FEDER- 031581), which aims to develop a solution capable of using biofeedback from the programmer to detect software code areas more prone to error. This thesis aims to assess the quality and reliability of Heart Rate Variability (HRV) and Pupillography (Pupil Diameter time series) measurements for cognitive stress discrimination in a code inspection context, which can be acquired using non-intrusive methods.

In order to accomplish the solution described, we need to find the ideal time resolution for each HRV feature which optimizes the detection of cognitive stress variations without compromising its reliability in a code inspection context. However, the studies found in the literature related to this topic were developed with the subjects at rest or performing elementary tasks in controlled environments. In order to find out which HRV features are adequate to be used in real-life applications, such as the mentioned high cognitive dynamic code inspection context, and to understand their time frame limitations, statistical and classification analysis approaches were followed. A total of 31 HRV features, extracted using time frames of variable sizes (ranging from 3 minutes to 10 seconds) in a code inspection context, were analyzed through these two approaches.

From the statistical approach, we could identify five features as the most reliable for the smallest time frames considering the present context: the mean NN, the HF, the LF, the LFpeak and the totPow features. Furthermore, we also determined that the 30-second window was the smallest time frame considered to have reliable measurements. The classification approach used Support Vector Machine (SVM) classifiers to analyze the impact of the extracting window in the complexity classification of software code sections. The HRV features were associated with the corresponding code section gazed at the extraction time, and statistical transformations of these features were computed. The F1-Scores obtained for the different classifications ranged from 0.62 to 0.75 across all windows. Furthermore, excluding the 10-second corresponding results, a window that proved to be too short of a time frame in the current context, the mean F1 scores obtained ranged between 0.66 and 0.75, indicating that it is possible to achieve similar classification performances using smaller time frames.

Regarding the pupillography measurements, in the literature, there is a lack of consensus in the guidelines about the pupillography frequency bands, with several authors using and reporting different bands for this signal analysis. With this in mind, we searched through several pupillography frequency band combinations to find the low-frequency (LF) and high-frequency (HF) bands that maximized the correlation with the HRV LF and HF bands. Following this procedure, we were capable of selecting adequate LF and HF band limits for the feature extraction in the present code inspection context: the LF band from 0.13Hz to 0.28Hz and the HF band from 0.28Hz to 0.35Hz. The features extracted from these bands were associated with the corresponding code section, and statistical transformations of these features were computed. An SVM classifier was trained using these transformed features, achieving a 0.76 mean F1-Score with a standard deviation of 0.07 which was the best performance in the overall study, having the highest mean F1-Score with the lowest variability. These results indicate that it could be possible to achieve an entire non-intrusive method using pupillography features for code complexity classification.

**Keywords**: software engineering; human error; biofeedback; cognitive demanding tasks; code comprehension; bio-signal processing; Heart Rate Variability (HRV); ultra-short-term HRV features; pupillography; frequency-bands limits

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**MRI** Magnetic Resonance Imaging. 32, 101

**PCA** Principal Component Analysis. 77

**PD** Pupil Diameter. 18, 80, 81

**PLR** Pupillary Light Response. 23

**PNS** Peripheral Nervous System. 5, 6

**PPG** Photoplethysmogram. 18, 27

**PSD** Power Spectrum Density. xvi, 24, 41, 82, 85

**SDLC** Software Development Life Cycle. 17

**SVM** Support Vector Machine. 31, 52, 54, 89, 90, 91, 104

**VLF** very-low-frequency. 19, 21, 41, 42, 67, 71

# 1

# Introduction

## 1.1 Context and Motivation

Currently, the world is facing a software boom driven by the current continuous developments in the different technology fields. The internet is now globally available, making everything interconnected, which leads to a vast market opportunity in e-commerce, advertisement, and telehealth, among other sectors. Furthermore, recent advances in automation, machine learning and artificial intelligence (AI) technologies allowed the emergence of new technology solutions such as speech recognition, autonomous driving and even clinical decision support systems.

In order to fulfil the demands and requirements of the different specific software development fields, and due to the programming inner cognitively demanding characteristics, software engineers and developers are constantly under tremendous amounts of pressure and stress. This intense environment is prone to human error in the form of residual software faults (bugs), which are currently one of the software industry's most significant problems. A study conducted in 2018 estimated the existence of an average of about 15 to 50 errors per 1000 lines in completed code [1]. This fact translates into effective costs, both in human time and effort as well as in financial costs. Code comprehension and bug detection tasks consume up to 70% of the programmers working time [2] and according to a Tricentis' research [3], in 2017 was estimated that the total net losses due to software bugs ascended to about 1.7 trillion dollars across the globe.

Software bugs commonly occur in the most complex code sections, so code complexity measures can be included in the fault avoidance techniques [4]. However, assessing the code complexity level is not a straightforward process. Two of the most commonly employed metrics are Halstead's effort metric, and McCabe's cyclomatic metric [5]. Unfortunately, these measures are not sensitive enough to specific contexts and fail to consider interactions between different program components [5]. The Biofeedback Augmented Software Engineering (BASE) (POCI - 01-0145 - FEDER- 031581) project attempts a different approach using the programmers' biofeedback to identify the code sections' complexity according to the subject's cognitive stress.

The most complex code sections are also the code areas which require a higher mental effort and cognitive load from the programmers. This cognitive stress can be measured through the Autonomous Nervous System (ANS) physical responses, which can include subtle changes in the heart rate [6], or pupil size variations [7]. In this way, a biofeedback approach using these biological measurements could be used to achieve software capable of identifying the code sections with higher bug probability and consequently, that should be reviewed. Finding the link between the code complexity and the physiological signals controlled by the ANS that can be extracted using non-invasive devices is precisely one of the goals of the BASE project.

## 1.2 Objectives

In order to achieve the software goal with the described traits, one of the first dilemmas faced is which time resolution should be used in the analysis of the measurements. Smaller time windows may provide greater time resolution in the analysis, making it possible to capture ANS dynamics during smaller gazing periods at a specific code area; information that could not be accessed using larger time windows. However, if the time frame is too small, the measurements' reliability may be compromised. In this way, one of the objectives of this thesis is to **comprehend how the time window size chosen affects the reliability of different Heart Rate Variability (HRV) features and how it influences the complexity classification of software code.**

During the research phase of the project it was also found a lack of consensus regarding the frequency bands of the pupil size variation signal (pupillography). With this in mind, another goal of this work is to **define the optimal frequency bands of the Pupillography that should be applied to study cognitive stress during high cognitively demanding contexts.**

## 1.3 General outline of the thesis

This document is structured as follow: **Chapter 2** briefly describes the physiological background and concepts related to the Autonomous Nervous System, the Heart Rate Variability and the pupil size variation; **Chapter 3** dissects the state of the art of cognitive stress detection through Ultra-Short-Term HRV features and pupil size variation measurements; **Chapter 4** describes the experiment design and data collected; **Chapter 5** is divided into three main sections regarding the methods, results and discussion related to the HRV measurements different approaches followed to assess the quality and reliability of these measurements; **Chapter 6** is also divided into three main sections regarding the methods, results and discussion related to the pupillography measurements bands' limits definition and assessment of these measurements discriminative code complexity ability; **Chapter 7** presents the study limitations and suggests the future work that should be conducted; finally, **Chapter 8** delineates the main conclusions of the present thesis.

## 1.4 Main Contributions

This thesis includes several original contributions. Two different papers were published in the scope of the present work. Furthermore, we are currently preparing a third paper that will be submitted to Scientific Reports journal.

The six pages two columns article *"Impact of Ultra-Short-Term HRV Features in Software Code Sections Complexity Classification"*, depicting the window HRV analysis reduction effects in code sections complexity classification, was submitted, accepted and orally presented at the 21st IEEE Mediterranean Electrotechnical Con-

ference (IEEE MELECON 2022), held in Palermo, Italy, on June 14-16, 2022 [8].
The study *"How Reliable Are Ultra-Short-Term HRV Measurements during Cognitively Demanding Tasks?"*, which uses statistical and correlation analysis to evaluate the reliability of ultra-short-term HRV measurements under cognitively demanding tasks context, was submitted to **Sensors** on June 22, and was published on August 30, 2022 [9].
Besides the highlighted contributions regarding the reliability and classification impact of ultra-short-term HRV features, the present thesis also provides insight into the pupillography frequency bands that should be used for the pupillography signal analysis under highly cognitively demanding contexts. Furthermore, it includes the analysis of the pupillography features discriminative power for code complexity classification. This work is currently being prepared to be submitted to the Scientific Reports journal.

# 2

# Physiology Background Concepts

## 2.1 Human Nervous System

The Nervous System is the structure in charge of everything that happens in the Human Body. This system is highly complex and organized, compiling several functions like reasoning, movement control, sensory responses and performing an integrative role across the different physiologic systems. Regarding its configuration, two major subdivisions compose the Human Nervous System: the Central Nervous System (CNS) and the Peripheral Nervous System (PNS) [10].

The PNS is responsible for carrying the signals exchanged between the different body components, whereas the CNS is accountable for receiving these messages, processing the contained information, and sending back signals answering the received stimuli [11].

Anatomically speaking, the CNS is formed by both the brain and the spinal cord. The brain is essentially composed of a complex network of wired neurons supported by glia cells and is encapsulated by the skull. This organ has four main constituents [10]:

- **The cerebrum** - is a significant part of the brain. The cerebrum is composed of the right and the left hemispheres, each in control of the opposing side of the body. The hemispheres are arranged in four lobes: Frontal, Temporal, Parietal and Occipital.
- **The brain stem** - the cerebrum is connected to the spinal cord by this structure. “Medulla oblongata” is another name for it.
- **The cerebellum** - positioned beneath and behind the cerebrum.

- **The diencephalon** - the thalamus and hypothalamus are part of this structure.

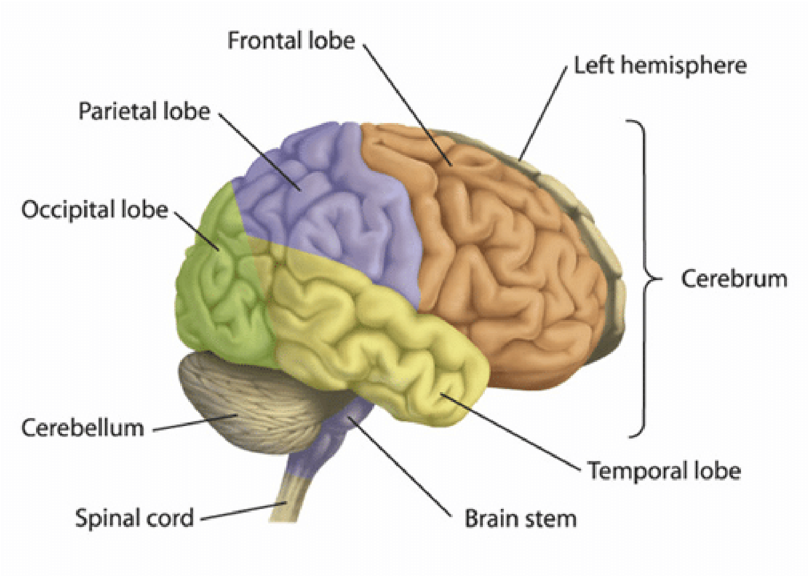


**Figure 2.1:** Representation of the cerebrum, divided into four lobes (Frontal, Temporal, Parietal and Occipital), the Cerebellum, the Brain stem and the Spinal cord. Adapted from [12]

The spinal cord is integrated into the vertebral column and connects the brain to the rest of the body, having in its composition 31 segments, and a pair of spinal nerves surges from each segment. The sensory and motor nerves are housed in the spinal cord [13].

Concerning the PNS, essentially every nervous tissue (sensory neurons, nerves, autonomic ganglia, enteric plexuses and others), excluding the brain and the spinal cord, are part of this system. The PNS has two distinct constituents, the Somatic Nervous System (SNS), which manages the voluntary and conscious activities (including sensory neurons and motor nerves), and the Autonomous Nervous System (ANS). This project thesis develops itself around the ANS dynamics.

### 2.1.1 Autonomic nervous system

As its name suggests, the ANS is behind the regulation of the body's involuntary visceral functions, having a central role in the homeostasis maintenance [14]. The cardiac pulsation, blood pressure, breathing mechanisms, digestion, or pupillary

response, are examples of these functions [11]. The ANS structuration has three divisions: the Enteric, the Sympathetic and the Parasympathetic Nervous Systems. The Enteric Nervous System is responsible for the gastrointestinal system's involuntary operation, and functions independently from the CNS [10]. The Sympathetic and the Parasympathetic Nervous Systems are the two major ANS subdivisions.
The Sympathetic Nervous System takes over during stressful or emotional situations, activating the so-called fight-or-flight mode, which prepares the body to face dangerous events. The responses of the sympathetic system can include: increasing blood pressure and heart rate, where in some circumstances, the body is capable of doubling its heart rate in 3 to 5 seconds [11]; sweating; pupil dilation, which has the effect of improving human long-range vision [10]; activate the adrenal medulla and redirect the blood flow to increase the available energy in the muscles and brain.
On the other hand, the Parasympathetic Nervous System has an antagonistic function and constantly works to assist the body in maintaining homeostasis. It is responsible for energy conservation and restoring body stability after the sympathetic system activation caused by stressful situations [10]. Some processes regulated by the parasympathetic system are the redirection of more blood to the intestines to promote the digestion processes; decreasing the heart rate and blood pressure (in extreme circumstances the arterial pressure can be drastically reduced during 10 to 15 seconds causing fainting [11]) and pupils' constriction, which enables closer vision improvement [10].

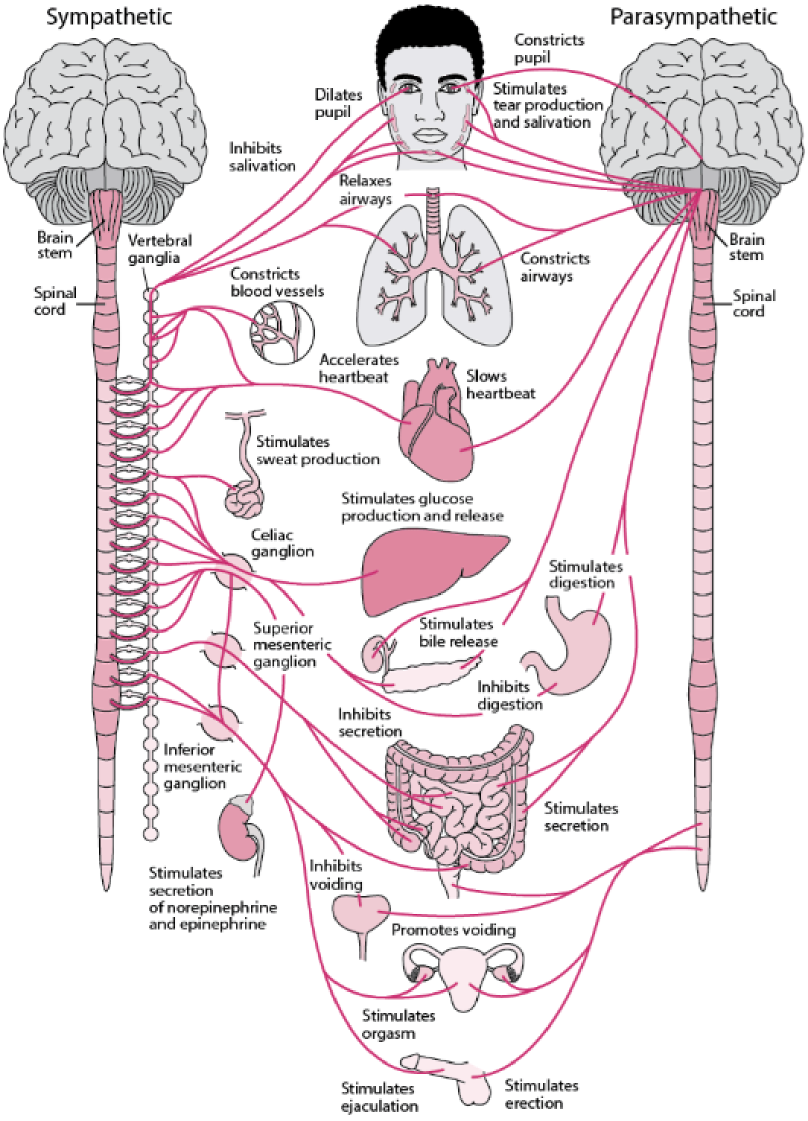


**Figure 2.2:** Schematic representation of the ANS and the several functions it controls. Adapted from [15]

## 2.2 Cognitive Stress

In a normal and calm environment, the human body establishes a dynamic equilibrium called homeostasis [16]. As seen in section 2.1.1, the parasympathetic nervous system takes control during this equilibrium [10]. Stress is an actual or anticipated disruption of this equilibrium [16]. This disruption occurs when the existing mental and physiological resources are insufficient to meet current demands, such as a response to an actual physical dangerous situation or when a person is a few minutes away from the deadline of important work submission [17]. The stress reaction allows the body to adjust to the stimuli that provoked the homeostasis disruption by supplying the body with additional suppressing the immune system [18]. During these situations of stress, the sympathetic nervous system takes over while the parasympathetic is suppressed [17].

Cognitive stress is the stress provoked by increased cognitive load demands, such as the cognitive stress induced when solving arithmetic equations [18, 19]. According to the Cognitive Load Theory [20], cognitive architecture is divided into working memory and long-term memory. Working memory is the limited memory employed for all conscious activities, such as reading. On the other hand, long-term memory corresponds to the memories and knowledge stored that can be accessed by the working memory. The cognitive load is the quantity of information being processed by the subject and can not exceed the working memory limit. Different factors can influence the cognitive load, such as the subject's characteristics, the environment or the complexity of the task being performed. A task with higher complexity will require an increased mental effort (i.e. cognitive capacity allocated to the task) than a lower complexity task, producing a higher cognitive load which induces more cognitive stress [20].These changes in cognitive stress levels manifest themselves through Autonomous Nervous System (ANS) variations and are reflected in biosignals controlled by the ANS, such as the Heart Rate Variability (HRV) or the pupil diameter variation [10, 19].

## 2.3 The Heart

The Human Heart can be found in the thoracic cavity, slightly placed left to the body's centre midline and is the central organ of the circulatory system. This system is composed essentially of the heart and a system of arteries, veins and capillaries, and is responsible for supplying the different body parts with nutrients and oxygen transported in the bloodstream. The heart is accountable for pumping the blood through the body's blood vessels so that it reaches all the body components, acting as a "muscular pumping mechanism" [21].

In terms of anatomy, the heart has four cavities, two ventricle chambers (left and right), and two atrium chambers (left and right). During the circulation process, the blood arrives from the body in the heart's right atrium, then to the right ventricle, where it is pumped in the direction of the lungs for gas exchanges (pulmonary circulation). The oxygen-enriched blood returns to the heart through the left atrium, from where it goes to the left ventricle so that it can be pumped to the rest of the human body (systemic circulation). In order to prevent blood reflux, the heart also has a set of valves which force the blood to flow in the correct direction. Two atrioventricular valves connect the atriums and ventricles (tricuspid valve on the right side and bicuspid valve on the left), and semilunar valves that manage the blood flow between the heart and other body parts (pulmonary and aortic valves) [21].

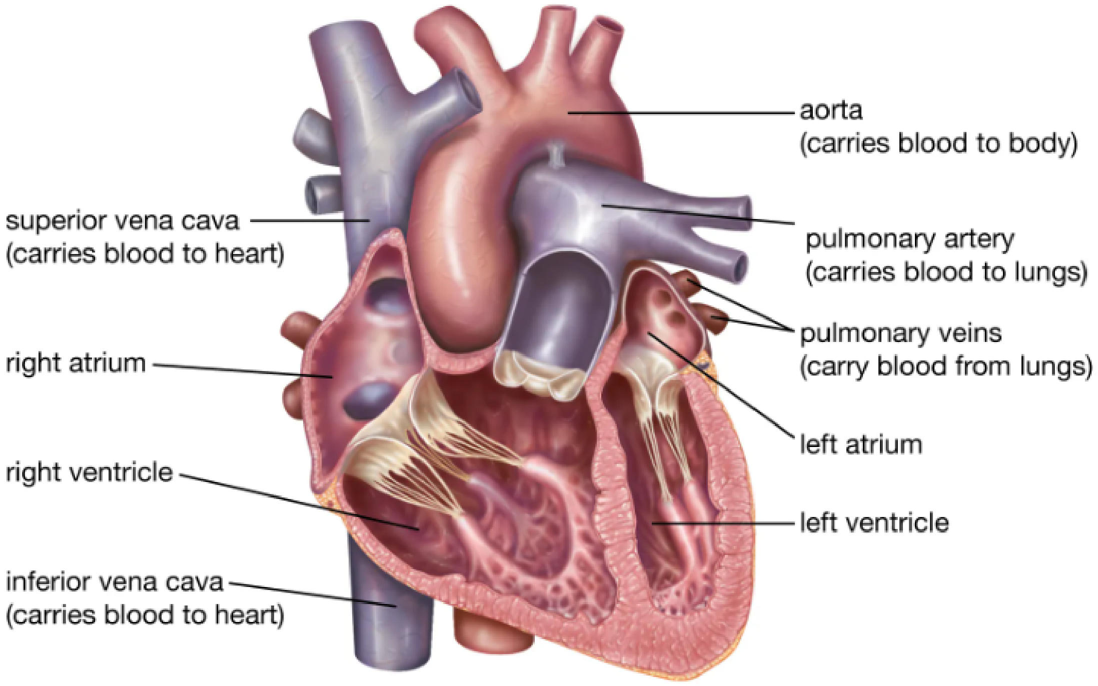


**Figure 2.3:** Schematic representation of the heart. Adapted from [22]

### 2.3.1 Electrocardiogram

The Electrocardiogram (ECG)) was reportedly first introduced in 1902 and, despite consisting in an initial less advanced form, allowed the medical and scientific community to collect objective information relative to how the human heart works. Later developments through the 20th century's first half led the ECG to evolve and improve to the current 12-lead electrocardiogram widely spread form [23]. Nowadays, the ECG plays a massive role in the first line of diagnostic and health monitoring for patients with cardiovascular malfunctions, such as arrhythmias, myocardial ischemia or infarction, to name a few [24]. Furthermore, with the advancement of new technologies, namely in the several fields of machine learning and data analysis, the ECG is currently being used for smart health monitoring systems, and other smart solutions, including cognitive stress assessment [25].

The ECG is a technique that records the electrical activity generated by the functioning of the heart through electrodes positioned at the surface of the body [26]. In the standard 12-lead ECG, which is currently one of the most clinical used configurations, ten electrodes are positioned on the patient's chest and limbs. The process behind the electrical signal collection is the cardiac cycle which, in short, consists of the atria and ventricles depolarization/contraction (the systole) and repolarization/relaxation (the diastole). This activity produces an electrical current in the heart that spreads along the adjacent tissues and is then captured by the electrodes positioned on the surface of the body throughout the ECG exam.

During the ECG signal collection, a set of waves correspondent to the cardiac cycle sequenced moments is captured. This set is the PQRST complex and represents an entire cardiac cycle. This complex starts with the P wave, which results from the atria depolarization, and the sequence continues with the QRS complex, which represents the ventricular depolarization. Finally, the T wave reflects the ventricular repolarization [26].

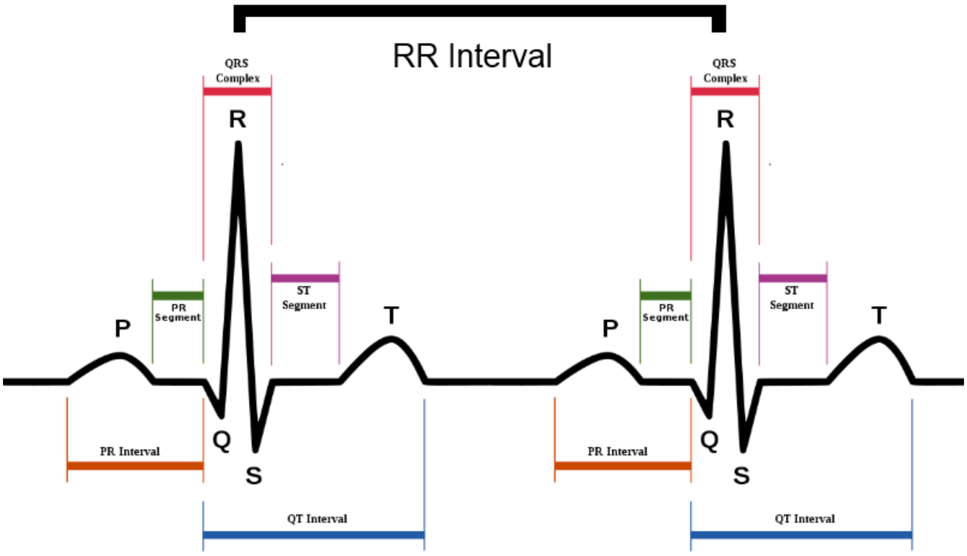


**Figure 2.4:** Schematic representation of two heartbeats captured by the ECG. Adapted from [27]

From the ECG signal is possible to extract different information, including the variance of the PQRST complex, the intervals between heartbeats and the heart rate, among other features that can be processed to evaluate the patient's health condition. One of the most studied measurements that can be computed from the ECG is the Heart Rate Variability (HRV), which has several applications ranging from health conditions evaluation to cognitive stress detection [28, 29].

### 2.3.2 Heart Rate Variability

The Heart Rate Variability (HRV) is a time series bio-signal that can be computed through the detection of the R-peaks present in the ECG signal. By definition, the HRV is the variance in the time duration of the intervals between consecutive heartbeats, known as R-R intervals (or NN intervals), which are measured in milliseconds (ms) [30]. The HRV can be influenced by several different physiological factors, such as the subjects' age, gender, ethnic group, lifestyle behaviour or the presence of chronic diseases [31]. Heart and respiratory rates have also been observed to affect the HRV [32, 33]. A healthy heart is expected to present chaotic dynamics. It is characterized by a higher HRV so that it can rapidly react to different internal or external stimuli, for instance, to respond to acute ischemia, imbalances in the metabolism or alterations in physical or mental activity [28]. Several health conditions have been observed to cause the decrease in the HRV, diminishing the heart's capability of replying to these stressful events [34].

In order to achieve the stability of the cardiovascular system and efficiently respond to sudden changes such as the previously mentioned, the ANS controls the heart rate, the blood pressure and other elements of this system that influence the HRV [28]. The R-R intervals are believed to behave like an index of the autonomic control [35] since they are influenced by the dynamic interaction between the parasympathetic and the sympathetic systems signals delivered to the heart (via the sinoatrial node). The parasympathetic activation reduces the heart rate in a process mediated by the synaptic release of acetylcholine. In contrast, sympathetic activation increases the heart rate, which is mediated by the synaptic release of noradrenaline. The metabolization of the acetylcholine occurs with short latency, and so the parasympathetic activation is associated with the higher frequencies of the HRV, whilst the sympathetic activation is related to the lower frequencies of the HRV since the noradrenaline reabsorption and metabolization are slower [28]. In this way, the HRV spectrum analysis can provide a measure of the sympathovagal balance [36]. The R-R intervals have also been reported to have a strong relationship with a subject cognitive load [37].

The mentioned facts make the HRV a non-invasive solid marker of the ANS activity. Several studies point to the HRV potential to be used not only for diagnosis and prognosis of health problems [38], but also for assessing a subject cognitive load. Several features can be extracted from the HRV time series across Time, Geometrical, Non-linear and Frequency Domains that can be used to assess the different ANS dynamics. These measurements can be extracted using different recording periods and are divided into three main categories, the long-term HRV measurements, lasting 24 hours; the short-term HRV measurements, which are recorded for around 5 minutes; and the ultra-short-term HRV measurements, which are the ones with the segments of analysis lasting under 5 minutes [39]. The time resolution of the HRV features will be further discussed in Chapter 3.

## 2.4 The Eye

The human eye is the sense organ responsible for collecting visual images. After being captured, the images are delivered from the eye to the brain, where they are processed and interpreted, allowing the human to be aware of his surroundings. Anatomically speaking, the eye resembles a sphere assembled with a small portion of a second transparent sphere with a higher curvature (cornea). Under the cornea lies the iris, which consists of a coloured ring of tissue; as the cornea is transparent, the iris colour defines the colour of the eyes. The structure that captures the light necessary to collect the visual images is the pupil, which is located in the centre of the iris. The pupil has a dark appearance, resultant of the fact that the light, which passes by the pupil into the eye, is almost entirely not reflected [40]. The pupil size (diameter) changes according to the environment's light intensity and other factors regulated by the ANS, namely by the sphincter and the dilator muscles [41]. The sympathetic activation induces pupil dilation, which improves the human long-range vision ability, while the parasympathetic activation results in pupil constriction, leading to improvements in the closer vision capabilities [10].

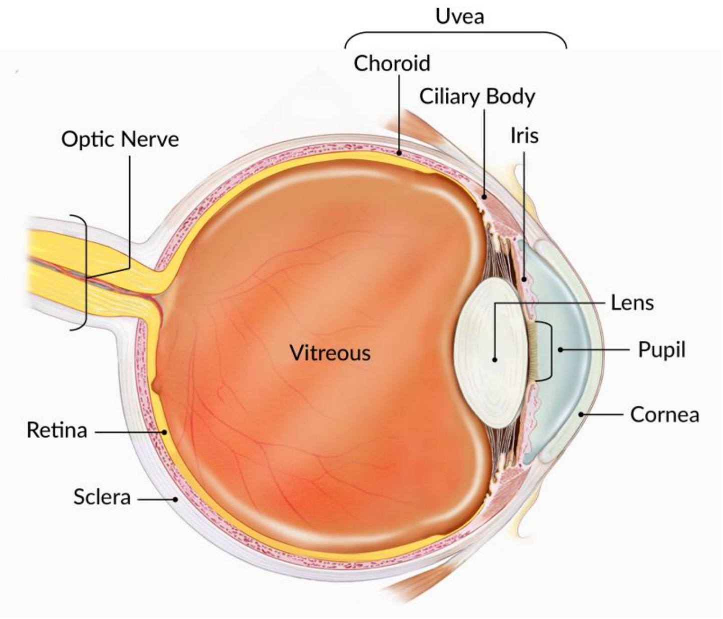


**Figure 2.5:** Schematic representation of the eye. Adapted from [42]

### 2.4.1 Eye-tracking and Pupillography

The optical channel carries almost 80% of all human collected sensory perceptions, and besides being the most abundant, the visual data is also the fastest to reach the brain. These exciting facts make the sight a very appetizing sense in many physiologic research fields [43]. However, despite some reported studies dating back to as early as the 19th century, the Eye-tracking technology was only first introduced at the start of the 20th century [44]. This technology at that time allowed the tracing of horizontal eye movements.

In the present day, the most common eye tracking devices use infrared cameras to detect the reflection on the pupil and cornea of near-infrared lights directed to the eyes. The eye tracking devices can provide x and y coordinates for the eye movements, allowing the detection of the exact location to where the subject is staring, with some devices also being capable of collecting the pupil size (diameter or area) [45, 46].

The collection of the pupil size variation by the eye tracking device results in a time-series signal, the pupillography signal. As mentioned before, the pupil size is controlled by muscles that are regulated by the ANS, i.e., which are managed by the parasympathetic and sympathetic nervous systems [47]. In this way, from the eye activity can be extracted different measurements, such as the blink duration, the blink rate, the pupil-size variability, the pupil diameter, the fixed staring duration, or other features resultant from the pupillography spectrum analysis, that can be used to assess the subject's ANS dynamics and act as a measure of the cognitive stress. These traits make pupillography increasingly popular, and many studies have been recently developed using this non-intrusive signal as an index of mental effort [7, 48].

# 3

# State of the Art

## 3.1 Overview

Software development is a cognitively demanding task requiring a lot of focus, decision making and logical reasoning. Conventionally, software development involves (usually) seven phases: planning, analysis, design, development, testing and maintenance. These phases compose the Software Development Life Cycle (SDLC); through the years, different strategies have been created to model the SDLC. Two of the most commonly used SDLC models are the waterfall and the agile models. The waterfall model essentially consists of a sequence of stages where a stage's output is the next stage's input. The agile models are processed incrementally and iteratively, allowing a quick response to requirements changes. Both methodologies have characteristics optimized for different software requirements. The agile models are ideal for developing small and useful software, having shorter delivery time frames, and the waterfall model is the best approach regarding larger projects with precise specifications [49]. However, software faults are practically unavoidable even when using adequate planning and delivering methodologies, making software reliability one of the most prevalent concerns in this industry [4, 50].

Nowadays, society relies hugely upon technology systems to perform various activities, including critical activities such as aviation control or assisted surgery. Behind these systems often lies complex software, so the concerns about software reliability have constantly been increasing, leading to the appearance of different approaches to avoid software faults (bugs), including verification, validation, software testing, and proof methodology strategies. The BASE project focuses exactly on fault avoidance, namely in fault prevention, fault removal, and fault forecasting [4]. It intends to prevent the faults before the software deployment using biofeedback from the programmers to identify the code sections more prone to having bugs.

Software faults typically occur in the most complex code areas, and as so, code complexity measures can be included in the fault avoidance techniques [4]. However, assessing the code complexity level is not a straightforward process. A study dating back to 1988 evaluating software complexity measures, performed by Elaine J. Weyuker, reports the strengths and weaknesses of different complexity metrics, including Halstead's effort metric and McCabe's cyclomatic metric [5]. The first method is based on the number of mental comparisons required to generate a program [51]. While the second is related to the decision structure of a program, accounting for the different control paths existing in the program [51, 52]. Unfortunately, both measures are revealed not to be sensitive enough to specific contexts and fail to consider interactions between different program components [5]. Using biofeedback from the programmer, the BASE project intends to identify the code sections' complexity according to the subject's cognitive stress. With this approach, the project expects to account for the several factors involved in software development and even allow personalized software faults detection models, signaling code areas where the programmer had a higher cognitive load as being more prone to bugs.

Several studies have been conducted throughout the years aiming to detect cognitive stress using biosignals. Some approaches include pattern recognition methods using electroencephalogram (EEG) spectral features to distinguish different levels of cognitive load [53]. Cognitive stress is known to affect ANS activity, with both the increment in sympathetic activity and the reduction in the parasympathetic activity being linked to a decline in performance during the execution of cognitively demanding tasks [37, 54]. This way approaches using biosignals controlled by the ANS are becoming increasingly popular. These signals can often be collected using non-intrusive devices, and some examples are Pupil Diameter (PD), Electrocardiogram (ECG), Electrodermal Activity (EDA) and Photoplethysmogram (PPG) [55,56]. In the present study, we will focus on the HRV (extracted from the ECG) and on the pupil size variation (Pupillography) to assess stress manifestations.

## 3.2 Heart Rate Variability measurements

Across different papers that approach the HRVanalysis, the most common features referenced in the time domain are mean NN (mNN), SDNN, SDSD, RMSSD, NN50 and pNN50 (see HRV features terminology in table 5.1). Regarding Power Spectrum Density analysis, the frequency domain is divided into three bands, the very-low-frequency band (VLF: under 0.04Hz), the low-frequency band (LF: 0.04 to 0.15Hz), and the high-frequency band (HF: 0.15 to 0.40Hz) [57]. The features extracted from each band most referenced in the literature are the total power and the peak. The ratio between the LF power and the HF power is also frequently mentioned.

Some of the previously mentioned features have already been linked to physiological dynamics. Starting with the VLF band is mentioned to be a heart's intrinsic nervous system consequence. The SDNN, as mentioned in [57], is influenced by every cyclic component responsible for variability in the recording period. This feature is highly correlated with the LF band, and the two are associated with both the sympathetic and parasympathetic systems dynamics. The LF band is as well linked to blood pressure regulation via baroreceptors. The features RMSSD, pNN50 and the HF band are also correlated and are closely influenced by the parasympathetic system. Thus, the ratio between LF power and HF power is believed to be a good measure of the balance between the sympathetic and parasympathetic systems. Although this belief is not consensual and this relationship is not as straightforward as some once believed, we can still look at this ratio as a metric of one system's predominance over another [6].

In addition to time and frequency domains, several authors also pursued the extraction of measures in the non-linear space in order to unveil non-linear HRV patterns. Based on the studies present in the literature, several measurements have been selected, focusing on their consistency when extracted using small time frames (e.g., 5 mins) [35, 58–60], which are: Approximate Entropy, Poincare' plot parameters (SD1 and SD2), Point Transition Measure, Katz Fractal Dimension and Higuchi Fractal

Dimension from the non-linear domain, Stress Index, HRV Triangular Index and TINN from the geometric domain were the ones selected.

### 3.2.1 Ultra-short-term HRV

Heart Rate Variability is conventionally used for the analysis of cardiac diseases in recordings lasting 24 hours, the long-term HRV measurements, or in 5 minutes recordings, the short-term HRV measurements [39]. Short-term HRV features (approximately 5 minutes in length) are already a standard and are currently well accepted as suitable time frames for extracting accurate HRV measurements [57]. However, the need to extract HRV measurements using time frames shorter than 1 minute (ultra-short-term HRV features) has grown due to several reasons [38,59,61]. Among these are the need to reduce the time spent and costs in the extraction of these indexes; the fact that they are incompatible with the dynamics of the physiological mechanisms to be captured (e.g., cognitive load spikes during code comprehension tasks execution); or the need to extract these features in new environments using modern wearable devices [61]. Also, the interest in using HRV in software engineering is growing very fast, and applications such as the identification of problematic code areas (that may have bugs and need revision) require a swift response in assessing programmers' cognitive load using HRV features [62]. To ensure a real-time response and to detect acute cognitive stress changes, we need time analysis windows as short as possible to achieve the required time resolution. This way, a wide range of new applications can benefit from the advances in the ultra-short-term measurements field and several studies have been conducted focusing on investigating the reliability of these ultra-short-term HRV features compared to the short-term ones.

In order to evaluate the ultra-short-term HRV measurements' reliability as a surrogate of the short-term HRV, different analyses can be performed. A procedure proposed by Pechia et al. [35] included a correlation analysis to test the existence of a significant association between features. If the correlation was significant and the correlation coefficient was above 0.07, perform a Bland-Altman plot to analyse the degree of bias. In case the data dispersion remains within the 95% line of agreement, the final step was to perform an effect size statistic (Cohen's d Statistic to parametric data or Cliff's Delta Statistic to non-parametric data). The feature is then considered a good surrogate if the effect size statistic test only detects minor differences. The mentioned procedure agrees with Shaffer et al. [38], which recommend using correlation/regression analyses paired with a Bland-Altman plot. Both works agree that only a correlation analysis is not enough to determine if an ultra-short-term HRV feature is a good surrogate of the short-term HRV. In fact, the two compared measurements can be highly correlated but have significantly different values.

A 2017 study carried out by Castaldo et al. [59] used Bland-Altman plots and Spearman's rank correlation analysis to assess which ultra-short-term HRV features are a valid surrogate of the short-term HRV. The study also built a machine learning model using ultra-short-term HRV features to discriminate between stress and rest states. The conclusions were that mean HR, the standard deviation of HR, mNN, SDNN, HF, and SD2 are appropriate short-term HRV surrogates for cognitive stress assessment. The paper also highlighted a machine learning model obtained using the mNN, the standard deviation of HR, and the HF features, which achieved an accuracy above 88%.

In Salahuddin et al. [39], the authors used mobile-derived ECG recording to extract several HRV measurements and the Kruskal-Wallis test to analyse the reliability of these measurements. It was "assumed that short-term analysis was not significantly different to the 150 seconds analysis if the p-value was greater than 0.05", and the goal was to find until which window span a feature is a good estimative of the 150-second window. The authors concluded that mean HR and RMSSD extracted using 10 seconds were not significantly different from the estimates using 150 seconds. This finding was also confirmed when using 20 seconds windows for extracting pNN50, HF, LF/HF, LFnu and HFnu features, 30 seconds for LF features and 50 seconds for VLF features. As for the remaining features studied by the authors, a minimum time frame of 60 secs was necessary for extracting features that were not significantly different from the 150 seconds reference features. This study data was recorded during the subject's day-to-day activities, like normal daily work, study, physical activities, and sleep.

In Baek et al. [63], a similar approach has been used to evaluate the reliability of ultra-short-term HRV measurements as short-term (5 minutes) HRV surrogates. The data was acquired in 5 minutes recordings while the subjects were "sitting at rest in a comfortable chair". In order to accomplish the proposed goal, the authors computed the p-value by the Kruskal–Wallis test, the Pearson correlation $r$ and the Bland–Altman plot analysis comparing the 5 minutes short-term measurements with the ultra-short-term ones with different time frames (270, 240, 210, 180, 150, 120, 90, 60, 30, 20, and 10 seconds). The highlighted features with the best results in this study were the mean HR, where 10 seconds windows were used to get results comparable to the 5 minutes analysis, the HF, which required 20 seconds windows, and the RMSSD, which required 30 seconds windows.

Following similar approaches, other works, such as the publications by Landreani et al. [61], Li et al. [64], Salahuddin et al. [65], Nussinovitch et al. [66] and McNames et al. [67], converged on a set of conclusions, where mean HR, mNN, SDSD, RMSSD, pNN50, HF, LF/HF, LFnu and HFnu were shown to be reliable under the 60 seconds recordings.

**Overall Remarks**

For the reported works, it is possible to conclude that ultra-short-term measurements are far from being consensual. Due to their extraction particularities only some features keep their stability under small window constraints. Additionally, it is still unclear what is the time frame limit for each HRV feature that can be applied to compute a reliable surrogate of its counterpart extracted from 5 minutes recordings. Furthermore, the studies found related to this topic have some limitations since they were developed with the subjects either at rest or performing elementary tasks in controlled environments, which are not expected in real-life contexts. In this work, we aim to elucidate these aspects and validate them under stressful and intellectually demanding environments, more precisely with the subjects performing software code inspection tasks (i.e., bug detection), which is a highly complex, dynamic, and cognitively demanding task. Our work's primary goal is to investigate the ultra-short-term HRV features to determine whether HRV-based tools can effectively be used in software development environments. To this extent, our present study investigates the smallest time frame, i.e., the shortest time resolution, where each feature is reliable. We also expect to investigate the ultra-short-term HRV measurements discrimination ability between two levels of acute cognitive stress (low and high complexity code sections) and how it is affected by the time frame reduction.

Another relevant aspect that is worth to be mentioned is that the existing studies perform an inter-subject statistical analysis of the features, i.e., perform the correlation or statistical analysis after concatenating the features collected from different subjects. This fact can lead to biased correlation values since it captures the inter-subject feature tendencies that may overwhelm the actual feature tendencies. In order to avoid this kind of bias, our study performs an intra-subject and intra-run feature statistical analysis.

## 3.3 Pupillography measurements

Pupillography measurements are becoming increasingly popular for clinical applications. One of the main advantages of this bio-signal is that it can be collected through non-invasive infrared video devices. Pupillography can be used, for instance, to detect visual field or afferent pupillary defects, accurately evaluating the pupil size variation (diameter) produced by changes in the environmental light intensity (Pupillary Light Response - PLR). Other pupillography applications include quantifying the autonomic effects of pharmaceuticals or measuring emotional responses such as fear, anger or stress, which have pupil dilation effects [68].

Several studies have investigated the relationship between mental activity and pupil size. A study by Hess et al. [69], dating back to 1964, experimented with using pupil size variation as a measure of cognitive load during simple problem-solving and concluded that the pupil size increases with the problem difficulty level (increase in cognitive load). In another study by Chen et al. [7], eight different features of eye activity were extracted as a measure of human cognitive load to discriminate between two cognitive stress levels. The features present in the experiment were the mean and the standard deviation of the pupil size, the blink latency and the blink rate, the fixation time and rate, and finally, the saccade size and speed. The investigation concluded that the present features have significant discriminative power in discerning the two cognitive stress levels.

In Pedrotti et al. [70], the authors followed a different approach using pupillography frequency domain features to distinguish four different driving tasks with different levels of complexity. The experiment showed a four-way parallel neural network classifier that obtained 79.2% precision using these features. This study also mentioned the lack of consensus regarding the frequency bands of the pupillography and their approach using wavelet transforms. In fact, several papers can be found using different bands for the frequency band analysis.

In an investigation conducted by Lüdtke et al. [71], after the detection and removal of blink artefacts, a frequency analysis of the pupillography was performed using a Fast Fourier Transform (FFT) in the 0.0 to 0.8Hz frequency region. The region was split into eight bands of 0.1Hz length, which were used to discern sleepiness from alert states. Significant differences were found between the alert and the sleepy groups using Mann–Whitney U-test. Another different study by Nakayama et al. [72] demonstrated the increase in Power Spectrum Density (PSD) in the regions of 0.1 to 0.5Hz and 1.6 to 3.5Hz with the complexity escalation of oral calculation tasks.

In Murata et al. [47], the low-frequency band was considered from 0.05 to 0.15Hz, while the frequencies ranging from 0.35 to 0.40Hz belonged to the respiration frequency band. The ratio between those two bands was employed to evaluate the cognitive load. Furthermore, in a 2015 study by Peysakhovich et al. [73], a similar approach used the LF/HF ratio to assess the subjects' cognitive load. The ratio was concluded to be sensitive to cognitive load and not affected by changes in the environmental light. This study considered frequencies ranging from 0.0 to 1.6Hz as part of the LF band, while frequencies within 1.6 to 4.0Hz belong to the HF band.

In order to use frequency domain features in the pupillography analysis, we need to define the correct frequency bands' limits in our experiment context. A 2004 study by Lee et al. [74] used HRV as ground truth to evaluate different pupil size estimation methods. In the experiment, the approach was to compute the correlation between the LF/HF ratio extracted from the HRV and the same ratio extracted from the pupil size variation spectrum, using different methods for pupil size estimation. The LF and HF pupillography bands considered were the same as the HRV (LF: 0.04 to 0.15Hz and HF: 0.15 to 0.40Hz), and the study obtained a maximum correlation of 69% between the LF/HF extracted from the HRV and the same ratio computed from the pupil size variation spectrum.

**Overall Remarks**

Through our research, we can conclude that the frequency bands' limits of the pupillography signal are neither well defined nor consensual since it is possible to find several studies using different bands' limits. In order to capture the acute ANS variations with the cognitive stress increase or decrease, namely the balance between the sympathetic and parasympathetic nervous systems, it is essential to use the correct frequency bands. As stated in subsection 2.2.2, sympathetic activation is related to the lower frequencies (LF band), and parasympathetic activation is associated with the higher frequencies (HF band) [28]. In the HRV signal, the two the LF band (0.04 to 0.15Hz) and the HF band (0.15 to 0.40Hz) are well-documented [57]. Knowing that the HRV and the pupil size variation are both controlled by the ANS [10], in the present study, we decided to follow an approach using the HRV as the ground truth for the pupillography frequency bands definition. The approach consists of varying the pupillography LF and HF band limits to find the pupillography frequency bands that maximize the correlation between these pupillography bands' power and the same bands' power computed from the HRV.

# 4

# Data Acquisition Protocol

## 4.1 Participants

The data used in the current work was collected in the scope of the BASE project and aimed at the research of error making and error discovery during software inspection tasks. These datasets contain biometrical signals extracted using functional magnetic resonance imaging (fMRI) and other non-invasive sensors including Electrocardiogram (ECG), Electroencephalogram (EEG), Photoplethysmogram (PPG), Electrodermal activity (EDA) and eye movements (Eye Tracking). In the current study, we will focus on the analysis of the HRV computed from the ECG signals and on the Pupillography signals obtained from the eye tracking during the periods associated with the code sections inspected by the subjects.

In order to collect the data used in the study, we opened a call for participation in the experiment. Through this process, we obtained 49 candidates consisting of a mixture of students (pursuing PhDs and MSs in different computer science fields), academic professors, and professional specialists in the software sector (code reviewers). The candidates were then interviewed and screened to guarantee their fitment to the study objectives. During the interview, demographic and biometric characteristics (e.g., age), professional status, programming experience, availability and motivation were accessed. Subsequently, each candidate's proficiency level has also been accessed based on the score provided by two questionnaires: 1) Programming experience questionnaire and 2) Technical questionnaire. The first questionnaire aimed to assess the candidate's programming experience based on the candidate's coding volume in the last three years. The second questionnaire's goal, composed of

10 questions, was to assess the candidate's coding skills. The programming experience gives us an overall idea of the experience in the past years from the candidate:

1. Experience in SW programming (Number of years)
2. Coding lines programmed in any language in the last 3 years (approximate number)
3. Coding lines programmed in C in the last 3 years (approximate number)
4. Coding lines written in the most extensive C program written (approximate number)

On the other hand, the technical questionnaire was used for the candidate characterization regarding the present knowledge and coding skills, which is, therefore, more helpful in selecting and classifying the candidates. Based on the results obtained in these questionnaires, the candidates with a score below 3 (out of 10) were considered not eligible since they were not representative of software industry professionals. The remaining ones were characterized as non-experienced (score between 4 and 7) and experienced (score between 8 and 10).

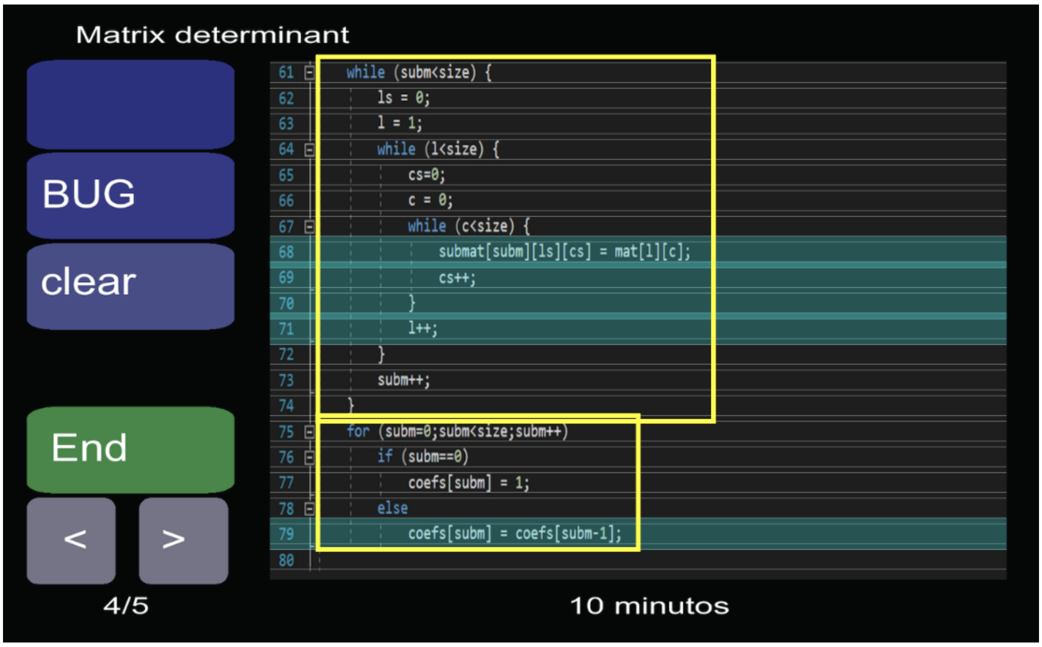
In summary, 21 male subjects, ranging from 19 to 40 years, with a median of 22 years, were selected for the experiments after the screening process.

All subjects provided a written informed consent, and all the data has been anonymized. This study was approved by the Ethical Committee of the Faculty of Medicine of the University of Coimbra, following the Declaration of Helsinki and the standard procedures for studies involving human subjects.

## 4.2 Experimental protocol and setup

The selected candidates were submitted to 4 different runs of code inspection tasks using 4 code snippets written in C code language (selected randomly at each run). Each run starts with a fixation cross in the middle of the screen for 30 seconds. Subsequently, three tasks are presented to the subject: a natural language reading (literary excerpt) task, a neutral (bug-free and straightforward code) code reading task, and one code inspection (code with bugs) task. The order of the presentation was randomly selected to avoid biasing the results, following a randomized control crossover design. Between each task and at the end of each run, a fixation cross is presented to the subject for 30 seconds. The description of each task is provided as follows:

1. **Natural language reading** - In this task, a text in natural language is presented to the subject (selected randomly from the existing 4 different texts) for 60 seconds. The presented texts were selected in order to have neutral characteristics and avoid measurement fluctuations induced by narrative-triggered emotions.
2. **Simple code snippet reading** - In this task, the subject is presented with a simple and iterative code snippet (selected randomly from the set of 4 different neutral code snippets) for 300 seconds. The presented code snippets were selected with the objective of inducing the subject into a low cognitive effort state which will be used as a reference state during the posterior analysis.
3. **Code inspection** - In this task, a code snippet in C language is displayed to the subject (selected randomly from a set of 4 different code snippets of different complexities) for a maximum of 600 seconds. In this task, the subject is asked to analyze and inspect the code aiming for bug detection (see code snippet example in figure 4.1).



```
while (subm<size) {
    ls = 0;
    l = 1;
    while (l<size) {
        cs=0;
        c = 0;
        while (c<size) {
            submat[subm][ls][cs] = mat[l][c];
            cs++;
        }
        l++;
    }
    subm++;
}
for (subm=0;subm<size;subm++)
    if (subm==0)
        coefs[subm] = 1;
    else
        coefs[subm] = coefs[subm-1];
```

**Figure 4.1:** Example of a code snippet inspected during the experiment.

The schematic representation of each run is provided below (see figure 4.2).

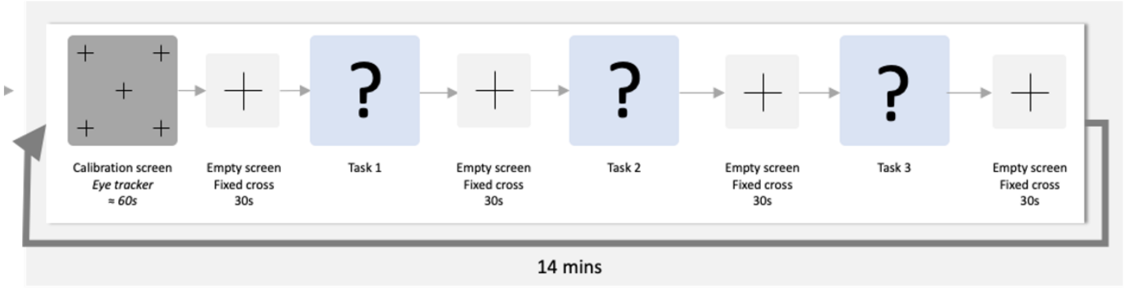


**Figure 4.2:** Schematic representation of an experiment run.

Each run had a duration of about 21 minutes, meaning that the whole protocol lasted about 1 hour and 20 minutes. During the experiment, the subjects were alone in a quiet, isolated room when performing the tasks. Furthermore, the subjects were informed apriori about all the protocol and processes of the experiment and were instructed not to take anything that could stimulate/inhibit them the day before the experiment. The code inspection tasks were presented to participants using the Vizard software [75].

The equipment used to collect the Electrocardiogram (ECG) signal was the Maglink RT (Neuroscan) with a sampling frequency of 10 kHz [76] (see equipment set up in Figure 4.3). For the ECG signal acquisition, the electrodes from Neuroscan equipment were positioned in the V1 and V2 locations. The EyeLink 1000 Plus Eye Tracker (with Long Range mount display) with a sampling frequency of 500 Hz was the equipment utilized to acquire the pupilogram and eye movements [77].

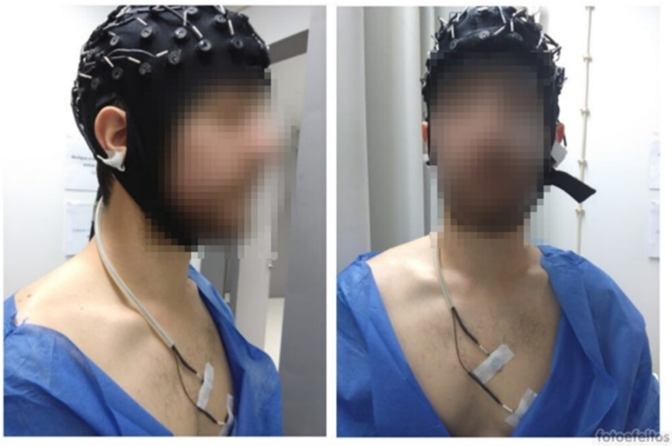


**Figure 4.3:** Equipment set up used in the experiment.

# 5

# Heart Rate Variability measurements

This chapter contains the study conducted to investigate the ultra-short-term HRV measurements reliability following two different approaches. The first approach used a statistical analysis (statistical significance tests, correlation tests and Bland-Altman plots) to investigate the lower time frame where each feature is reliable. The second approach examines the impact of reducing the window size in the complexity classification of software code sections. This chapter is divided into three main sections: Methods, where the methodology used in the experiment is described; Results, where the obtained results are presented; and Discussion, where the results obtained are discussed.

## 5.1 Methods

This section describes the practical steps and methods performed to investigate the reliability of the ultra-short-term HRV measurements. The ECG signal is pre-processed and segmented to obtain the R-R time series (HRV) before the ultra-short-term HRV measurements extraction using 18 different time frames. These measurements' reliability is then investigated using a statistical analysis approach (statistical significance tests, correlation tests and Bland-Altman plots) and a classification approach (with linear SVM classifiers). Figure 5.1 represents the general flow chart of the experimental steps followed for the HRV measurements quality and reliability study using the statistical and classification approaches.

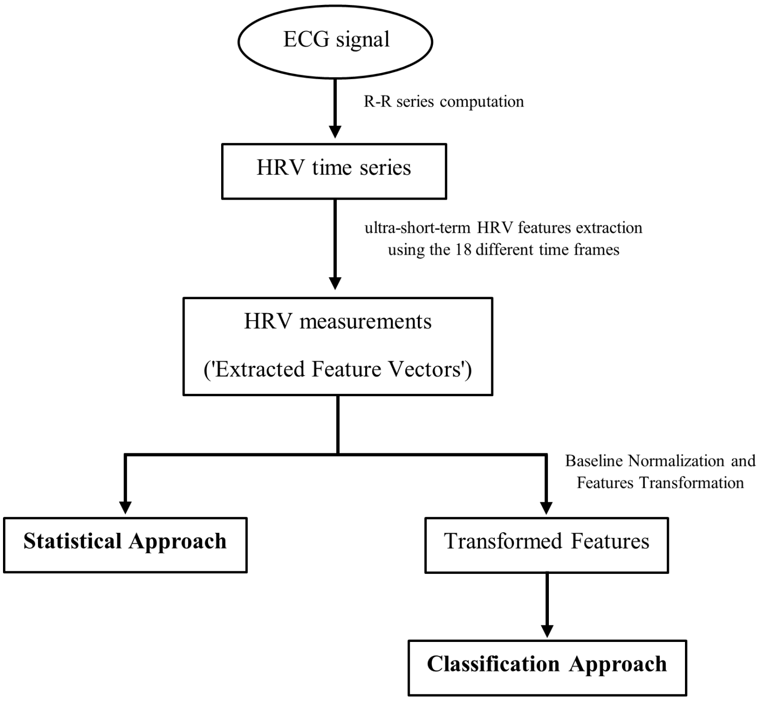

ECG signal
R-R series computation
HRV time series
ultra-short-term HRV features extraction using the 18 different time frames
HRV measurements ('Extracted Feature Vectors')
Baseline Normalization and Features Transformation
Statistical Approach
Transformed Features
Classification Approach

**Figure 5.1:** General Flow Chart of the experimental steps followed for the HRV measurements quality and reliability study.

### 5.1.1 Pre-processing and ECG segmentation

During the present experiment, several different biosignals were collected. Functional Magnetic Resonance Imaging (fMRI) was one of the exams performed. This exam forces the experiment to be conducted inside an MRI scanner. The fMRI has a noise effect in the ECG signal, which induces the gradient artifact (GA) on this signal. In order to remove the GA, we performed an average artifact subtraction (AAS) method based on the algorithm from Niazy et al. [78], which allows the mitigation of the GA effect. Furthermore, the MRI scanner produces a magnetic field which impacts and alters the ECG morphology, making the T-wave more extensive than the QRS complex and reducing the R-wave amplitude. These changes in the ECG morphology lead to the failure of the traditional QRS detection algorithms, which ultimately leads to the incorrect R-R intervals computation.

In these scenarios, the R-peak detection method proposed by Christov et al. [79] is frequently employed. This algorithm has proven to be robust to the changes in the ECG morphology, and it achieves high performance in the R-peak detection on ECG signals collected inside the MRI scanner. This way, we utilised the Christov et al. [79] algorithm for the R-peak detection and visually inspected the data obtained to examine the R-peak detection method quality. After concluding the R-peak detection process, we computed the R-R intervals to achieve the HRV time series.

### 5.1.2 Feature Extraction

In order to carry out the HRV analysis, following pre-processing and ECG segmentation, we proceed with the features extraction from the Code inspection data collected during each subject run. This data corresponds to the most cognitively demanding task of the current experiment, where the subject inspects code snippets, having sections of different complexity levels, aiming to find software faults (bugs). A total of 31 features across Time, Geometrical, Non-Linear and Frequency domains were extracted using a sliding window of variable size and a jumping step of 1 second. The sliding window size used was ranged from 3 minutes to 10 seconds, being iteratively reduced by 10 seconds, making up a total of 18 different windows (see Figure 5.2). All 31 features were extracted by applying the 18 different sliding windows. The 3 minutes (180-second) sliding window was used as the gold standard in the statistical and classification approaches. This time frame was the larger window size since the study was performed during a highly complex, dynamic and cognitively demanding task (code inspection). In the present context, a 5 minutes time frame is a considerably large window. A window of this size would capture physiological data corresponding to more than one code section, where the subject could feel different difficulty levels, leading to inaccurate results since it would capture different ANS dynamics.

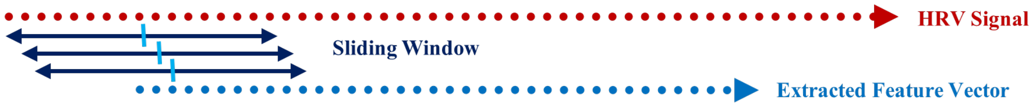


**Figure 5.2:** Schematic representation of the extraction of a feature using one of the sliding windows. In the end we obtain a total of 558 feature vectors, corresponding to the 31 features times 18 window sizes, for each experiment run.

The described extraction procedure produces vectors of individual measurements from the HRV data collected during the Code inspection task (to facilitate referencing, we will call these the ‘Extracted Feature Vectors’). Each individual measurement is computed based on a RR signal portion with the size of the sliding window employed. The individual measurements are then associated with the time instant corresponding to the center of the RR signal portion used to compute the respective individual measurement.

It is important to mention that the same RR signal originates ‘Extracted Feature Vectors’ of different lengths accordingly to the time frame applied in the extraction process. The vector obtained with the 180-second sliding window is the one with fewer individual measurements, while the vector extracted with the 10-second sliding window is the larger, having more individual measurements. In this study, the ‘Extracted Feature Vectors’ are directly used in the statistical approach. In the statistical analysis, these measurements do not need to be normalized since the statistical analysis is performed run by run and subject by subject, not being impacted by the inter-run and inter-subject variability. For the classification approach, these ‘Extracted Feature Vectors’ were baseline normalized before the feature transformation used in the classification process.

#### 5.1.2.1 Feature Description

The set of 31 features in study includes six features from the Time Domain, three from the Geometrical Domain, six from the Non-Linear Domain, and 16 from the Frequency Domain. The different features were selected based on the current literature on the subject of Ultra-Short-Term HRV measurements and are the result of a search for the most reliable features extracted using small time frames. This section briefly describes the 31 features present in the current study (summarized in Table 5.1).

1. **Time Domain Features**

   In this experiment, we computed the standard six most referenced features in the literature from the time domain: mean of NN (or RR) intervals (**mNN**), standard deviation of NN (or RR) intervals (**SDNN**), the standard deviation of the differences between heart beats (**SDSD**), the root mean square of the differences between heart beats (**RMSSD**), the number of consecutive RR intervals differing more than 50 milliseconds (**NN50**) and the proportion of consecutive RR intervals differing more than 50 milliseconds (**pNN50**) [30,35, 57,80].

- The mean NN (mNN) is precisely the mean of the RR intervals, which can be computed by [80]:

$$mNN = \frac{1}{N}\sum_{n=1}^{N} RR_n \quad (ms) \tag{5.1}$$

- The standard deviation of RR intervals (SDNN) can be obtained by the formula [80]:

$$SDNN = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(RR_n - mNN)} \quad (ms) \tag{5.2}$$

  This measurement represents the HRV time series short-term and long-term variations [80].

- In the standard deviation of the differences between heartbeats (SDSD), we first must compute the difference between adjacent heartbeats and then calculate the standard deviation of the resulting values [57]. It is worth mentioning that the vector resulting from the computation of the difference between adjacent heartbeats will have one element less relative to the original RR vector. The formula used for the SDSD computation is the following [80]:

$$SDSD = \sqrt{E\{\Delta RR_n^2\} - E\{\Delta RR_n\}^2} \quad (ms) \tag{5.3}$$

  In contrast to the SDNN, the SDSD characterizes the short-term (beat-by-beat) variability [80]. It is also important to note that for stationary RR series: $E\{\Delta RR_n\} = E\{\Delta RR_{n+1}\} - E\{RR_n\} = 0$. This means that SDSD equals the root mean square of the differences between heartbeats (RMSSD) for stationary RR series [80]. The SDSD and the RMSSD measurements present a high correlation between the two.

- The root mean square of the differences between heart beats (RMSSD) is calculated by the following expression [80]:

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}(RR_{n+1} - RR_n)^2} \quad (ms) \tag{5.4}$$

- In order to obtain the NN50 feature, the number of consecutive RR intervals differing more than 50 milliseconds is counted. The pNN50 feature is the percentage of consecutive RR intervals differing more than 50 milliseconds [80]:

$$NN50 = count(|RR_{i+1} - RR_i| > 50ms) \tag{5.5}$$

$$pNN50 = \frac{NN50 \times 100}{N-1} \quad (\%) \tag{5.6}$$

2. **Geometrical Domain Features**

The RR intervals sequence can form a geometrical pattern that geometrical domain features can characterize [81]. From the geometrical domain, three features were extracted, the HRV Triangular Index (**TI**), the Triangular Interpolation of RR (or NN interval) Histogram (**TINN**) and the Baevsky's Stress Index (**SI**).

- To determine the HRV Triangular Index (TI), we start by computing a Histogram of RR intervals with a bin size of 1/128 seconds [81]. Then the HRV Triangular Index will be given by the total number of RR intervals D divided by the absolute frequency Y of its most frequent value X (the mode) [81]:

$$TI = \frac{D}{Y} \tag{5.7}$$

- The Triangular Interpolation of NN Histogram (TINN) consists of approximating the distribution histogram by a triangle, with the same bin size as the TI measurement. Its value corresponds to the baseline width of the NN interval histogram [81]. In order to compute the TINN feature, the following expression is used [82]:

$$TINN = M\text{–}N \tag{5.8}$$

In the above expression, M and N represent the triangular function T vertices, where $T(t) = 0$ for $t <= N$ and for $t >= M$, and in the modal bin, $T(X) = Y$. T obtains the values of linear functions with the connection of the points (N, O) with (X, Y) and (X, Y) with (M, O). The M and N values are defined by the triangular function best fitting the sample distribution [82].

- Regarding the Baevsky's Stress Index (SI), the expression used for its computation is the following [80, 83]:

$$SI = \frac{Amo \times 100\%}{2Mo \times MxDMn} \tag{5.9}$$

In the previous expression, Amo stands for the mode amplitude which is denoted in the percentage form, Mo represents the mode, being this the RR interval that appears more frequently in the series and MxDMn which constitute the variation scope, this last variable presents the degree of RR interval variability of the heart.

3. **Non-Linear Domain Features**

Referring to the non-linear measurements, we computed a total of 6 features: the Approximate Entropy (**ApEn**), **SD1** and **SD2** from Poincare Plot Parameters, the Point Transition Measure (**PTM**), the Katz Fractal Dimension (**KFD**) and the Higuchi Fractal Dimension (**HDF**).

- The Approximate Entropy (ApEn) is a measure of the complexity or irregularity of the RR series. The algorithm for the computation of this feature consists of given a series of N RR intervals $(RR_1, RR_2, \ldots, RR_N)$, a series of vector of length m $(X_1, X_2, \ldots, X_{N-m+1})$ is defined from the RR intervals: $X_i = [RR_i, RR_{i+1}, \ldots, RR_{i+m-1]}$. Then the distance $d[X_i, X_j]$

between the vectors $X_i$ and $X_j$ will be established as the maximum absolute difference between their respective scalar components [84].
After this step, for every vector $X_i$, the relative number of vectors $X_j$ for which $d[X_i, X_j] \leq r$, $C_m^i(r)$ is calculated using the expression that follows:

$$C_m^i(r) = \frac{number\ of\{d[X_i, X_j] \leq r\}}{N - m + 1} \tag{5.10}$$

Here r represents the tolerance value. Subsequently, the index $\phi^m(r)$ is computed by taking the natural logarithm of each $C_m^i(r)$ and averaging them over i, as shown in the following expression:

$$\phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} ln\, C_m^i(r) \tag{5.11}$$

In the end, the approximate entropy is estimated using the next formula [84]:

$$ApEn(m,r,N) = \phi^m(r) - \phi^{m+1}(r) \tag{5.12}$$

- The Poincare Plot is a geometrical and non-linear method regularly used to analyze HRV and to represent the correlation between successive inter-beats intervals. This method gives origin to a scattered plot that represents the beat-to-beat intervals against the prior intervals, i.e., a point in the Poincare Plot will be the $RR_i$ on the x-axis and the $RR_{i+1}$ on the y-axis. In order to reach a quantitative analysis of the Poincare Plot, we proceed with the adjustment of an ellipse to the scattered plot, obtaining the standard descriptor, SD1 and SD2 [58].

  SD1 is the standard deviation of the Poincare plot perpendicular to the line-of-identity [35]. This feature represents the standard deviation of short-term inter-beats interval variability, which is referenced to emphasize the short-term dynamics of HRV [58]. In order to perform the SD1 computation, the following expression is used:

$$SD1 = \frac{\sqrt{2}}{2} std(x_i - x_{i+1})^2 \tag{5.13}$$

In this expression std represents the standard deviation of the HRV time series and x stands for the time interval between successive beats.
On the other hand, SD2 is the standard deviation of the Poincare plot along the line-of-identity [35]. In contrast to the SD1, this feature represents the standard deviation of long-term inter-beats interval variability, referenced to illustrate the long-term dynamics of HRV. The expression applied to compute SD2 is the following [58]:

$$SD2 = \sqrt{2std(x_i)^2 - \frac{1}{2} std(x_i - x_{i+1})^2} \tag{5.14}$$

- The Point Transition Measure (PTM) is a new feature proposed by Zubair et al. [58] with the goal of quantifying not only the spatial information but also the temporal variation at the point-to-point level of the Poincare plot. In this way, the PTM tries to overcome the SD1, and SD2 measurements limitation since these two features only present spatial information and do not include temporal variation. In order to compute the PTM measurement, two successive points of the Poincare plot are used, with a moving window being applied to draw these two successive points.
  To compute this feature, in the first place, we must calculate the vector's length and angle between two points. Then, the effects of these values are integrated using the following expression so that a single value can be obtained [58]:

$$PTM = \frac{1}{N} \sum_{i=1}^{N} (l_i)(sin(\frac{\theta_i}{4})) \tag{5.15}$$

In this equation, the N value corresponds to the total number of vectors, l is the length computed between two successive points,  corresponds to the angle of the vector and the application of the number 4 is derived from the repetition of the angle.

- Two features regarding the Fractal Dimension were computed using two different algorithms, the Katz Fractal Dimension (KFD) and the Higuchi Fractal Dimension (HFD). These algorithms numerically classify waveforms, such as the HRV time series, by assessing their fractal dimensionality [85,86].

  With the purpose of computing the Katz Fractal Dimension, two variables need to be known: the total length of the curve 'L' and the 'd', which stands for the planar extent or diameter of the waveform. The total length of the curve 'L' is simply the sum of the distance between successive waveform points [85]. To this extent, the Euclidian distance is used:

$$d(s_1,s_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{5.16}$$

  In this expression, $(x_1 - x_2)$ is equivalent to one in all samples. 'L' is then computed as the sum of all Euclidian distances between two successive points.

  The distance between the beginning point of the sequence and the point that offers the greatest distance serves as a rough approximation for the waveform's planar extent or diameter, 'd'.

$$d = max(dist(1,i)) \quad i = 2,\ldots,N \tag{5.17}$$

  With N corresponding to the total number of points. Having these two parameters, we can proceed to the Kratz Fractal Dimension computation, applying the expression [85]:

$$KFD = \frac{log_{10}(n)}{log_{10}(n) + log_{10}(d/L)} \tag{5.18}$$

  In the previous equation, 'n' represents the number of steps in the waveform, which is equal to the subtraction of 1 to the total number of points $(n = N - 1)$.

  The Higuchi Fractal Dimension (HFD) consists of computing the mean length of the curve for every k-sample set and, after that, in log-log scale, making a plot of the length curve against the k parameter. The resulting slope of the graphic will express the Higuchi Fractal Dimension (HFD) [86].

In this way, the estimation of the optimal maximum parameter k, *kmax*, is needed. In order to find this parameter, a search was conducted throughout plotting HFD values against different *kmax* [86]. We concluded that no more improvements were observed after a *kmax* value equal to 50, and the HFD value remained stable. It is also important to mention that the *kmax* parameter must be less than half the window length since it is the maximal distance between compared instances [87]. The analysis windows in the current study range from 10 to 180 seconds. In this way, a *kmax* value of 50 cannot be used in any window with a size below 100 seconds. Saying so, regarding the windows from 180 to 100 seconds, the $kmax = 50$ was applied in the HFD computation. Meanwhile, the *kmax* used was precisely half the respective window size for the other time frames in the present study.

4. **Frequency Domain Features**

In order to investigate how Power distributes as a function of frequency, a Power Spectral Density (PSD) estimation was performed using Burg's autoregressive method with order 16. The order was assessed using the partial autocorrelation sequence. The frequency bands recommended by the 1996 Task Force [57] were the ones used in this study: the very-low-frequency band (VLF), which includes frequencies under 0.04Hz, the low-frequency band (LF), with frequencies ranging between 0.04Hz and 0.15Hz, and high-frequency band (HF) were the frequencies are within the 0.15Hz to 0.4Hz range. After the three different frequency bands computation, the Power features were extracted by calculating the area under the PSD curve [30, 35, 57].

- The total Power (**totPow**) is the sum of the Power of every frequency band.

- The **Peak** is the maximum Power across every frequency of interest.
- **VLF**, **LF** and **HF** are respectively the sums of the total Power on the very-low-frequency, low-frequency, and high-frequency bands.
- VLF normalized (**VLFnu**), LF normalized (**LFnu**), and HF normalized (**HFnu**) are the frequency bands relative Power. These features correspond to the VLF, LF and HF measurements normalized by the total Power.
- **VLFpeak**, **LFpeak** and **HFpeak** are respectively the maximum Power on the very-low-frequency, low-frequency, and high-frequency bands.
- VLFpeak normalized (**VLFpeak-nu**), LFpeak normalized (**LFpeak-nu**) and HFpeak normalized (**HFpeak-nu**) are the frequency bands relative peaks. These features correspond to the VLFpeak, LFpeak and HFpeak divided by the Peak.
- The ratio between the LF band's maximum Power and the HF band's (**LFpeak/HFpeak**) consists of the quotient between the LFpeak and the HFpeak features. In a similar way, the ratio between the total Power of the LF band and the total Power of the HF band (**LF/HF**) is the quotient between the LF and the HF measurements.

**Table 5.1:** Set of Features used in the current study presenting the designation used across the document, the units of measurement, a description of the feature and the papers reporting that feature for the analyse of HRV.

| HRV Features Initials | Units | HRV Features Description | References |
|---|---|---|---|
| Time Domain | | | |
| mNN | [ms] | mean of NN (or RR) intervals | [35] |
| SDNN | [ms] | standard deviation of NN (or RR) intervals | [35, 57] |
| SDSD | [ms] | standard deviation of the differences between heart beats | [35, 57] |
| RMSSD | [ms] | the root mean square of the differences between heart beats | [35, 57] |
| NN50 | – | number of consecutive RR intervals differing more than 50 milliseconds | [35, 57] |
| pNN50 | [%] | proportion of consecutive RR intervals differing more than 50 milliseconds | [35, 57] |
| Geometrical Domain | | | |
| TI | – | HRV Triangular Index - integral of the NN interval histogram divided by the height of the histogram | [35, 57, 81, 82] |
| TINN | – | Triangular Interpolation of RR (or NN interval) Histogram - baseline width of the NN interval histogram | [35, 57, 81, 82] |
| SI | – | The Baevsky's Stress Index | [80, 83] |
| Non-Linear Domain | | | |
| ApEn | – | Approximate Entropy - measures the complexity or irregularity of the RR series | [84] |
| SD1 | [ms] | Standard Deviation of the Poincare' plot perpendicular to the line-of-identity | [35, 58] |
| SD2 | [ms] | Standard Deviation of the Poincare' plot along the line-of-identity | [35, 58] |
| PTM | – | Point Transition Measure - quantifies the temporal variation at the point-to-point level of the Poincare plot | [58] |
| KFD | – | Katz Fractal Dimension | [85] |
| HFD | – | Higuchi Fractal Dimension | [86] |
| Frequency Domain | | | |
| VLF | [ms$^2$] | Very-Low Frequency band power ($\leq 0.04Hz$) | [30, 35, 57] |
| LF | [ms$^2$] | Low Frequency band power (0.04 - 0.15 Hz) | [30, 35, 57] |
| HF | [ms$^2$] | High Frequency band power (0.15 - 0.4 Hz) | [30, 35, 57] |
| VLFnu | n.u. | VLF power normalized | [30, 35, 57] |
| Lfnu | n.u. | LF power normalized | [30, 35, 57] |
| HFnu | n.u. | HF power normalized | [30, 35, 57] |
| VLFpeak | [ms$^2$] | VLF power frequency peak | [30, 35, 57] |
| LFpeak | [ms$^2$] | LF power frequency peak | [30, 35, 57] |
| HFpeak | [ms$^2$] | HF power frequency peak | [30, 35, 57] |
| VLFpeak-nu | n.u. | VLF power frequency peak normalised | [30, 35, 57] |
| LFpeak-nu | n.u. | LF power frequency peak normalised | [30, 35, 57] |
| HFpeak-nu | n.u. | HF power frequency peak normalized | [30, 35, 57] |
| totPow | [ms$^2$] | Total Power | [30, 35, 57] |
| Peak | [ms$^2$] | Overall frequency power peak | [30, 35, 57] |
| LF/HF | – | Ratio of LF and HF band powers | [30, 35, 57] |
| LFpeak/HFpeak | – | Ratio of LF and HF band power frequency peak | [30, 35, 57] |

#### 5.1.2.2 Normalization

In order to reduce the inter-subject and inter-run variability, i.e., the variability between different subjects and the internal variability of a subject during the experiment (different runs), all the features extracted during the "code inspection" task were normalized before the feature transformation used in the classification approach.

The data collected during the "natural language reading" periods was used as a baseline for the normalization process. With this intent, the features present in the current study (table 5.1) were also extracted from the data collected during the "natural language reading" task periods, using a similar procedure as explained in section 5.1.2. To facilitate the reference, let us call these the 'rest features' and the features regarding the "code inspection" task the 'code features'. In the "natural language reading" period, the subjects are supposed to be in a low cognitive stress state, which corresponds to an optimal state for the normalization process. The normalized features were obtained by calculating the ratio between each "code feature" and the corresponding "rest feature" median. The median has been selected to perform this computation since the data does not follow a normal distribution (assessed using the Kolmogorov–Smirnov test).

#### 5.1.2.3 Feature Transformation

As mentioned in chapter 4, during the "Code inspection" task, the subjects inspected a code snippet in C language, aiming for bug detection. The code snippets contain different code sections with different complexity levels. The code sections are labelled as low or high complexity according to a classification performed by a panel of experts. In order to capture and enhance the cognitive stress state presented by the subjects at the different code complexity sections, statistical feature transformations were computed. To this extent, the individual measurements present in the 'Extracted Feature Vectors' (produced in the feature extraction process, see section 5.1.2), after the normalization process (section 5.1.2.2), were grouped based on all the instants the subject was looking to a specific section during a run. From each group, a set of features was computed employing statistic transformations:

- Simple statistic transformations:
  - mean
  - standard deviation
  - maximum
  - minimum
  - median
  - quantile 0,50
  - quantile 0,75
  - quantile 0,85
  - quantile 0,95
- Peak statistic transformations, where the grouped measurements local maxima are extracted, and then the simple statistic transformations are computed:
  - peaks mean
  - peaks standard deviation
  - peaks maximum
  - peaks minimum
  - peaks median
  - peaks quantile 0,50
  - peaks quantile 0,75
  - peaks quantile 0,85
  - peaks quantile 0,95
  - peaks rate (ratio of local maxima)

The scheme present in figure 5.3 illustrates the statistical transformations method followed.

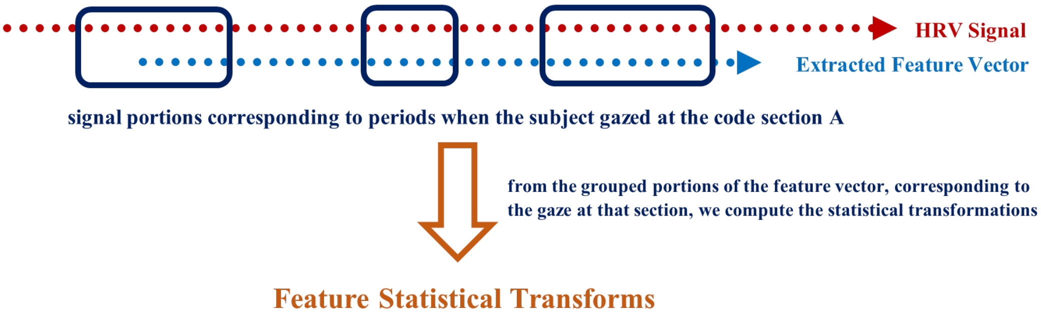

HRV Signal
Extracted Feature Vector
signal portions corresponding to periods when the subject gazed at the code section A
from the grouped portions of the feature vector, corresponding to the gaze at that section, we compute the statistical transformations
Feature Statistical Transforms

**Figure 5.3:** Feature statistical transformations scheme.

This process resulted in 589 features (31 features (section 5.1.2) x 19 statistical transformations) for each of the 18 different sliding window sizes used in the initial extraction (to simplify reference, let us call these the 'Transformed Features'). In resume, 18 datasets were built, one for each time frame in the current study, where each section gazed during a run is an instance, labelled according to their difficulty, and the dataset features are the 589 'Transformed Features'. The resultant datasets are the ones used in the classification approach.

### 5.1.3 Statistical analysis

In order to investigate the smallest time frame, i.e., the finest time resolution, where each ultra-short-term HRV feature is reliable, under our experiment context, the first approach conducted was the statistical analysis of these measurements. To this extent, an intra-subject and intra-run statistical analysis of the 'Extracted Feature Vectors' (features resultant from the initial extraction process, section 5.1.2) was performed using statistical significance tests, correlation tests and Bland-Altman plots. In the present study, the 'Extracted Feature Vectors' using the 180-second sliding window are used as the gold standard in the statistical analysis. Figure 5.4 represents the flow chart of the experimental steps followed to evaluate the ultra-short-term HRV measurements' reliability through the statistical analysis approach. When performing statistical analysis, the first important step is to choose the proper tests according to the data distribution. In order to determine if the features extracted follow a normal distribution, the Kolmogorov-Smirnov test was performed individually by measurement in each experiment run. The test's null hypothesis is that the data follows a standard normal distribution. At a 5% significance level, we obtained the rejection of the null hypothesis for every measurement in all the runs. The conclusion is that our data does not follow a standard normal distribution, so the statistical significance and correlation tests applied must be non-parametric.

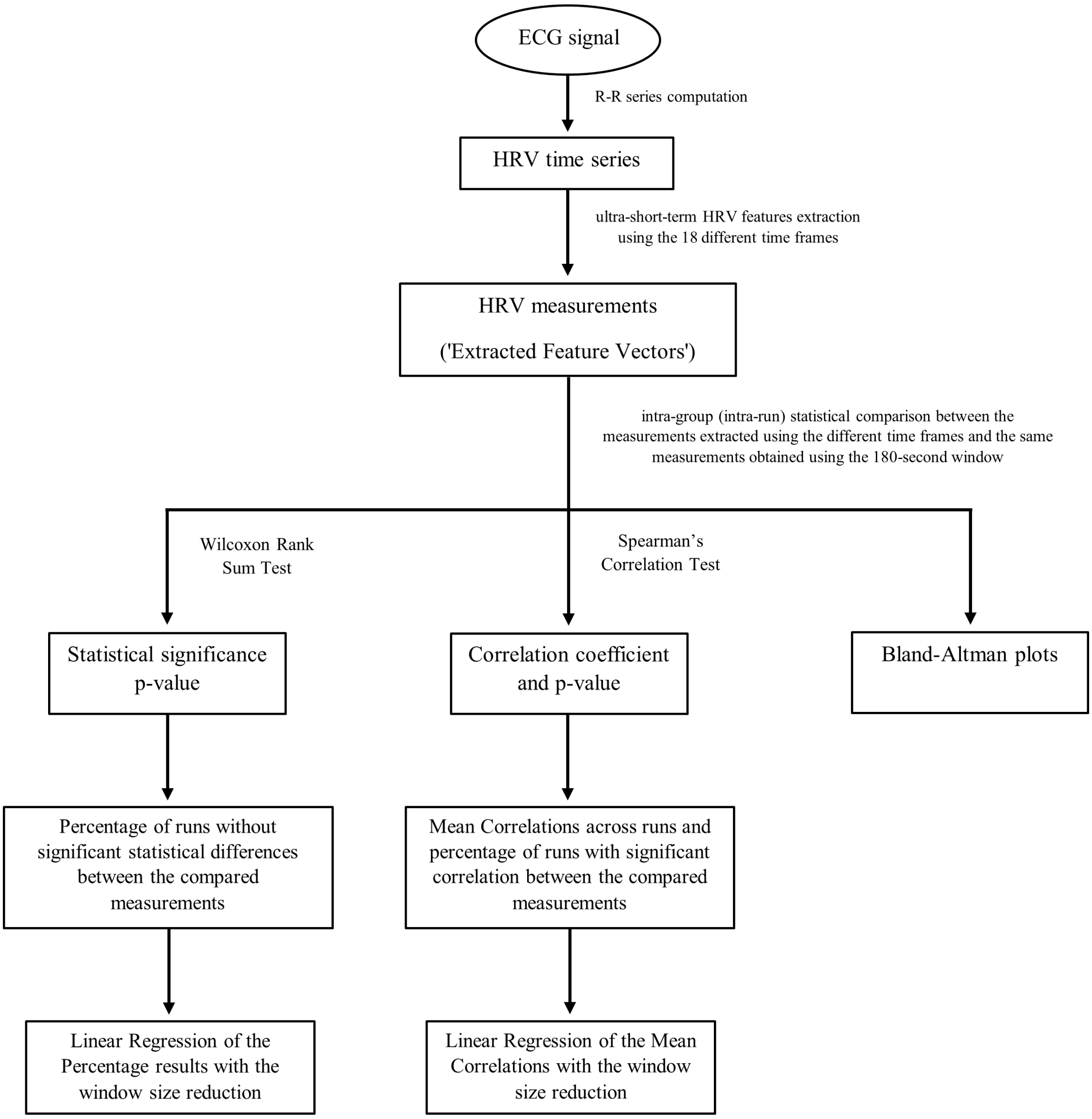

ECG signal
R-R series computation
HRV time series
ultra-short-term HRV features extraction using the 18 different time frames
HRV measurements ('Extracted Feature Vectors')
intra-group (intra-run) statistical comparison between the measurements extracted using the different time frames and the same measurements obtained using the 180-second window
Wilcoxon Rank Sum Test
Spearman's Correlation Test
Statistical significance p-value
Correlation coefficient and p-value
Bland-Altman plots
Percentage of runs without significant statistical differences between the compared measurements
Mean Correlations across runs and percentage of runs with significant correlation between the compared measurements
Linear Regression of the Percentage results with the window size reduction
Linear Regression of the Mean Correlations with the window size reduction

**Figure 5.4:** Flow chart of the experimental steps followed to evaluate the ultra-short-term HRV measurements' reliability through the statistical analysis approach.

#### 5.1.3.1 Statistical Significance test

The Wilcoxon rank sum test was performed to assess the sliding window size stability limit for each feature, i.e., to assess the smallest time frame where each feature is reliable. In this test, the measurements extracted using the different time frames were placed against the measurements computed using the 180 seconds sliding window. The test was performed independently for every experiment run and to all the 31 features in the study. With the explained procedure, we can inspect how the variation of the window size in the features extraction process affects the different measurements, assuming the 180 seconds window as the reference. The Wilcoxon rank sum null hypothesis is that the samples compared belong to distributions with equal medians.

Completing this procedure, a p-value was obtained for all the 31 features extracted using the 18 different sliding windows, totalizing a matrix of 31x18 p-values for each experiment run of the different subjects. If the p-value was below 0.05, the null hypothesis was rejected. Otherwise, the null hypothesis was accepted. To analyze the global extension of this test across the different runs, we computed the percentage of runs where each feature extracted with a specific time frame and the same feature obtained using the 180-second sliding window do not present significant statistical differences. These percentages were arranged in heatmap tables where the effect of reducing the sliding window size on the extraction can be observed.

In order to further investigate and quantify the time frame reduction effect, a graph was done for the results obtained from each feature. The 18 different time frames (window sizes used) were placed as the independent variable on the xx axis, and the percentage of runs without significant statistical differences as the dependent variable on the yy axis. It is essential to mention that the 180-second time frame is positioned at the origin of the xx axis, and each unit of this axis corresponds to a reduction of 10 seconds in the time frame used, being the minimum (10 seconds time frame) the last result on the xx axis. The yy axis ranges between 0 and 100%. Then a linear regression was performed for the results of each feature, and the respective coefficients of determination ($R^2$) were computed for each linear regression.

#### 5.1.3.2 Correlation test

In order to complement the insight obtained with the significance test and investigate the smallest time frame where each feature behavior is still representative of the 180-second correspondent measurement under our experiment context, Spearman's correlation test was also performed. Through the application of Spearman's correlation test, both a p-value and a correlation coefficient are obtained. The p-value is used to determine if a significant correlation exists between the data compared, while the correlation coefficient measures how correlated they are. This test compared the measurements obtained with the different sliding windows to those acquired with the 180 seconds reference time frame. Again, the procedure was done independently for each run.

An essential difference between Spearman's correlation test and the significance test of the previous section is that, in this correlation test, we must compare two vectors of the same length. As explained in section 5.1.2, extracting a measurement from an RR signal with a sliding window of 180 seconds produces a vector inferior in length compared to extracting the same measurement using a sliding window of an inferior time frame. Hence, with this test, we use the portion of the extracted feature vectors, computed using the smaller time frames, corresponding to the instants of the measurements in the vector resulting from the 180 seconds extraction (see Figure 5.5 illustration).

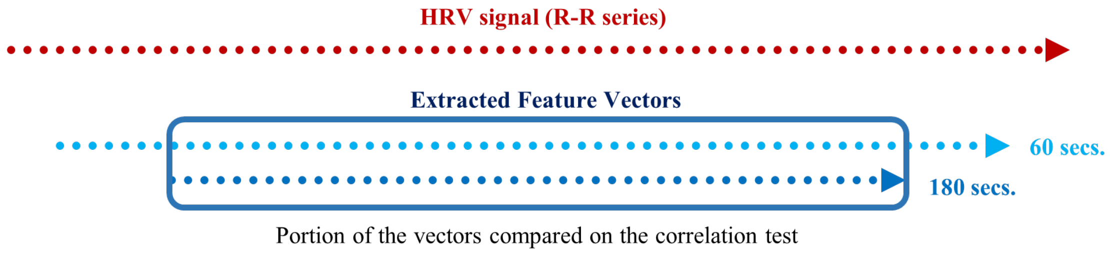

HRV signal (R-R series)
Extracted Feature Vectors
60 secs.
180 secs.
Portion of the vectors compared on the correlation test

**Figure 5.5:** Schematic of the portion of two feature vectors compared on the correlation test, extracted using respectively 180 and 60 seconds sliding windows, with 1 second steps.

After the correlation tests completion, for each experiment run of the different subjects, a p-value and a correlation coefficient are obtained for all the 31 features extracted using the 18 different sliding windows in the study (matrix 31x18 p-values and matrix 31x18 p-values correlation coefficients, for each experimental run). After this step, we computed the percentages of runs where significant correlation existed, and these percentages were arranged on a matrix. The matrix lines correspond to the features and the columns to the sliding window size used in their extraction.

Regarding the correlation coefficients obtained through this procedure, we calculated the means of these values across the different runs. Due to the presence of runs with different sample sizes, the means were computed using Fisher's z Transformation. This method allows us to give more weight to the feature extracted in runs with a larger number of samples [88]. With this step, the average correlation across runs was obtained between the 31 features extracted with the 180 seconds window and the same 31 features extracted with the other time frames in the study. The Fisher's mean values were placed in tables where the lines correspond to the features and the columns to the sliding window sizes used in their extraction. In these tables, we can efficiently observe how reducing the sliding window time span affects the correlation values.

Similarly to the process done in the previous subsection, a linear regression was performed with the mean correlation results obtained for each feature. This time, the correlation means are used as the dependent variable on the yy axis, which ranges from 0 to 1. The coefficients of determination ($R^2$) associated with the linear regressions were also computed.

#### 5.1.3.3 Bland-Altman plots

Following Pechia et al. [35] and Shaffer et al. [38] recommendations, we proceeded with the Bland-Altman plot analysis to evaluate the features' degree of bias. A relevant difference between our approach and the existing studies is that we perform an intra-subject and intra-run feature analysis, i.e., we test the features' correlation and statistical differences using different time frames within the same experiment run. In order to maintain the same intra-group analysis approach, we performed a Bland-Altman plot for non-parametric data for each feature extracted with the different time frames compared with the respective feature extracted with the 180-second window. This procedure was repeated for every experiment run. In the resultant Bland-Altman plots, it is possible to observe the level of agreement between the compared measurements.

### 5.1.4 Classification

In order to further investigate the impact of the time frame reduction in the HRV features extraction, we followed an alternative approach using these features for the complexity classification of software code sections. To this extent, we propose the performance analysis of several classifiers fed with the 'Transformed Features' (section 5.1.2.3), resulting from the statistical transformations of standard HRV features, computed at different time resolutions, i.e., using sliding windows ranging from 180 down to 10 seconds. These classifiers aim to discriminate between low and high complexity software code sections. We anticipate that by experimenting with diverse time frames, we will be able to investigate how the extracting window reduction affects the ultra-short-term HRV features and, ultimately, the classifiers' performance. In this study, the performance obtained by the classifiers using the features computed with the 180-second time frame was used as the reference. Figure 5.6 represents the flow chart of the experimental steps followed to investigate the ultra-short-term HRV impact on the classifiers' performance.

Regarding the current approach, it is essential to emphasize the use of 'Transformed Features'. The 'Transformed Features' are computed after the individual measurements present in the 'Extracted Feature Vectors' (produced in the feature extraction process, see section 5.1.2), being normalized (section 5.1.2.2) and concatenated based on all the instants the subject was gazing at a particular section during a run. This procedure allows us to capture and enhance the cognitive stress state presented by the subjects in the different code complexity sections. However, this makes our classifiers' approach conclusions specific to software code inspection tasks or similar contexts since we are evaluating specifically the transformed features discriminative ability between code sections of different complexity. In this way, this approach's conclusions are not as generalizable as the statistical analysis approach described in section 5.1.3, which does not involve statistical transformations and evaluates the features behavior and tendencies with the time frame reduction.

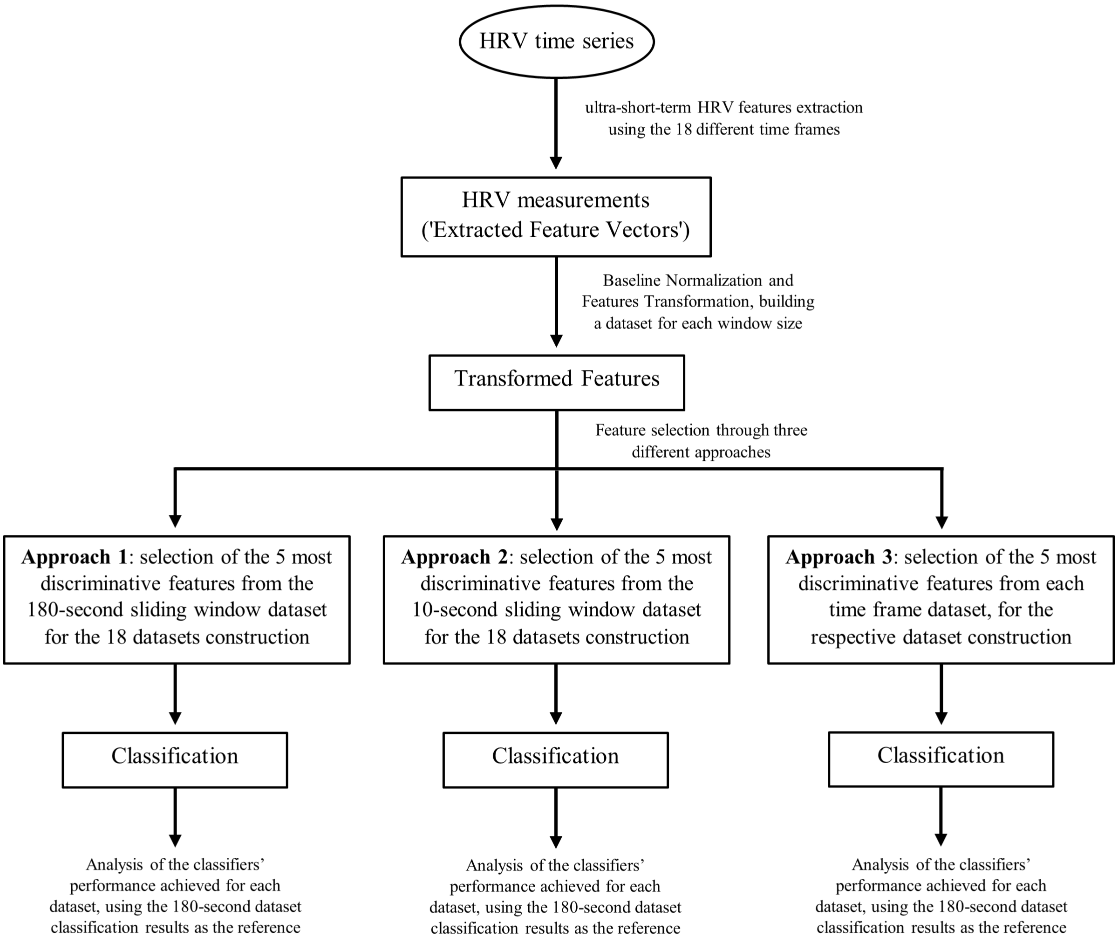

HRV time series
ultra-short-term HRV features extraction using the 18 different time frames
HRV measurements ('Extracted Feature Vectors')
Baseline Normalization and Features Transformation, building a dataset for each window size
Transformed Features
Feature selection through three different approaches
Approach 1: selection of the 5 most discriminative features from the 180-second sliding window dataset for the 18 datasets construction
Approach 2: selection of the 5 most discriminative features from the 10-second sliding window dataset for the 18 datasets construction
Approach 3: selection of the 5 most discriminative features from each time frame dataset, for the respective dataset construction
Classification
Classification
Classification
Analysis of the classifiers' performance achieved for each dataset, using the 180-second dataset classification results as the reference
Analysis of the classifiers' performance achieved for each dataset, using the 180-second dataset classification results as the reference
Analysis of the classifiers' performance achieved for each dataset, using the 180-second dataset classification results as the reference

**Figure 5.6:** Flow chart of the experimental steps followed to investigate the ultra-short-term HRV impact on the classifiers' performance.

### 5.1.4.1 Cross-validation scheme

The classification procedure was performed using a Nested Leave-One-Subject-Out cross-validation scheme and SVM classifiers with linear kernel. The SVM classifier with linear kernel was selected for performing the current analysis since it is a simple and interpretable algorithm that allows the inspection of the weight given to each feature. In this way, the linear SVM is ideal for avoiding overfitting and comparing the results based on the features obtained using different extracting windows, minimizing the influence of other factors.

The Nested Leave-One-Subject-Out essentially consists of two loops, the inner and the outer loop. The outer loop is used to evaluate the classification performance, using the samples corresponding to one subject for testing and the remaining subjects for training. The inner loop uses the training dataset, dividing it into an inner testing dataset with the samples corresponding to one subject and an inner training dataset with the remaining samples, for selecting the best model parameters. To this extent, the inner loop is used to perform a grid search, which allows the optimization of several classification parameters. Among these parameters are the classifier's regularization parameter (C), the number of features selected that optimizes the classification results and the selected features. The regularization parameter and the number of features used for classification are selected by testing different models and identifying the model which produces the most robust performance. The features are selected based on a statistical test evaluating the features' discriminative ability between the different complexity code sections (see 5.1.4.2 section). These parameters optimization and the classification procedure is performed independently for each time frame dataset.

#### 5.1.4.2 Feature selection

At this point of the study, 589 transformed features were contained in each time frame dataset (section 5.1.2.3). In order to reduce the number of features by dataset, and since the data did not follow a normal distribution, we performed the non-parametric Kruskal-Wallis test for each transformed feature split into two groups based on the complexity label (at each inner loop iteration of the nested cross-validation scheme used). The Kruskal-Wallis test algorithm returns the p-value for the null hypothesis that both groups come from the same distribution. The null hypothesis was rejected for a p-value below 0.05. A feature is designated as discriminative if the samples compared are considered samples from different distributions, i.e., if the null hypothesis is rejected and the feature's samples regarding the low and high complexity code sections have significant statistical differences.

From this procedure, it was observed that the datasets resulting from the smaller time frames had a substantially higher number of discriminative transformed features compared to the datasets corresponding to the more extensive sliding extracting windows. Through the grid search procedure, we concluded that five is the number of discriminative features that optimizes the classification results for most of the time frame datasets used in the classification process. In this way, we searched the five most discriminative transformed features by time frame, i.e., with the lowest p-values obtained in the Kruskal-Wallis analysis, at each inner loop of the nested cross-validation scheme used. Then, we computed the transformed features' occurrence percentages as the five most discriminative. The five most occurring transformed features in this search were selected as the top 5 most discriminative transformed features of the respective time frame. Concluded this step, in order to produce an appropriate comparison between the classification models, we followed three different approaches:

- **Approach 1)** Selection of the 5 most discriminative transformed features from the 180-second sliding window dataset, and used these 5 features in the construction of the 18 datasets;
- **Approach 2)** Selection of the 5 most discriminative transformed features from the 10-second sliding window dataset, and used these 5 features in the construction of the 18 datasets;
- **Approach 3)** Selection of the 5 most discriminative transformed features from each window size dataset, and use them in the respective time resolution dataset construction.

Each of the multiple datasets produced by the described approaches was used to train and test an SVM classifier with a linear kernel. This procedure allows us to evaluate the impact of the time resolution on the different classifiers' performances.

#### 5.1.4.3 Classifier Train and Test

The classification procedure was performed using a Nested Leave-One-Subject-Out cross-validation scheme (section 5.1.4.1) for every 18 datasets resultant from the three feature selection approaches (section 5.1.4.2). Each classifier's performance is evaluated in the outer loop of the nested cross-validation scheme.

In order to assess the performance of each classifier, the F-measure (or $F_1$-score) was the evaluation metric selected. This robust metric is defined as the harmonic average between the Precision (P) and the Recall (or sensitivity, R) metrics, computed by class [89]. In this way, the F-measure permits the False Positives (FP) and False

Negatives (FN) general evaluation for each class. The following expressions allow the F-measure calculation [89]:

- Precision (P) is the number of True Positives (TP) divided by the sum of the True Positives (TP) plus the False Positives (FP), i.e., the number of samples classified correctly of a given class divided by the number of samples classified as being part of that given class:

$$P = \frac{TP}{TP + FP} \tag{5.19}$$

- Recall (or sensitivity, R) corresponds to the number of True Positives (TP), i.e., the number of samples classified correctly of a given class, divided by the sum of the True Positives (TP) plus the False Negatives (FN), i.e., divided by the number of samples which are part of that given class:

$$R = \frac{TP}{TP + FN} \tag{5.20}$$

- Ultimately, the F-measure (or $F_1$-score) can be computed by the harmonic average of the last two metrics [89]:

$$F_1 = \frac{2PR}{P + R} \tag{5.21}$$

The confusion matrix present in table 5.2 further enlightens the True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) nomenclature meaning, to better understand the performance metrics.

**Table 5.2:** Confusion Matrix used in the classifiers' performance evaluation.

| | | Actual Values | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | True Positives (TP) | False Positives (FP) |
| | Negative (0) | False Negatives (FN) | True Negatives (TN) |

The F-measure was computed for both training and test in each outer loop of the Nested Leave-One-Subject-Out cross-validation method. This strategy produces a vector of $F_1$-scores for each window size dataset constructed following the three different feature selection approaches. The size of this vector will be equivalent to the number of subjects in the study. In order to assign an F-measure score for each time frame dataset, the average of the F-measures obtained for the respective time frame was computed. The standard deviation was also calculated to quantify the performance variability of the classifiers.

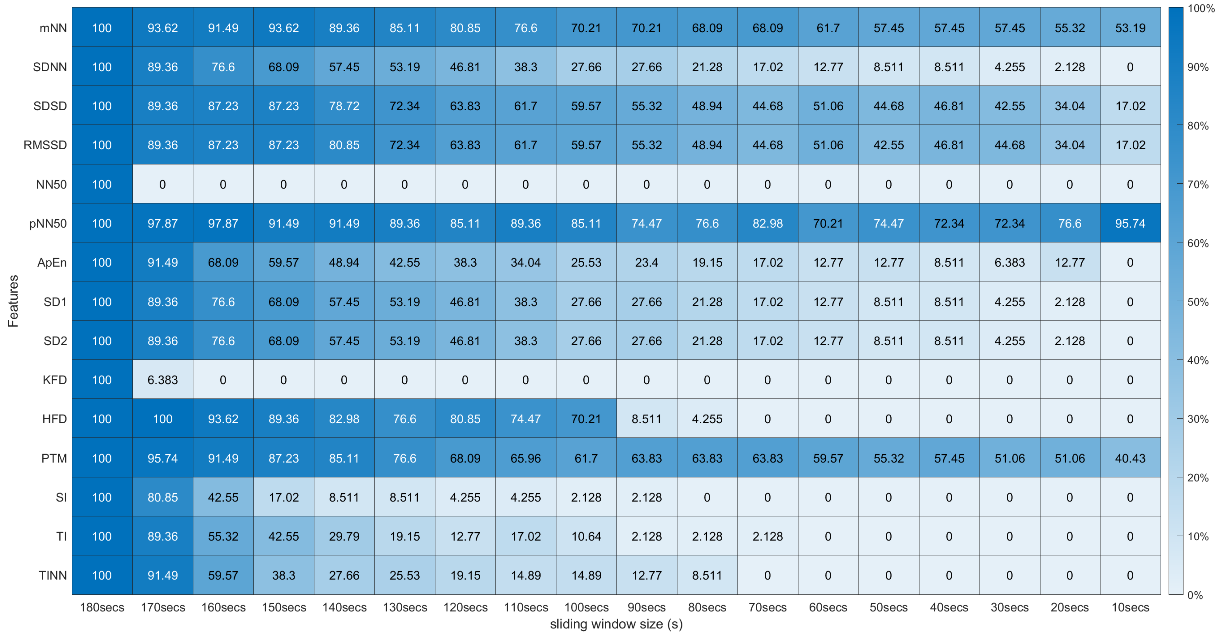
In order to compare the results obtained from the different time frames, a statistical significance analysis was conducted comparing the $F_1$-scores results distributions from each window size dataset against the 180-second dataset, assumed as reference. Through the Kolmogorov-Smirnov test application, it was possible to determine that the F1-Score results vector for each time frame did not follow a normal distribution. Consequently, the non-parametric Wilcoxon rank sum test was selected to verify if there were significant statistical differences between the performances obtained with the classifiers fed with the features from each time frame dataset and the classifiers fed with the features from the 180-second dataset. The $F_1$-scores performance vectors compared were considered from different distributions if the p-value obtained was below 0.05, i.e., if the Wilcoxon rank sum null hypothesis was rejected. This procedure was repeated for three different approaches followed.

## 5.2 Results

### 5.2.1 Reliability of ultra-short HRV measurements - statistical analysis approach

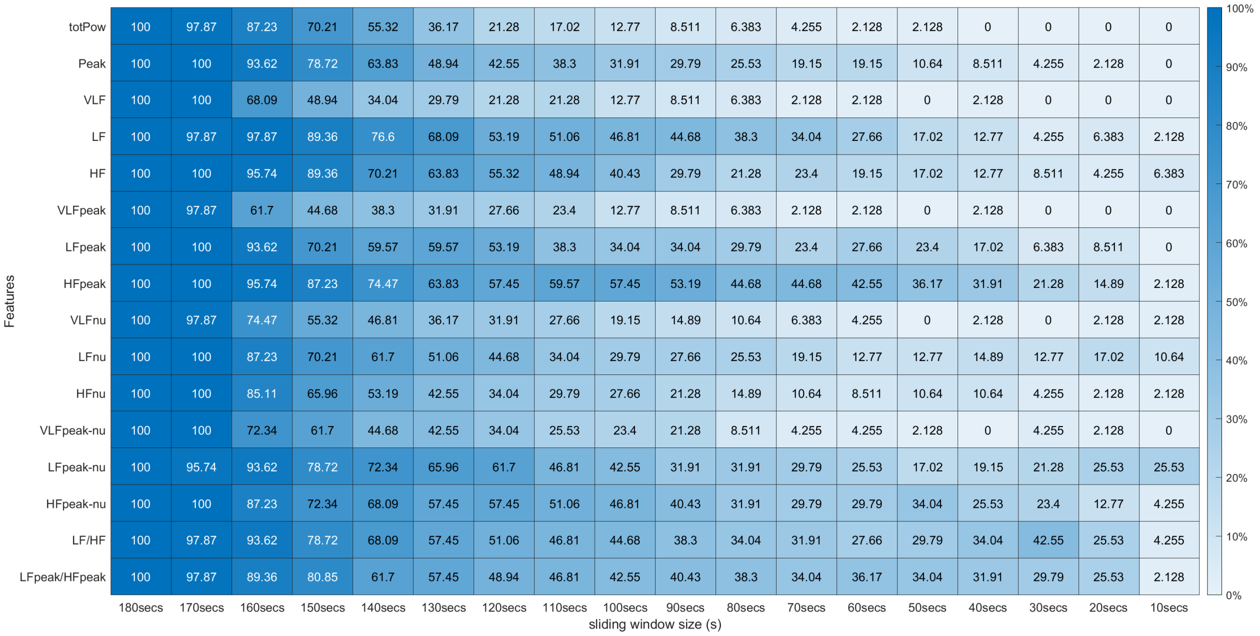
#### 5.2.1.1 Statistical Significance test results

Figures 5.7 and 5.8 summarize the results obtained using the procedure introduced in section 5.1.3.1. In particular, figure 5.7 summarizes the results related to the Time, Geometrical and Non-Linear Domain features, whereas figure 5.8 presents the results achieved using the Frequency Domain features. The values on each cell correspond to the percentage of runs where there is no significant difference between the feature (row) extracted using a respective window size (column) and the same feature obtained using the 180 seconds reference sliding window.



| Features | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mNN | 100 | 93.62 | 91.49 | 93.62 | 89.36 | 85.11 | 80.85 | 76.6 | 70.21 | 70.21 | 68.09 | 68.09 | 61.7 | 57.45 | 57.45 | 57.45 | 55.32 | 53.19 |
| SDNN | 100 | 89.36 | 76.6 | 68.09 | 57.45 | 53.19 | 46.81 | 38.3 | 27.66 | 27.66 | 21.28 | 17.02 | 12.77 | 8.511 | 8.511 | 4.255 | 2.128 | 0 |
| SDSD | 100 | 89.36 | 87.23 | 87.23 | 78.72 | 72.34 | 63.83 | 61.7 | 59.57 | 55.32 | 48.94 | 44.68 | 51.06 | 44.68 | 46.81 | 42.55 | 34.04 | 17.02 |
| RMSSD | 100 | 89.36 | 87.23 | 87.23 | 80.85 | 72.34 | 63.83 | 61.7 | 59.57 | 55.32 | 48.94 | 44.68 | 51.06 | 42.55 | 46.81 | 44.68 | 34.04 | 17.02 |
| NN50 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pNN50 | 100 | 97.87 | 97.87 | 91.49 | 91.49 | 89.36 | 85.11 | 89.36 | 85.11 | 74.47 | 76.6 | 82.98 | 70.21 | 74.47 | 72.34 | 72.34 | 76.6 | 95.74 |
| ApEn | 100 | 91.49 | 68.09 | 59.57 | 48.94 | 42.55 | 38.3 | 34.04 | 25.53 | 23.4 | 19.15 | 17.02 | 12.77 | 12.77 | 8.511 | 6.383 | 12.77 | 0 |
| SD1 | 100 | 89.36 | 76.6 | 68.09 | 57.45 | 53.19 | 46.81 | 38.3 | 27.66 | 27.66 | 21.28 | 17.02 | 12.77 | 8.511 | 8.511 | 4.255 | 2.128 | 0 |
| SD2 | 100 | 89.36 | 76.6 | 68.09 | 57.45 | 53.19 | 46.81 | 38.3 | 27.66 | 27.66 | 21.28 | 17.02 | 12.77 | 8.511 | 8.511 | 4.255 | 2.128 | 0 |
| KFD | 100 | 6.383 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HFD | 100 | 100 | 93.62 | 89.36 | 82.98 | 76.6 | 80.85 | 74.47 | 70.21 | 8.511 | 4.255 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PTM | 100 | 95.74 | 91.49 | 87.23 | 85.11 | 76.6 | 68.09 | 65.96 | 61.7 | 63.83 | 63.83 | 63.83 | 59.57 | 55.32 | 57.45 | 51.06 | 51.06 | 40.43 |
| SI | 100 | 80.85 | 42.55 | 17.02 | 8.511 | 8.511 | 4.255 | 4.255 | 2.128 | 2.128 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| TI | 100 | 89.36 | 55.32 | 42.55 | 29.79 | 19.15 | 12.77 | 17.02 | 10.64 | 2.128 | 2.128 | 2.128 | 0 | 0 | 0 | 0 | 0 | 0 |
| TINN | 100 | 91.49 | 59.57 | 38.3 | 27.66 | 25.53 | 19.15 | 14.89 | 14.89 | 12.77 | 8.511 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Figure 5.7:** Wilcoxon Rank Sum Test (Time, Non-Linear and Geometrical Domain)[1]



| | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| totPow | 100 | 97.87 | 87.23 | 70.21 | 55.32 | 36.17 | 21.28 | 17.02 | 12.77 | 8.511 | 6.383 | 4.255 | 2.128 | 2.128 | 0 | 0 | 0 | 0 |
| Peak | 100 | 100 | 93.62 | 78.72 | 63.83 | 48.94 | 42.55 | 38.3 | 31.91 | 29.79 | 25.53 | 19.15 | 19.15 | 10.64 | 8.511 | 4.255 | 2.128 | 0 |
| VLF | 100 | 100 | 68.09 | 48.94 | 34.04 | 29.79 | 21.28 | 21.28 | 12.77 | 8.511 | 6.383 | 2.128 | 2.128 | 0 | 2.128 | 0 | 0 | 0 |
| LF | 100 | 97.87 | 97.87 | 89.36 | 76.6 | 68.09 | 53.19 | 51.06 | 46.81 | 44.68 | 38.3 | 34.04 | 27.66 | 17.02 | 12.77 | 4.255 | 6.383 | 2.128 |
| HF | 100 | 100 | 95.74 | 89.36 | 70.21 | 63.83 | 55.32 | 48.94 | 40.43 | 29.79 | 21.28 | 23.4 | 19.15 | 17.02 | 12.77 | 8.511 | 4.255 | 6.383 |
| VLFpeak | 100 | 97.87 | 61.7 | 44.68 | 38.3 | 31.91 | 27.66 | 23.4 | 12.77 | 8.511 | 6.383 | 2.128 | 2.128 | 0 | 2.128 | 0 | 0 | 0 |
| LFpeak | 100 | 100 | 93.62 | 70.21 | 59.57 | 59.57 | 53.19 | 38.3 | 34.04 | 34.04 | 29.79 | 23.4 | 27.66 | 23.4 | 17.02 | 6.383 | 8.511 | 0 |
| HFpeak | 100 | 100 | 95.74 | 87.23 | 74.47 | 63.83 | 57.45 | 59.57 | 57.45 | 53.19 | 44.68 | 44.68 | 42.55 | 36.17 | 31.91 | 21.28 | 14.89 | 2.128 |
| VLFnu | 100 | 97.87 | 74.47 | 55.32 | 46.81 | 36.17 | 31.91 | 27.66 | 19.15 | 14.89 | 10.64 | 6.383 | 4.255 | 0 | 2.128 | 0 | 2.128 | 2.128 |
| LFnu | 100 | 100 | 87.23 | 70.21 | 61.7 | 51.06 | 44.68 | 34.04 | 29.79 | 27.66 | 25.53 | 19.15 | 12.77 | 12.77 | 14.89 | 12.77 | 17.02 | 10.64 |
| HFnu | 100 | 100 | 85.11 | 65.96 | 53.19 | 42.55 | 34.04 | 29.79 | 27.66 | 21.28 | 14.89 | 10.64 | 8.511 | 10.64 | 10.64 | 4.255 | 2.128 | 2.128 |
| VLFpeak-nu | 100 | 100 | 72.34 | 61.7 | 44.68 | 42.55 | 34.04 | 25.53 | 23.4 | 21.28 | 8.511 | 4.255 | 4.255 | 2.128 | 0 | 4.255 | 2.128 | 0 |
| LFpeak-nu | 100 | 95.74 | 93.62 | 78.72 | 72.34 | 65.96 | 61.7 | 46.81 | 42.55 | 31.91 | 31.91 | 29.79 | 25.53 | 17.02 | 19.15 | 21.28 | 25.53 | 25.53 |
| HFpeak-nu | 100 | 100 | 87.23 | 72.34 | 68.09 | 57.45 | 57.45 | 51.06 | 46.81 | 40.43 | 31.91 | 29.79 | 29.79 | 34.04 | 25.53 | 23.4 | 12.77 | 4.255 |
| LF/HF | 100 | 97.87 | 93.62 | 78.72 | 68.09 | 57.45 | 51.06 | 46.81 | 44.68 | 38.3 | 34.04 | 31.91 | 27.66 | 29.79 | 34.04 | 42.55 | 25.53 | 4.255 |
| LFpeak/HFpeak | 100 | 97.87 | 89.36 | 80.85 | 61.7 | 57.45 | 48.94 | 46.81 | 42.55 | 40.43 | 38.3 | 34.04 | 36.17 | 34.04 | 31.91 | 29.79 | 25.53 | 2.128 |

**Figure 5.8:** Wilcoxon Rank Sum Test (Frequency Domain)[1]

Figures 5.9 and 5.10 are the graphic presentations of the linear regressions obtained respectively for the Time, Geometrical and Non-Linear Domain features results and for the Frequency Domain features results. The features' lines chosen to be presented were the ones considered representative of the overall results. In the appendix section A, it is possible to consult the slopes, the yy interceptions and the coefficients of determination obtained for every feature in the study.

[1] Percentage of runs where the feature (line) extracted with a respective window size (column) did not present significant statistical differences compared to the same feature extracted using the 180-second window.

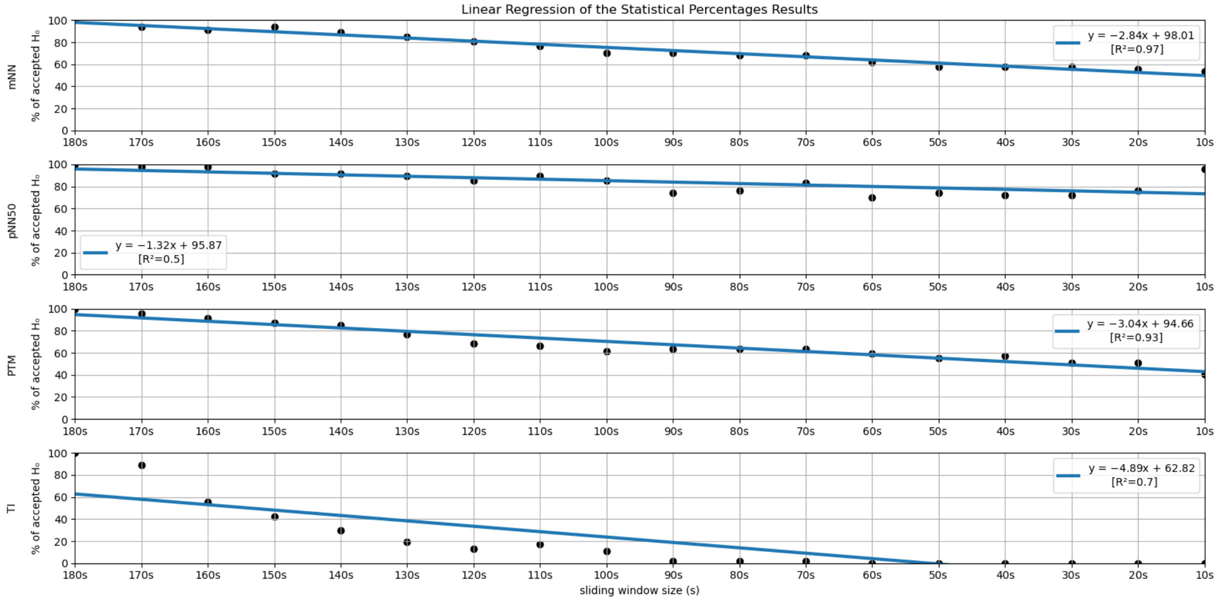


**Figure 5.9:** Linear Regressions of the Statistical Percentages obtained for the features mNN, pNN50, PTM and TI.

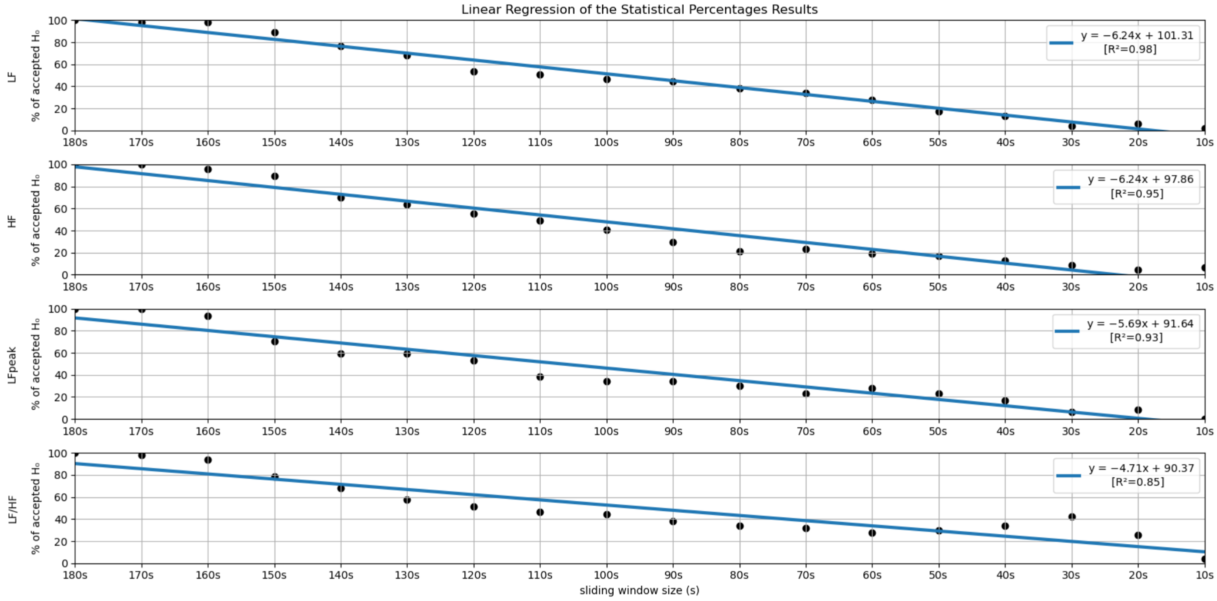
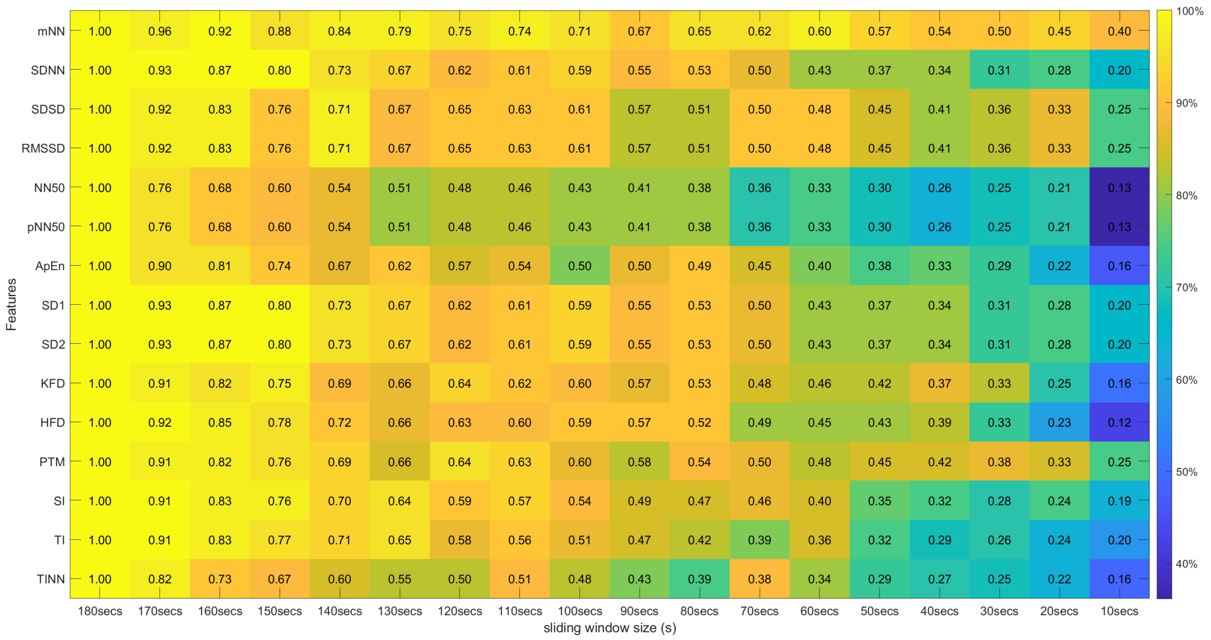


**Figure 5.10:** Linear Regressions of the Statistical Percentages obtained for the features LF, HF, LFpeak and LF/HF.

From both heatmaps in figures 5.7 and 5.8, it is possible to observe that reducing the sliding window size has a great impact on the significance test results in almost every feature. This drop represents a large decrease, through the time frame reduction, in the percentage of runs where there is no significant difference between extracting features using that window and the 180 seconds reference window. From figures 5.9 and 5.10, we can observe that the linear regressions obtained provide quantitative and visual support for this claim.

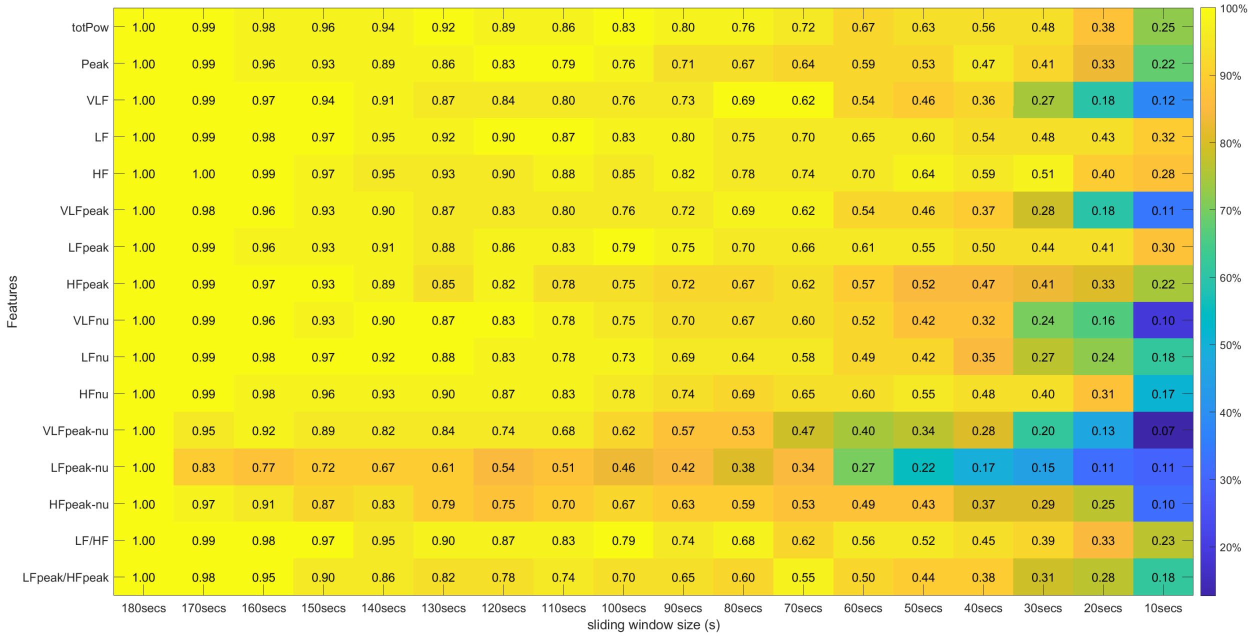
#### 5.2.1.2 Correlation test results

Figures 5.11 and 5.12 introduce the correlation testing results described in section 5.1.3.2, i.e., figure 5.11 corresponds to the correlation analysis of the Time, Geometrical and Non-Linear Domain features, whereas figure 5.12 corresponds to the correlation analysis of the Frequency Domain features. Each **cell value** corresponds to the Fisher's means, across the different experiment runs, of the correlations between the feature (row) extracted using a window size (column) and the same feature obtained using the 180 seconds sliding window. While the **heatmap colours** correspond to the percentage of runs where a significant correlation ($\alpha = 0.05$) exists between the feature (line) extracted using a given window size (column) and the same feature obtained using the 180 seconds sliding window.



| | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mNN | 1.00 | 0.96 | 0.92 | 0.88 | 0.84 | 0.79 | 0.75 | 0.74 | 0.71 | 0.67 | 0.65 | 0.62 | 0.60 | 0.57 | 0.54 | 0.50 | 0.45 | 0.40 |
| SDNN | 1.00 | 0.93 | 0.87 | 0.80 | 0.73 | 0.67 | 0.62 | 0.61 | 0.59 | 0.55 | 0.53 | 0.50 | 0.43 | 0.37 | 0.34 | 0.31 | 0.28 | 0.20 |
| SDSD | 1.00 | 0.92 | 0.83 | 0.76 | 0.71 | 0.67 | 0.65 | 0.63 | 0.61 | 0.57 | 0.51 | 0.50 | 0.48 | 0.45 | 0.41 | 0.36 | 0.33 | 0.25 |
| RMSSD | 1.00 | 0.92 | 0.83 | 0.76 | 0.71 | 0.67 | 0.65 | 0.63 | 0.61 | 0.57 | 0.51 | 0.50 | 0.48 | 0.45 | 0.41 | 0.36 | 0.33 | 0.25 |
| NN50 | 1.00 | 0.76 | 0.68 | 0.60 | 0.54 | 0.51 | 0.48 | 0.46 | 0.43 | 0.41 | 0.38 | 0.36 | 0.33 | 0.30 | 0.26 | 0.25 | 0.21 | 0.13 |
| pNN50 | 1.00 | 0.76 | 0.68 | 0.60 | 0.54 | 0.51 | 0.48 | 0.46 | 0.43 | 0.41 | 0.38 | 0.36 | 0.33 | 0.30 | 0.26 | 0.25 | 0.21 | 0.13 |
| ApEn | 1.00 | 0.90 | 0.81 | 0.74 | 0.67 | 0.62 | 0.57 | 0.54 | 0.50 | 0.50 | 0.49 | 0.45 | 0.40 | 0.38 | 0.33 | 0.29 | 0.22 | 0.16 |
| SD1 | 1.00 | 0.93 | 0.87 | 0.80 | 0.73 | 0.67 | 0.62 | 0.61 | 0.59 | 0.55 | 0.53 | 0.50 | 0.43 | 0.37 | 0.34 | 0.31 | 0.28 | 0.20 |
| SD2 | 1.00 | 0.93 | 0.87 | 0.80 | 0.73 | 0.67 | 0.62 | 0.61 | 0.59 | 0.55 | 0.53 | 0.50 | 0.43 | 0.37 | 0.34 | 0.31 | 0.28 | 0.20 |
| KFD | 1.00 | 0.91 | 0.82 | 0.75 | 0.69 | 0.66 | 0.64 | 0.62 | 0.60 | 0.57 | 0.53 | 0.48 | 0.46 | 0.42 | 0.37 | 0.33 | 0.25 | 0.16 |
| HFD | 1.00 | 0.92 | 0.85 | 0.78 | 0.72 | 0.66 | 0.63 | 0.60 | 0.59 | 0.57 | 0.52 | 0.49 | 0.45 | 0.43 | 0.39 | 0.33 | 0.23 | 0.12 |
| PTM | 1.00 | 0.91 | 0.82 | 0.76 | 0.69 | 0.66 | 0.64 | 0.63 | 0.60 | 0.58 | 0.54 | 0.50 | 0.48 | 0.45 | 0.42 | 0.38 | 0.33 | 0.25 |
| SI | 1.00 | 0.91 | 0.83 | 0.76 | 0.70 | 0.64 | 0.59 | 0.57 | 0.54 | 0.49 | 0.47 | 0.46 | 0.40 | 0.35 | 0.32 | 0.28 | 0.24 | 0.19 |
| TI | 1.00 | 0.91 | 0.83 | 0.77 | 0.71 | 0.65 | 0.58 | 0.56 | 0.51 | 0.47 | 0.42 | 0.39 | 0.36 | 0.32 | 0.29 | 0.26 | 0.24 | 0.20 |
| TINN | 1.00 | 0.82 | 0.73 | 0.67 | 0.60 | 0.55 | 0.50 | 0.51 | 0.48 | 0.43 | 0.39 | 0.38 | 0.34 | 0.29 | 0.27 | 0.25 | 0.22 | 0.16 |

**Figure 5.11:** Spearman's Correlation Test (Time, Non-Linear and Geometrical Domains)[2]



| Features | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| totPow | 1.00 | 0.99 | 0.98 | 0.96 | 0.94 | 0.92 | 0.89 | 0.86 | 0.83 | 0.80 | 0.76 | 0.72 | 0.67 | 0.63 | 0.56 | 0.48 | 0.38 | 0.25 |
| Peak | 1.00 | 0.99 | 0.96 | 0.93 | 0.89 | 0.86 | 0.83 | 0.79 | 0.76 | 0.71 | 0.67 | 0.64 | 0.59 | 0.53 | 0.47 | 0.41 | 0.33 | 0.22 |
| VLF | 1.00 | 0.99 | 0.97 | 0.94 | 0.91 | 0.87 | 0.84 | 0.80 | 0.76 | 0.73 | 0.69 | 0.62 | 0.54 | 0.46 | 0.36 | 0.27 | 0.18 | 0.12 |
| LF | 1.00 | 0.99 | 0.98 | 0.97 | 0.95 | 0.92 | 0.90 | 0.87 | 0.83 | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.54 | 0.48 | 0.43 | 0.32 |
| HF | 1.00 | 1.00 | 0.99 | 0.97 | 0.95 | 0.93 | 0.90 | 0.88 | 0.85 | 0.82 | 0.78 | 0.74 | 0.70 | 0.64 | 0.59 | 0.51 | 0.40 | 0.28 |
| VLFpeak | 1.00 | 0.98 | 0.96 | 0.93 | 0.90 | 0.87 | 0.83 | 0.80 | 0.76 | 0.72 | 0.69 | 0.62 | 0.54 | 0.46 | 0.37 | 0.28 | 0.18 | 0.11 |
| LFpeak | 1.00 | 0.99 | 0.96 | 0.93 | 0.91 | 0.88 | 0.86 | 0.83 | 0.79 | 0.75 | 0.70 | 0.66 | 0.61 | 0.55 | 0.50 | 0.44 | 0.41 | 0.30 |
| HFpeak | 1.00 | 0.99 | 0.97 | 0.93 | 0.89 | 0.85 | 0.82 | 0.78 | 0.75 | 0.72 | 0.67 | 0.62 | 0.57 | 0.52 | 0.47 | 0.41 | 0.33 | 0.22 |
| VLFnu | 1.00 | 0.99 | 0.96 | 0.93 | 0.90 | 0.87 | 0.83 | 0.78 | 0.75 | 0.70 | 0.67 | 0.60 | 0.52 | 0.42 | 0.32 | 0.24 | 0.16 | 0.10 |
| LFnu | 1.00 | 0.99 | 0.98 | 0.97 | 0.92 | 0.88 | 0.83 | 0.78 | 0.73 | 0.69 | 0.64 | 0.58 | 0.49 | 0.42 | 0.35 | 0.27 | 0.24 | 0.18 |
| HFnu | 1.00 | 0.99 | 0.98 | 0.96 | 0.93 | 0.90 | 0.87 | 0.83 | 0.78 | 0.74 | 0.69 | 0.65 | 0.60 | 0.55 | 0.48 | 0.40 | 0.31 | 0.17 |
| VLFpeak-nu | 1.00 | 0.95 | 0.92 | 0.89 | 0.82 | 0.84 | 0.74 | 0.68 | 0.62 | 0.57 | 0.53 | 0.47 | 0.40 | 0.34 | 0.28 | 0.20 | 0.13 | 0.07 |
| LFpeak-nu | 1.00 | 0.83 | 0.77 | 0.72 | 0.67 | 0.61 | 0.54 | 0.51 | 0.46 | 0.42 | 0.38 | 0.34 | 0.27 | 0.22 | 0.17 | 0.15 | 0.11 | 0.11 |
| HFpeak-nu | 1.00 | 0.97 | 0.91 | 0.87 | 0.83 | 0.79 | 0.75 | 0.70 | 0.67 | 0.63 | 0.59 | 0.53 | 0.49 | 0.43 | 0.37 | 0.29 | 0.25 | 0.10 |
| LF/HF | 1.00 | 0.99 | 0.98 | 0.97 | 0.95 | 0.90 | 0.87 | 0.83 | 0.79 | 0.74 | 0.68 | 0.62 | 0.56 | 0.52 | 0.45 | 0.39 | 0.33 | 0.23 |
| LFpeak/HFpeak | 1.00 | 0.98 | 0.95 | 0.90 | 0.86 | 0.82 | 0.78 | 0.74 | 0.70 | 0.65 | 0.60 | 0.55 | 0.50 | 0.44 | 0.38 | 0.31 | 0.28 | 0.18 |

**Figure 5.12:** Spearman's Correlation Test (Frequency Domain)[2]

Figures 5.13 and 5.14 allow a visual inspection of the linear regressions obtained respectively for the Time, Geometrical and Non-Linear Domain features correlation means and for the Frequency Domain features correlation means. Following the same scheme as sub-section 5.2.1.1 (statistical significance), we selected a few representative linear regressions examples to be graphically presented. The appendix section B tables contain the values obtained for the slopes, the yy interceptions and the coefficients of determination.

[2] **Heatmap Colours**: Percentage of runs where there exists significant correlation between the feature (row) extracted using the respective window size (column) and the same feature obtained using the 180 second sliding window.

**Cell Values**: Means, across the different runs, of the correlation coefficients between the feature (row) extracted using the respective window size (column) and the same feature obtained using the 180 second sliding window.

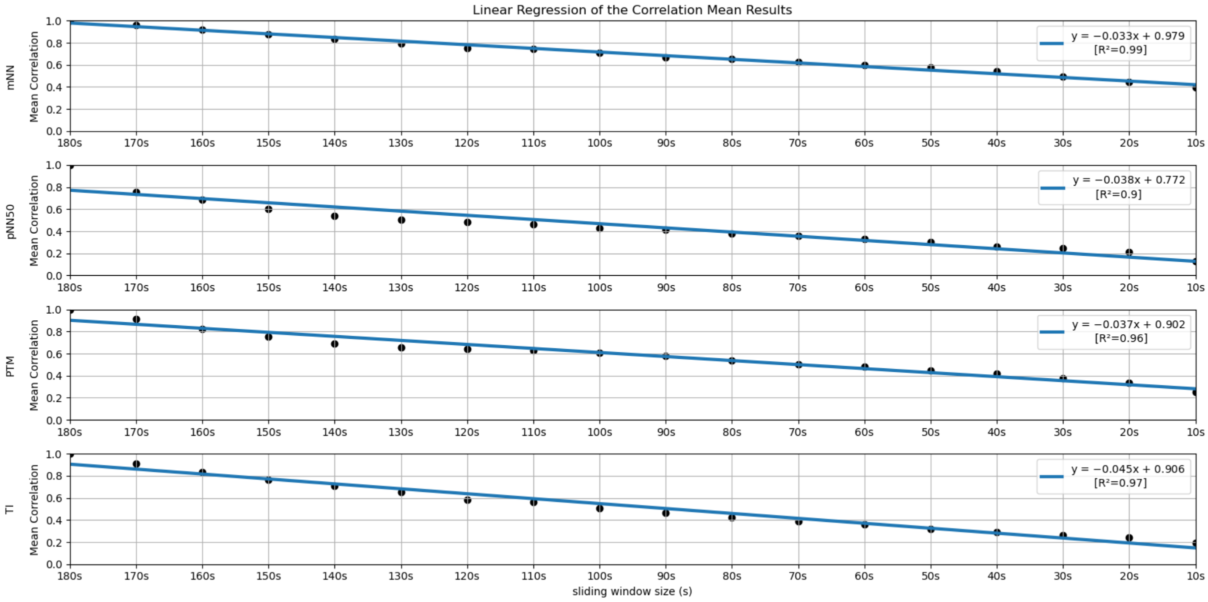


**Figure 5.13:** Linear Regressions of the Mean Correlations across runs obtained for the features mNN, pNN50, PTM and TI.

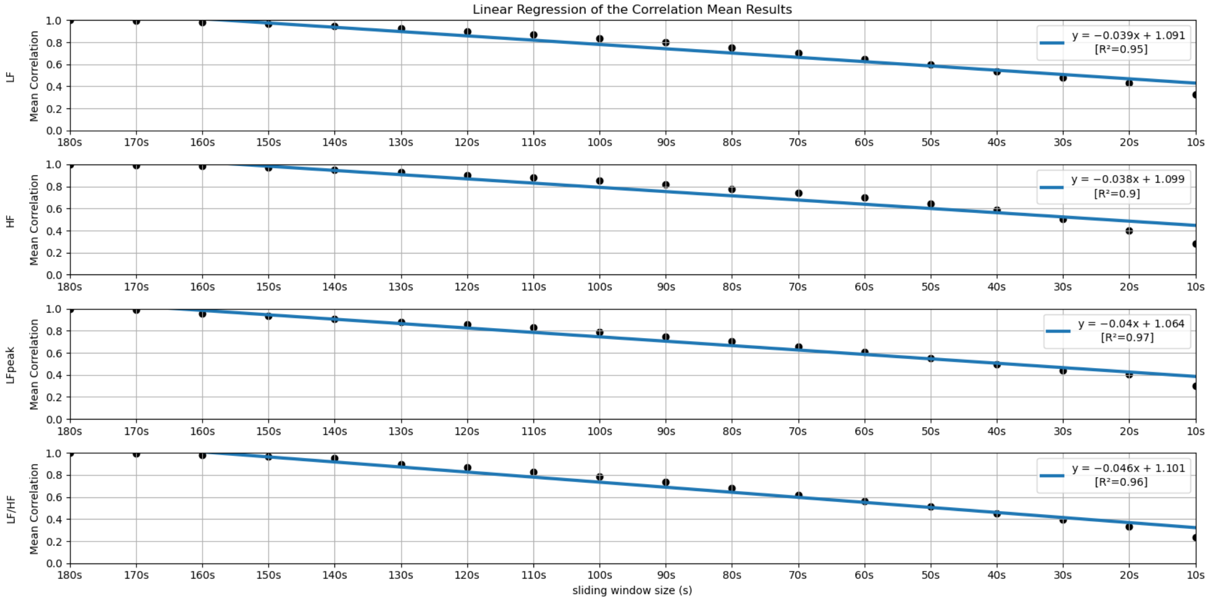


**Figure 5.14:** Linear Regressions of the Mean Correlations across runs obtained for the features LF, HF, LFpeak and LF/HF.

#### 5.2.1.3 Bland-Altman plots results

For illustrative purposes, figure 5.15 depicts the Bland-Altman plots achieved for the feature LF/HF, extracted with the 120, 90, 60, 30 and 10 seconds time frames compared to the same feature extracted using a 180 seconds window. The data used to perform these plots corresponds to a single experiment run of an individual subject. The Bland-Altman plots allow us to observe the degree of bias present between the compared features and if the data dispersion remains within the 95% line of agreement.

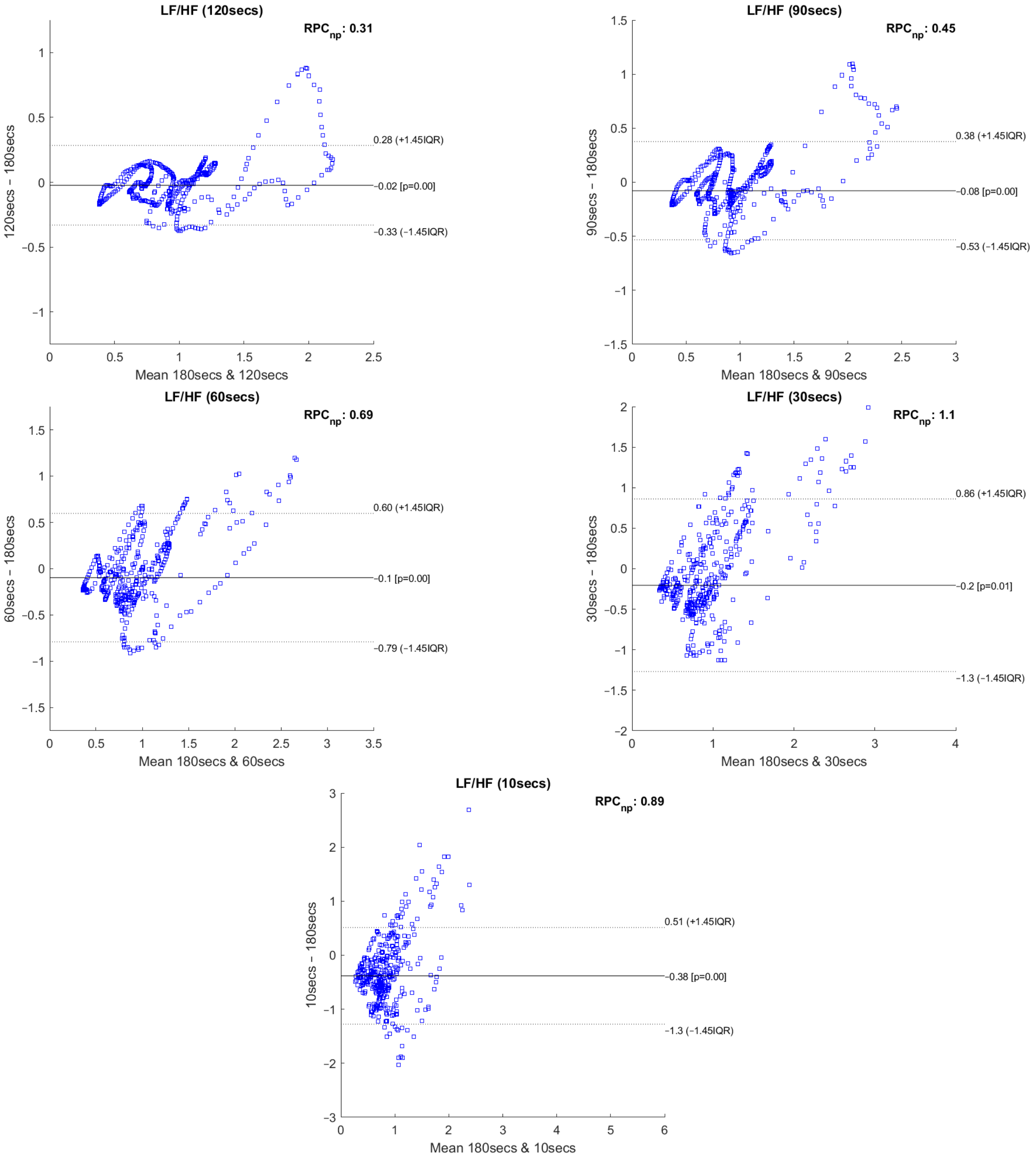


**Figure 5.15:** Bland-Altman plots of the LF/HF feature extracted with 120, 90, 60, 30 and 10 seconds compared to the LF/HF extracted with 180 seconds time frame, regarding a single experiment run of an individual subject.

### 5.2.2 Impact of Ultra-Short-Term HRV Features in Software Code Sections Complexity Classification

Tables 5.3 and 5.4 contain respectively the top 5 most discriminative transformed HRV features from the 180 seconds and from the 10 seconds time frames datasets. The occurrence is the percentage of times the feature is present in the five most discriminative features set at the inner loop of the nested cross-validation scheme. The top 5 mean position corresponds to the average position the feature occupied when present in the five most discriminative features set, which varies between 1 and 5. The five most discriminative transformed HRV features from the other time frames in the study can be consulted in the appendix C.

**Table 5.3:** 5 most discriminative transformed HRV features from the 180 seconds dataset (selected using the Kruskal Wallis test at each Nested Leave-One-Subject-Out inner iteration).

**180 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 100.0 | 1.0 |
| mNN | peaks rate | 82.42 | 3.2 |
| HFD | peaks rate | 78.02 | 3.2 |
| LF/HF | peaks rate | 76.92 | 2.9 |
| TI | peaks rate | 56.04 | 4.0 |

**Table 5.4:** 5 most discriminative transformed HRV features from the 10 seconds dataset (selected using the Kruskal Wallis test at each Nested Leave-One-Subject-Out inner iteration).

**10 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HFpeak | peaks min | 94.29 | 2.0 |
| HFpeak-nu | min | 75.24 | 2.3 |
| KFD | peaks rate | 61.90 | 3.0 |
| HFD | min | 60.00 | 3.1 |
| ApEn | peaks min | 54.29 | 2.9 |

The results of the classification procedure described in section 5.1.4.3 are presented in figure 5.16 and table 5.5. Figure 5.16 corresponds to the graphical representation of the median and standard deviation of the F-measure values obtained for each window size dataset following the three feature selection approaches (see table in appendix D to consult the plotted values). In table 5.5, we present the results of the Wilcoxon rank sum test analysis of the F-measures distributions, using the results of the 180 seconds dataset of the respective approach as reference.

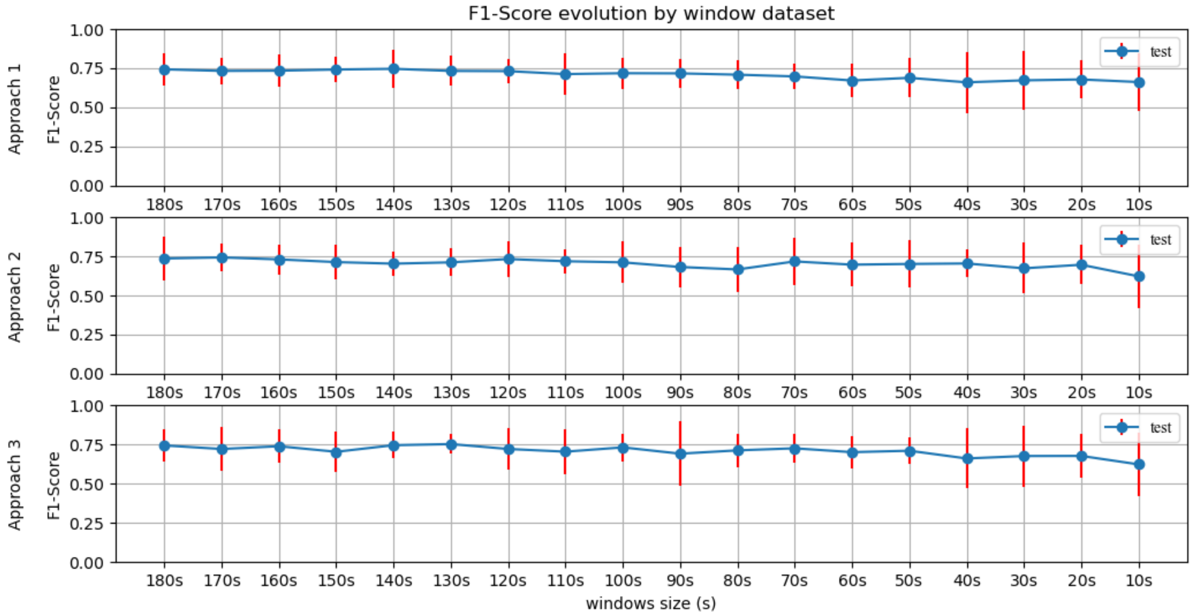
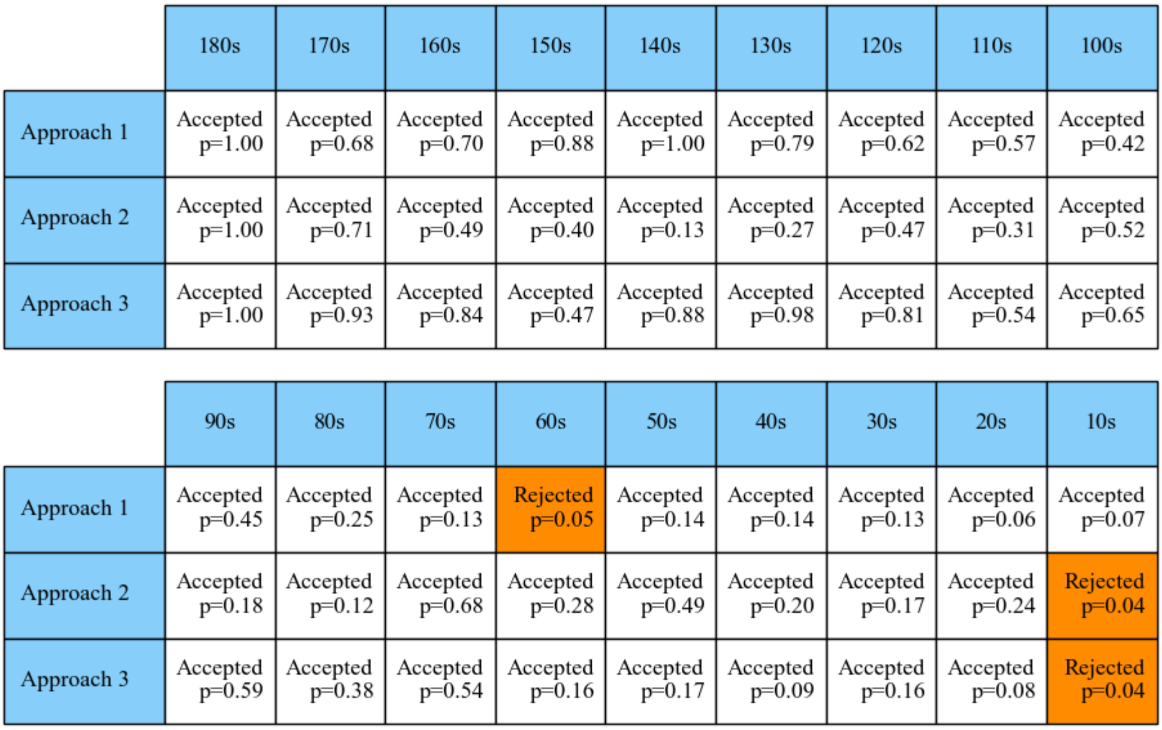


**Figure 5.16:** Mean F1-Scores and Standard Deviations plot obtained using the datasets based on the different time frames and approaches.

**Table 5.5:** Wilcoxon rank sum test analysis of the F1-Scores distributions compared to the 180 seconds dataset using the results of the 180 seconds dataset of the respective approach as reference.



| | 180s | 170s | 160s | 150s | 140s | 130s | 120s | 110s | 100s |
|---|---|---|---|---|---|---|---|---|---|
| Approach 1 | Accepted<br>p=1.00 | Accepted<br>p=0.68 | Accepted<br>p=0.70 | Accepted<br>p=0.88 | Accepted<br>p=1.00 | Accepted<br>p=0.79 | Accepted<br>p=0.62 | Accepted<br>p=0.57 | Accepted<br>p=0.42 |
| Approach 2 | Accepted<br>p=1.00 | Accepted<br>p=0.71 | Accepted<br>p=0.49 | Accepted<br>p=0.40 | Accepted<br>p=0.13 | Accepted<br>p=0.27 | Accepted<br>p=0.47 | Accepted<br>p=0.31 | Accepted<br>p=0.52 |
| Approach 3 | Accepted<br>p=1.00 | Accepted<br>p=0.93 | Accepted<br>p=0.84 | Accepted<br>p=0.47 | Accepted<br>p=0.88 | Accepted<br>p=0.98 | Accepted<br>p=0.81 | Accepted<br>p=0.54 | Accepted<br>p=0.65 |

| | 90s | 80s | 70s | 60s | 50s | 40s | 30s | 20s | 10s |
|---|---|---|---|---|---|---|---|---|---|
| Approach 1 | Accepted<br>p=0.45 | Accepted<br>p=0.25 | Accepted<br>p=0.13 | Rejected<br>p=0.05 | Accepted<br>p=0.14 | Accepted<br>p=0.14 | Accepted<br>p=0.13 | Accepted<br>p=0.06 | Accepted<br>p=0.07 |
| Approach 2 | Accepted<br>p=0.18 | Accepted<br>p=0.12 | Accepted<br>p=0.68 | Accepted<br>p=0.28 | Accepted<br>p=0.49 | Accepted<br>p=0.20 | Accepted<br>p=0.17 | Accepted<br>p=0.24 | Rejected<br>p=0.04 |
| Approach 3 | Accepted<br>p=0.59 | Accepted<br>p=0.38 | Accepted<br>p=0.54 | Accepted<br>p=0.16 | Accepted<br>p=0.17 | Accepted<br>p=0.09 | Accepted<br>p=0.16 | Accepted<br>p=0.08 | Rejected<br>p=0.04 |

## 5.3 Discussion

### 5.3.1 Reliability of ultra-short HRV measurements - statistical analysis approach

Regarding the statistical analysis and the Time Domain features (heatmaps in figures 5.7 and 5.8, and graphs in figures 5.9 and 5.10), it is observed that four features have significance levels remaining relatively stable throughout the sliding window duration variation, having a yy interception value close to 100 and a relatively gradual slope. The mentioned features are the SDSD, the RMSSD (basically the normalised version of the SDSD, which explains the similar percentages obtained in both measurements), the pNN50 and the mean NN (mNN). The latter two correspond to the features that exhibit the highest overall stability in the significance study. It is important to note that the linear regression obtained from the pNN50 significance results has a low $R^2$ value (0.50). However, this low value results from a clear outlier in the 10 seconds time frame.

Keeping our attention on the Non-Linear and Geometrical Domain features, these groups have the lowest percentages of runs without significant differences between the compared features. In some cases, linear regressions with yy interception values are much further away from 100 (ApEn, KFD, SI, TI, TINN); in other cases, with very sharp slopes (SD1, SD2, HFD), or with both these characteristics. This was an expected observation considering the literature regarding similar studies on the ultra-short-term HRV measurements. However, the Point Transition Measure (PTM) shows promising results since the yy interception is 94,66% and the slope -3,04, which is a relative soft slope in the overall context. The fact that this feature proposed in Zubair et al. [58] attempts to quantify the temporal variation at the Poincare plot's point-to-point level may help explain the much better significance results when compared to other Non-Linear measurements.

Lastly, regarding the features of the Frequency Domain, we can notice that the features corresponding to the Very-Low-Frequency band have the worst performance in this test. All these features have linear regression yy interception values between 70% and 80% and slope values below -5, which is a relatively sharp slope considering other features. This result is expected considering the current literature. It may be explained since the VLF band includes waves with 25 seconds periods and above (frequencies under 0.04Hz), which means that with a sliding window of fewer than 25 seconds, we cannot capture a full wave, which increases uncertainty. This complication may also be extended to the Low-Frequency Band. Considering the scope of the LF band, which periods range from 6.7 seconds to 25 seconds, with 10 seconds or a 20 seconds time frame, it is not possible to capture a complete oscillation period. Even the LF band may not be the best way to go regarding ultra-short-term HRV measurements. This is well reflected in the obtained results. According to Pechia et al. [35], it is recommended that spectral analyses are performed on stationary recordings lasting at least ten times more than the slower significant signal oscillation period. This may help to explain the quick drop in the acceptance percentage results. In fact, from the 110 seconds window, we soon observe that most features do not have even 50% of the runs without a significant difference to the 180 seconds reference extraction window.

The results obtained based on the Wilcoxon Rank Sum test (figures 5.7, 5.8, 5.9 and 5.10) should be carefully analysed since some features have characteristics that directly affect these test results, which compare the sample distribution and its medians. One example where the statistical significance test results are affected is when the feature being analysed depends, directly or indirectly, on the time frame used for its computation, for instance, the NN50 feature. This feature is the number of consecutive RR intervals differing more than 50 milliseconds. For the same cognitive state, a larger window is expected to catch a larger number of consecutive RR intervals differing more than 50 milliseconds. This fact affects the feature's median values across the time frame reduction, and significant statistical differences will be found comparing this feature extracted with two different sized windows. The features from the Frequency Domain, which compute the total power, are other examples (more oversized windows will expectedly have higher total power values for the same cognitive state). Also, the HFD relies on the parameter "kmax" for its computation, dependent on the window size employed. On the other hand, features which are normalised values, like pNN50, tend to have more consistent acceptance percentages through the window size reduction. In this way, a comparison using an isolated statistical significance test like the two-sided Wilcoxon rank sum, which compares the features medians, may lead to biased results in these features. Furthermore, we believe that having different medians does not mean the feature is not proper to be extracted with smaller windows, considering our current goal, i.e., finding the smallest time frame where each feature behaviour is still representative of the 180-second correspondent measurement, for cognitive stress levels discrimination purposes.

Another problem with the isolated use of the significance analysis was that in many features, the acceptance percentage decreases to zero very early as the window size decreases. This fact gives the false impression that, for instance, in the TI features, using 60-second or 10-second window essentially produces equivalent results. In this way, the correlation results corroborate some considerations made previously during the analysis of the significance tables. In addition, the linear regression obtained for the correlation results has more solid fits, having no $R^2$ values under 0.90, allowing more accurate conclusions. The correlation analysis may give us more insight into how a feature changes with the reduction of the window size and, in this way, help us to evaluate until each window size a particular feature remains reliable in our study conditions.

Regarding the correlation analysis (heatmaps in figures 5.11 and 5.12, and graphs in figures 5.13 and 5.14), starting with the Time Domain HRV feature set, the mNN is the only feature where the correlation means remains above 0.50 until the 60 seconds window (more precisely, its correlation mean remains above 0.50 until the 30 seconds time frame). This feature achieved the highest correlation in the smaller time frame in the study (0.40). From the literature, some studies concluded that the mNN is reliable until the 10 seconds time frame, such as the study by Salahuddin et al. [39], so the expectation was to see higher correlation values until smaller extraction time frames. The same can be said regarding the RMSSD and the pNN50 features. In the literature, these features are often mentioned as being reliable using 60 seconds time frames and lower [59, 63], yet, in the current experiment, the correlation means obtained for these features using the 60 seconds window were already below 0.50. However, in the study performed by Salahuddin et al. [39], some recommendations by Pechia et al. [35] and by Shaffer et al. [38] are not adopted, such as the recommendations regarding the features bias quantification, which may be increasing with the time frame reduction. Also, the mentioned study used a 150-second reference for the statistical analysis, while we used a 180-second time frame as a reference. Furthermore, the existing investigations, such as the study by Baek et al. [63], perform an inter-subject analysis of the features. This fact can lead to biased correlation values since it captures the inter-subject feature tendencies that may overwhelm the actual feature tendencies and increase the correlation of the features. In our present study, we perform an intra-subject and intra-run feature analysis, avoiding this kind of bias. This analysis difference explains the lower correlations obtained in the present study. It is also important to underline that, contrary to the most existing literature on the topic, our study is projected in a real-life, non-controlled environment, emulating contexts for real applications, like bug detection algorithms based on the features under study. Accounting for these considerations, we cannot expect as high correlation values or as clean and clear results as those obtained in more controlled and resting environments, requiring lower cognitive effort.

Regarding the Non-Linear and Geometrical Domain HRV features, some features would probably be overlooked considering only the significance results. Let us take as an example the significance values of the KFD, the SI, the TI and the TINN. The acceptance percentages in the significance test drop to very low values in the 170-, 140-, 130- and 120-second windows, respectively. However, in the correlation results, we can observe the existence of correlation until smaller windows. Actually, in the KFD feature, more than 50% correlation is observed until 80 seconds, and this feature has correlation values similar to HFD. This similarity is expected since both features compute the Fractal Dimension. If the significance test results were the only ones taken into consideration, we could have erroneously concluded that these features are very distinct. From the Geometrical and Non-Linear Domains (figure 5.13), the PTM was the feature which had a higher correlation on the smaller windows and with the softer slope (-0,037) of these two groups, which corroborates the considerations previously done.

Globally speaking, the features from the Frequency Domain are the ones that exhibit higher consistent correlation values for smaller windows. Several features from this domain have mean correlation values above 50% until windows of 40 seconds, with the HF obtaining more than 50% correlation mean also when using the 30 seconds sliding window. This observation is substantially different from the significance results, which could biasedly suggest that Time Domain's features are more reliable in the smaller time frames. The set of features HF, LF, LFpeak and totPow, are the features with the most promising correlation results from this domain, having correlation mean values greater than 25% at the 10 seconds time frame. Analyzing the slopes of the linear regressions achieved using the correlation means, it is observable that the frequency domain features exhibit higher yy interception values, maintaining a relative softer slope. These facts indicate that their tendencies are less impacted by the sliding window size reduction. Once more, as expected, the VLF band had the poorest results from the Frequency Domain set of features, with the steepest linear regression slopes, despite the correlation results not being as low as the literature would suggest until the 60 seconds window, compared to the other measurements. The set of features, with yy interception value of at least 0.95, with the softer slopes of the overall study where: the mNN (a = -0,033), the HF (a = -0,038), the LF (a = -0,039), the LFpeak (a = -0,040) and the totPow (a = -0,040). These features are also the ones with the higher correlation means in the 10 seconds time frame. Both these indicators can mean that this set of measurements is adequate to perform the intended analysis in a code inspection context.

In table 5.6 is presented a summary of the top 5 features by sliding window, according to the correlation mean values obtained. From this table, we can observe that the features from the Frequency Domain clearly stand out. In fact, only one feature from the Time Domain reached these tops, the mNN when the extracting sliding window was 60 seconds or under, being the feature with the highest correlation when using the 10 seconds sliding window. The HF is the more consistent feature with the higher correlation values until the 30 seconds time frame.

**Table 5.6:** Top 5 features by time frame regarding the correlation means.

| **120 secs** | | **90 secs** | | **60 secs** | | **30 secs** | | **10 secs** | |
|---|---|---|---|---|---|---|---|---|---|
| HF | 90% | HF | 82% | HF | 70% | HF | 51% | mNN | 40% |
| LF | 90% | totPow | 80% | totPow | 67% | mNN | 50% | LF | 32% |
| totPow | 89% | LF | 80% | LF | 65% | LF | 48% | LFpeak | 30% |
| LF/HF | 87% | LFpeak | 75% | LFpeak | 61% | totPow | 48% | HF | 28% |
| HFnu | 87% | HFnu | 74% | mNN | 60% | LFpeak | 44% | totPow | 25% |

From the significance and correlation analyses performed, it is observable that every HRV measurement present in the current study is affected by the time frame size used in their extraction. The Bland-Altman plots further corroborate this statement. These plots allow us to observe the generalised increase in the lines of agreement values of the features with the extracting window size decrease. Figure 5.15 corresponds to the Bland-Altman plots of the LF/HF feature extracted with 120, 90, 60, 30 and 10 seconds compared to the same measurement extracted with the 180 seconds time frame. In these illustrative plots, it is possible to observe this increase in the lines of agreement values with the sliding extracting window shortening. The number of measurements that fall out of the lines of agreement also increases with the time frame reduction. In the LF/HF feature, this effect is very clearly observable. In this way, we can conclude that the degree of bias increases with the analysis window size reduction compared to the 180-seconds measurements. However, in this study context, it is observed that the variability which occurs in some features might be due to the fact that the samples extracted from the 180-second window capture a more overall picture of the ANS dynamics, i.e., during a window of 180-second duration a higher degree of variability of the ANS activity might exist due to a higher degree of variability in cognitive stress during that period, in comparison to the samples extracted from the shorter windows, where a lower degree of the variability of cognitive stress is observed. This remark is in accordance with the increased variability observed as the time window duration is decreased. Therefore, given the task's nature and application, the existence of variability on some of the Bland-Altman plots might be considered a concern but not a limitation for the software engineering application, given the high correlations between the two time series of comparison (180-seconds vs shorter windows). These results show that the prevalent cognitive state in both windows is similar but not necessarily equal since larger windows will capture higher cognitive state fluctuations compared to shorter windows. Furthermore, these differences can be readily captured and compensated by current machine learning and statistical techniques used to model risk scores based on HRV.

**Overall Remarks**

Considering the results obtained, it is observed that the chosen time frame significantly impacts every feature in the study. The features from the frequency domain are the ones that maintain higher correlation levels until the smaller extraction window durations. From the set of the considered HRV features in this analysis, 13 features had at least 50% correlation when using the 60 seconds time frame (12 from the frequency domain and only the mNN from the time domain). The lower statistical significance results can be explained by the fact that features like HF or

LF compute the total power of the respective band. Using a window with a larger size will expectedly have higher total power values for the same cognitive state. Despite this fact, these features accurately represent the 180-second correspondent measurement behaviour, as observable in the correlation results. Furthermore, for cognitive stress levels discrimination purposes, we do not need an exact surrogate of the short-term measurements, and the feature behaviour and tendencies resultant from Autonomous Nervous System changes can be used to evaluate different cognitive stress levels.

Regarding the smaller window size in the study (10 seconds), only three features exhibited at least 30% correlation: the mNN, the LF and the LFpeak. Thus, a 10 seconds window time frame is too optimistic in our study context (high cognitive stress). The 30-second time frame is the smallest window with features with at least 50% correlation, and only two fulfilled this criterion, the HF and the mNN. The mNN, the HF, the LF, the LFpeak and the totPow features presented the softer linear regression slopes of the overall correlation analysis, with a yy interception value above 0.95, meaning they are less impacted by the time frame reduction. In this way, this set of five features has shown to be the most reliable for the smallest time frames considering the present context. The mNN feature has proven to be particularly robust to the reduction of the extracting window. This feature has a correlation mean of 50% using a 30 seconds window and showed no significant statistical differences in more than 50% of the experiment runs using all the sliding windows under study while maintaining a low degree of bias compared to the 180-second reference.

Considering all the results, in a cognitively demanding tasks context, a classifier built with features extracted using time frames under 30 seconds might lead to inconsistent results, with potential low scores and high deviations. However, further study is required to assess whether to discard features extracted using smaller time frames in machine learning contexts since these features may catch some shorter cognitive patterns that larger time frames may not be able to discriminate. An approach using classifiers trained with datasets, each composed of features extracted with a different time frame, may offer more extensive insight and help to answer the raised hypothesis. In the next section, we will precisely discuss the impact of the time frames used in the feature extraction process in the complexity classification of software code sections.

### 5.3.2 Impact of Ultra-Short-Term HRV Features in Software Code Sections Complexity Classification

Foremost, regarding the classification procedure (section 5.1.4), it should be underlined why, in the experiment context, we opted to use the statistical transformed HRV features instead of directly using the HRV features. When a programmer inspects software code with multiple sections having different complexities, he is expected to go back and forward through the sections, searching for code flaws and bugs. This behaviour makes that a single code section can be gazed at by the programmer several times. Also, during each gaze, the cognitive load presented by the subject may be different. To illustrate this logic, we may think of the following example: a programmer inspects a particular section for the first time and does not feel any difficulty, although, through the code inspection continuation, he has a doubt and suspects that the answer to his doubt may be in a previous section, so he returns to that initial section, he may now feel a more intense difficulty there, where previously that level of difficulty was not felt.

In the datasets used for the classification process, each data sample corresponds to a particular section gazed by a subject during an experiment run (see section 5.1.2.3). The features characterizing each sample are the statistical transformations of the HRV features extracted during the gazing periods. The importance of using transformed features is to capture and enhance the subject's cognitive state on each specific code section over the experiment. The use of different size sliding windows in the HRV features extraction (before the statistical transformations) is proposed since different time frames may access different ANS dynamics. Furthermore, smaller time frames may provide greater granularity in the HRV analysis, making it possible to capture the ANS dynamics during smaller code section gazing periods. This way, the analysis conducted in this section assesses the impact of the time frames used to extract ultra-short-term HRV features in the complexity classification of software code sections. More specifically, we propose to evaluate which time frames are more suited to employ in the features extraction process in this specific software development or similar contexts.

Through the analysis of figure 5.16, it is possible to observe that the three different approaches achieved similar results. In approach number one, the F1-Scores remain particularly stable until the 60 seconds time frame. Regarding the performances achieved using the 60 seconds time frame dataset, it is observed that the F1-Scores do not belong to the same distribution as F1-Scores obtained using the 180-second reference dataset (see table 5.5). This result appears to be an outlier since the distributions obtained with the time frames below 60 seconds belong to the same distribution as the 180-second ones. This observation is in accordance with the expected outcome since all the datasets used in this approach were based on the most discriminative transformed features of the 180 seconds time frame. From the results achieved with this approach, we can also observe that the performance of the classifiers did not change substantially with the time frame reduction. This statement is confirmed by the range of the F1 scores obtained being between 0.66 and 0.75. The most considerable difference found across the time frame reduction results is in the standard deviation values, which are significantly larger in the results related to the use of smaller time frames. The best result achieved using this approach corresponds to the use of the 140 seconds sliding window.

Analysing the results achieved following the second approach, it is possible to notice an increase in the performance of the classifiers fed with features extracted using smaller sliding windows compared to the first approach results. However, the classification based on the features extracted with the 10 seconds window decreased its performance. Also, the two larger windows obtained higher performances in this approach. These facts were not initially expected since we used the most discriminative features of the 10 seconds time frame dataset in this approach. However, they can be explained by the presence of a compensatory behaviour between transformed features.

Finally, regarding the performances achieved with the third approach, the conclusions are not much different from the other two. This approach used the most discriminative transformed features for each dataset time frame. Therefore, the different time frame classifiers are expected to achieve the higher F1-Scores from the three approaches for each extracting window size and present a lower variability (standard deviation). Nonetheless, one can observe some exceptions believed to be the result of the before-mentioned compensatory behaviour of the different sets of transformed features. The higher performance model of the overall study was obtained with the 130 seconds time frame following this approach. This model obtained a mean F1-Score of 0.75 and a standard deviation of 0.06, indicating that the 130 seconds time frame can be adequate for the current study classification objectives. On the other hand, a 10 seconds window has proven to be too short of a time resolution. This window obtained the lower mean F1-score of the study (0.62), and, in approaches 2 and 3, the distribution of its results did not belong to the same distribution as the 180 seconds reference results.

**Overall Remarks**

Considering the higher F1-Score mean value obtained for each time frame across the three different classification approaches, we can observe that the classification performance remains relatively stable with the extraction window size reduction. The mean F1-scores obtained ranged between 0.75 and 0.62 across all windows and approaches. Furthermore, excluding the 10-second corresponding results, a window that proved to be too short of a time frame in the current context, the mean F1-scores obtained ranged between 0.75 and 0.66. In this way, it is possible to conclude that the reduction of the time frame used for extracting the HRV features before the statistical transformations method application does not substantially affect the results obtained in the classification process. This conclusion is a relevant outcome since most published results regarding the window size impact on the HRV analysis, as well as our results (see section 5.2.1.1), suggest that the lower the window size, the higher the uncertainty of the HRV features. Moreover, the correlation between the feature extracted using the reference window (180 seconds) and the feature extracted using the smaller windows decreases with the time frame reduction (see section 5.2.1.1)—facts expected to impact the classification results. However, in the present study context, when applied in a classifier paired with the proposed statistical transformations method in the dataset construction, the results show that the performance degradation is much smaller than expected.

The higher-than-expected performance results might be explained by the complementary nature of the feature set selected, which can compensate for the uncertainty in each feature. This belief is reinforced since, although some sets of transformed features were considered the most discriminative set in a particular window size dataset (approach 3), other different sets of features ended up having the best F1-Scores. This compensatory effect has also proven to be present across the different time frames since the F1-Score values hold relatively stable with the window size reduction. A possible approach to further explore these compensatory effects may be applying a Principal Component Analysis (PCA) to the features, under the penalty of losing some of its interpretability.

Another explanation for the obtained results stability may reside in the statistical transformation method used in the dataset construction. This procedure reduces the probability of, in a sample, the programmer's difficulty sensation being different from the actual complexity label of the code section gazed. With this method implementation, in each sample's features, it is possible to account for every time a particular section is gazed at by the programmer. In this way, each code section sample accounts for the different difficulty levels, i.e., the different cognitive stress levels, felt at each gaze, helping further extend the compensatory behaviour discussed. Nevertheless, the standard deviation of F1-Scores obtained with the cross-validation method used reveals higher variability in the smaller time frame datasets.

Regarding the features selected as the top 5 most discriminative HRV transformed features by time frame (appendix C), one interesting observation is the predominance of the frequency domain features. Also, the HF feature (with the peaks rate transformation) is the most frequently selected feature across the different time frames. These observations are in accordance with the statistical approach correlation results, corroborating and strengthening the thesis and justifications presented in the 5.3.1 discussion subsection.

Considering the use case reported in this section, where the extraction of HRV features under ultra-short time periods is vital to capture fine events such as the inspection of short but complex code sections, a classifier fed with statistical transformations of features extracted using different time frames could be an optimal solution. It is essential to underline that the conclusions presented in this section are context-specific and should be carefully analysed and further studied.

# 6

# Pupillography measurements

This chapter contains the study developed for defining pupillography frequency bands' limits and investigating the pupillography features' discriminative potential for code section complexity classification. This chapter is divided into three main sections: Methods, where the methodology used in the experiment is described; Results, where the obtained results are presented; and Discussion, where the results obtained are discussed.

## 6.1 Methods

This section describes the experimental methods performed to define the adequate pupillography frequency bands' limits for the frequency features extraction used to feed a software code complexity classifier. Initially, the pupillography signal collected during the Code inspection was pre-processed to remove blink-related and other artefacts. Then the study of the pupillography frequency bands' limits definition was conducted. With the frequency bands defined, the pupillography frequency features were extracted, and statistical transformations of these features were computed. The resultant statistical transformed features were used to classify the complexity of code sections. Finally, the best features from the HRV and the pupillography signal were joined to build a final code section complexity classifier. The flow chart in figure 6.1 schematically represents the practical steps followed.

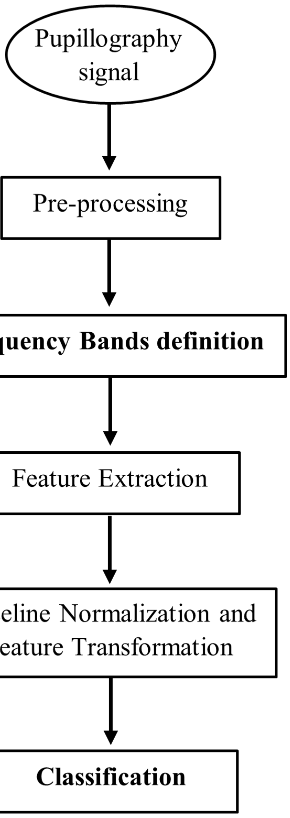

Pupillography signal
Pre-processing
Frequency Bands definition
Feature Extraction
Baseline Normalization and Feature Transformation
Classification

**Figure 6.1:** The flow chart of the practical steps followed for the pupillography signal analysis.

### 6.1.1 Pre-processing

In order to remove blink-related and other artefacts from the pupillography signal of the left eye (i.e., pupil diameter variation signal of the left eye), the same pre-processing methodology as the one implemented by Ricardo et al. [90] was used.

The first step of the pre-processing methodology implemented was to remove all the pupil diameter (PD) readings considered inaccurate. The readings considered inaccurate included the ones marked by the eye tracking as invalid ($\Delta inv$) and the adjacent readings 100 milliseconds before the onset and 100 milliseconds after the offset of the readings marked as invalid since they were observed to be questionably larger.

After the first inaccurate readings removal procedure, we also identified the existence of some PD readings in the pupillography signal, which represented pupil dilations happening too quickly or too deviated from the trend line to reflect the underlying physiological function. In order to detect and remove these inaccurate readings, we performed an outlier detection technique using boxplot analysis [91]. To this extent, we differentiated the pupil diameter signal (PD') and computed the lower (Q1) and upper (Q3) quartiles, i.e., the $25th$ and $75th$ percentiles. Following this step, we computed the interquartile range equal to the difference between the lower and upper quartiles: $IQR = Q3 - Q1$ [91]. Then, the PD reading corresponding to the $t$ instant is marked as an outlier when:

$$PD'(t) < Q1 - 1.5IQR \quad \vee \quad PD'(t) > Q3 + 1.5IQR \tag{6.1}$$

Concluded the exclusion of all the outliers and inaccurate readings, we performed a shape-preserving piecewise cubic interpolation to interpolate the excluded values. Then, we down-sampled the pupillography signal resultant from this procedure to 20 Hz. This way, the data size was significantly reduced without interfering with the study frequency of interest (between 0 and 10 Hz).

Furthermore, it is essential to minimize the impact of eye blinks and other external factors artifacts in the pupillography time series, which significantly affect the time and frequency analysis of this signal, as stated by Nakayama et al. [72]. To this extent, we employed an algorithm which uses Singular Spectrum Analysis (iterative SSA) [92–94] to fill the signal missing data. This algorithm decomposes the initial signal into multiple components which have a meaningful interpretation, such as oscillatory modes, trends, or noise. Then the missing data gaps are iteratively reconstructed based on an arbitrary number of components.

Finally, the pupillography signal was filtered using a high-pass filter with a cut-off frequency of $4 \times 10^{-4}$ Hz to reduce the impact of medium-term nonstationary components present in the time interval being analyzed [95].

### 6.1.2 Feature Extraction

For the feature extraction from the Code inspection pupillography data collected during each subject run, we used a sliding window with 180 seconds and a jumping step of 1 second. A total of 12 features were computed from the frequency domain. In the same way as the HRV features (section 5.1.2), the described procedure produces vectors of individual measurements from the pupil diameter time series collected during the Code inspection task (to facilitate referencing, we will call these the 'Extracted Feature Vectors'). Each individual measurement is computed based on a pupillography signal portion with the size of the sliding window employed. The individual measurements are then associated with the time instant corresponding to the center of the pupillography signal portion used to compute the respective individual measurement.

In this study, the 'Extracted Feature Vectors' are directly used in the frequency bands definition study. We used the HF, the LF, and the LF/HF features in the band definition study. After the bands were defined, all 12 features were extracted based on that defined bands.

#### 6.1.2.1 Feature Description

In order to study the Power distribution as a function of the frequency, we employed Burg's autoregressive method with order 16 to estimate the Power Spectral Density (PSD) of the pupil diameter time series. The order used in the Burg's method was assessed using the partial autocorrelation sequence. Completed this step, the LF and HF bands were computed using several frequency band limits and the LF, HF and LF/HF features were extracted based on that different bands. These features were employed for assessing the adequate bands' limits definition. After the adequate bands' limits were defined, the LF and HF bands were computed based on the adequate limits obtained. Then, the following Power features were extracted by calculating the area under the PSD curve:

- The total Power (**totPow**) is the sum of the Power of all frequencies considered of potential interest (0 to 1 Hz).
- The **Peak** is the maximum Power across every frequency of interest (0 to 1 Hz).
- **LF** and **HF** are, respectively, the sums of the total Power on the low-frequency and high-frequency bands.
- LF normalized (**LFnu**) and HF normalized (**HFnu**) are the frequency bands relative Power. These features correspond to the LF and HF measurements

normalized by the total Power.

- **LFpeak** and **HFpeak** are, respectively, the maximum Power on the low-frequency and high-frequency bands.
- LFpeak normalized (**LFpeak-nu**) and HFpeak normalized (**HFpeak-nu**) are the frequency bands' relative peaks. These features correspond to the LFpeak and HFpeak divided by the Peak.
- The ratio between the LF band's maximum Power and the HF band's maximum Power (**LFpeak/HFpeak**) consists of the quotient between the LFpeak and the HFpeak features. Finally, the ratio between the LF band's total Power and the HF band's total Power (**LF/HF**) is the quotient between the LF and the HF measurements.

#### 6.1.2.2 Baseline Normalization

In order to reduce the inter-subject and inter-run variability, i.e., the variability between different subjects and the internal variability of a subject during the experiment (different runs), all the pupillography features extracted during the "code inspection" task were normalized before the feature transformation used in the classification following the same procedure used on the HRV measurements. The data collected during the "natural language reading" periods was used as a baseline in the normalization process. With this intent, the features described in section 6.1.2.1 were also extracted from the pupil diameter data collected during the "natural language reading" task periods. To facilitate the reference, let us call these the 'rest features' and the features regarding the "code inspection" task the 'code features'. The normalized features were computed by calculating the ratio between each "code feature" and the corresponding experimental run "rest feature" median. The median has been selected to perform this baseline normalization since the data does not follow a normal distribution (assessed using the Kolmogorov–Smirnov test).

#### 6.1.2.3 Feature Transformation

In order to capture and enhance the cognitive stress state presented by the subjects at the different code complexity sections (low or high complexity) during the code inspection task, statistical feature transformations were computed. To this extent, the individual measurements present in the 'Extracted Feature Vectors' (produced in the feature extraction process, see section 6.1.2), after the normalization process (section 6.1.2.2), were grouped based on all the instants the subject was looking to a specific section during a run. From each group, a set of features was computed employing statistic transformations, the same procedure used for the HRV measurements (see section 5.1.2.3). The feature transformations computed were:

- Simple statistic transformations:
  - mean
  - standard deviation
  - maximum
  - minimum
  - median
  - quantile 0,50
  - quantile 0,75
  - quantile 0,85
  - quantile 0,95
- Peak statistic transformations, where the grouped measurements local maxima are extracted, and then the simple statistic transformations are computed:
  - peaks mean
  - peaks standard deviation
  - peaks maximum
  - peaks minimum
  - peaks median
  - peaks quantile 0,50
  - peaks quantile 0,75
  - peaks quantile 0,85
  - peaks quantile 0,95
  - peaks rate (ratio of local maxima)

This process resulted in 228 'Transformed Features' (12 features (section 6.1.2) x 19 statistical transformations). In summary, in the resultant dataset, each software code section gazed during a run is an instance, labelled according to their difficulty, and the dataset features are the 228 'Transformed Features'. The resultant datasets are the ones used in the classification.

### 6.1.3 Frequency Bands definition

In order to carry out the study of the adequate pupillography frequency bands' limits for frequency features extraction, we started by using Burg's autoregressive method with order 16 to estimate the Power Spectral Density (PSD) of the pupil diameter time series. To this extent, we used a 180-second sliding window with a jumping step of 1 second to analyse the data collected during the Code inspection task. This procedure forms a vector with the PSD of each window analysed. Then, we computed the mean of the PSD for each subject in the study and the overall PSD average. Figure 6.2 corresponds to the PSD averages (overall and by subject) plots regarding the code inspection experiment task data.

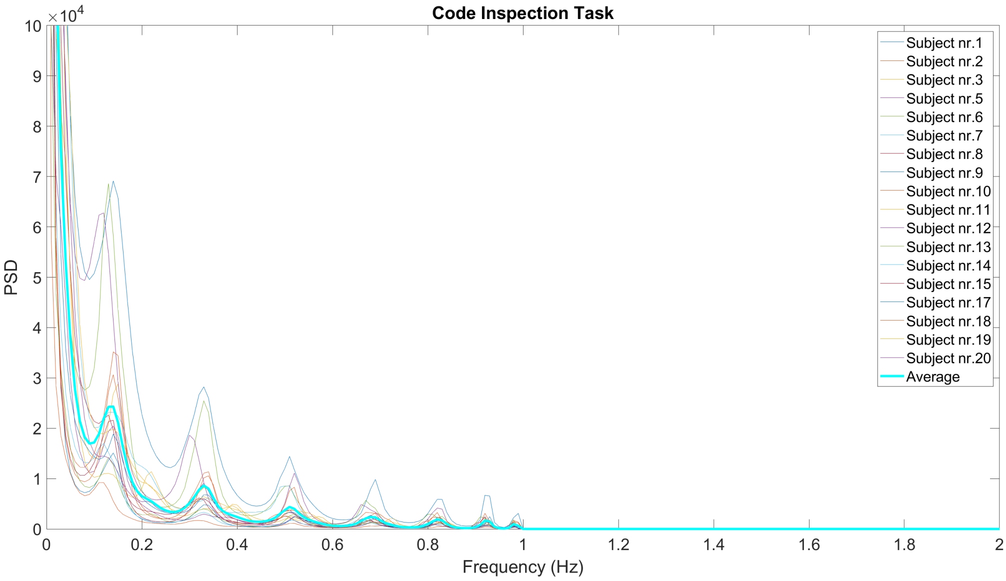


**Figure 6.2:** Plot of the PSD averages, computed by subject and overall, from the pupillography signal collected during the code inspection task.

From the figure 6.2 it is possible to conclude that the frequencies of interest for the pupillography spectrum analysis are between 0 and 1 Hz. With this in mind, in order to find the limits for the LF and HF bands, we computed the power of 51594 band combinations using the PSD computed with the 180-second sliding window with a jumping step of 1 second from the signal collected during the Code inspection task. The band combinations tested were built by defining an initial limit for the LF band (ranging from 0 to 0.20 Hz in 0.01 steps) and a final limit for this band (ranging from 0.10 to 0.50 Hz in 0.01 steps), always respecting the condition of the final limit being superior to the initial one. Then the initial limit of the HF band is defined as the final limit of the LF band, and the HF band's final limit is ranged from 0.30 to 1 Hz in 0.01 Hz steps, respecting the same condition of the final limit being superior to the initial one. The ratio between the power of the LF and HF resultant bands was also computed.

In order to select the most suitable combination of bands, we used the LF, HF and LF/HF measurements extracted from the HRV resultant from the data collected during the Code inspection task, using the 180-second sliding window with a jumping step of 1 second (section 5.1.2), as ground truth. We started by performing the Kolmogorov-Smirnov test individually by measurement in each experiment run to assess the data distribution. With this test, we concluded that the features did not follow a normal distribution, so the correlation test used to compare the HRV and the pupillography measurements should be non-parametric. This way, we computed Spearman's correlation test comparing the LF, HF, and LF/HF HRV features with the LF, HF and LF/HF pupillography features resultant from the frequency bands tested. The procedure was conducted independently for each run (intra-group analysis).

The Spearman's correlation test output is a p-value and a correlation coefficient. The p-value is used to determine if a significant correlation exists between the data compared, while the correlation coefficient measures how correlated they are. This way, after the correlation test procedure, for each experiment run of the different subjects, a p-value and a correlation coefficient are obtained for the LF, HF, and LF/HF for each band combination tested. After this step, we computed the percentages of runs having significant correlation and calculated the correlation coefficient means across the different runs for the LF, HF, and LF/HF measurements for each band combination tested. Due to the presence of runs with different sample sizes, the means were computed using Fisher's z Transformation, which allows us to give more weight to the feature extracted in runs with a larger number of samples [88].

With the correlation results obtained, to select the most suitable LF and HF band combination, two different approaches were followed:

- **Approach 1)** In this approach, the goal was to simultaneously maximize the LF and HF pupillography bands' power correlation with the LF and HF of the HRV signal. To this extent, we selected the top 10 bands' combinations based on the geometric mean between the LF and the HF correlation mean values across the runs. After this step, the LF, the HF, the LF/HF and the geometric mean of the LF and the HF ($\sqrt{LF * HF}$), correlation means and percentages of runs having a significant correlation of the obtained top 10 were arranged in heatmap tables.

- **Approach 2)** This approach aimed to maximize the pupillography LF/HF feature correlation with the LF/HF feature of the HRV signal. To this extent, we selected the top 10 bands' combinations based on the obtained LF/HF correlation means. After this step, the LF, the HF, the LF/HF and the geometric mean of the LF and the HF ($\sqrt{LF * HF}$), correlation means and percentages of runs having a significant correlation of the obtained top 10 were arranged in heatmap tables.

The top 10 bands' combinations results from the two approaches were observed and analyzed to select the most suitable LF and HF bands' limits for the frequency features extraction. We opted to conduct these two different approaches given the nature of the LF, HF and LF/HF features. As seen in the State of the Art chapter (chapter 3), the LF is more associated with the sympathetic nervous system, while the HF is linked to the parasympathetic nervous system. The LF/HF ratio represents the balance between the two systems. This way, if we maximize the isolated correlation of the LF and the HF features, but the LF/HF ratio obtains a low correlation, it can be due to the fact that we are not selecting the correct frequency bands' limits. The inverse logic is also true. Following the two described approaches and crossing the results obtained, we reduce the hypothesis of not selecting the correct frequency bands' limits. The flow chart in figure 6.3 summarizes the practical steps followed for the pupillography frequency bands definition.

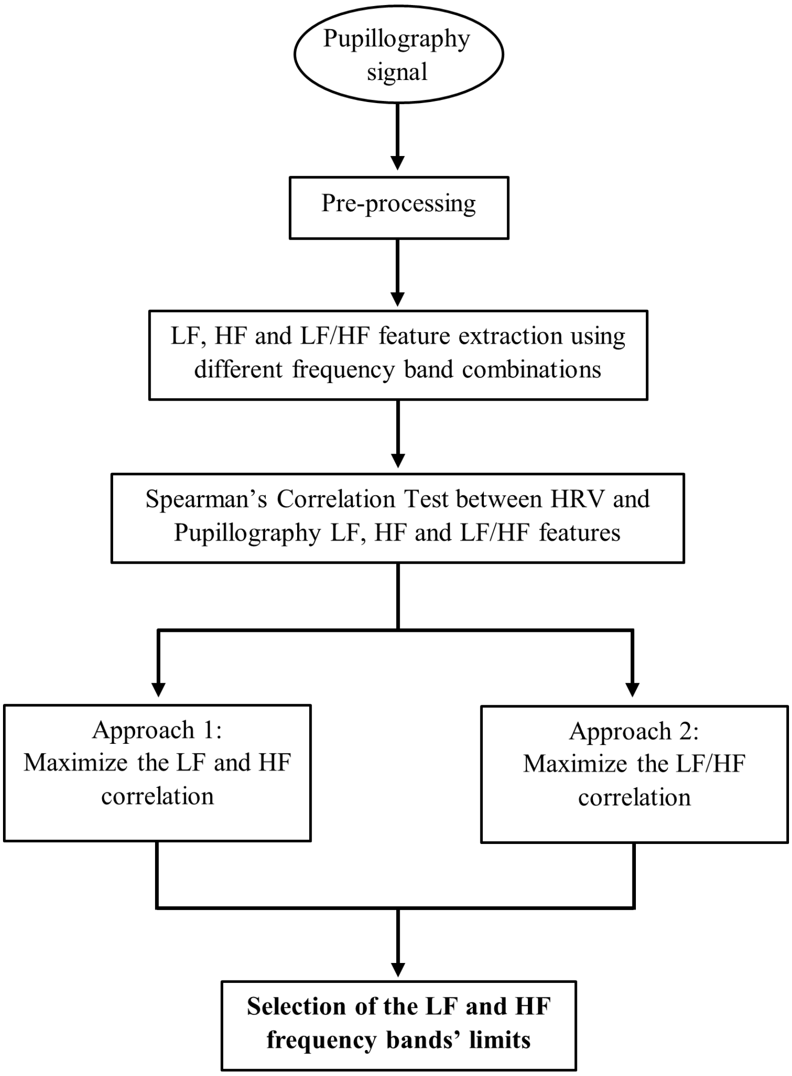

Pupillography signal
Pre-processing
LF, HF and LF/HF feature extraction using different frequency band combinations
Spearman's Correlation Test between HRV and Pupillography LF, HF and LF/HF features
Approach 1: Maximize the LF and HF correlation
Approach 2: Maximize the LF/HF correlation
Selection of the LF and HF frequency bands' limits

**Figure 6.3:** Flow chart of the practical steps followed for the pupillography frequency bands definition.

### 6.1.4 Classification

In order to investigate the pupillography transformed features' ability to discriminate between low and high complexity code sections, we built a classifier fed with the features based on the frequency bands' limits previously defined extracted with a 180-second sliding window and assessed its performance (figure 6.4). The performance obtained with these features was then compared to the performance achieved by the classifier fed by the HRV transformed features, resulting from the statistical transformation of the ultra-short-term HRV measurements extracted using a 180-second sliding window (section 5.1.4).

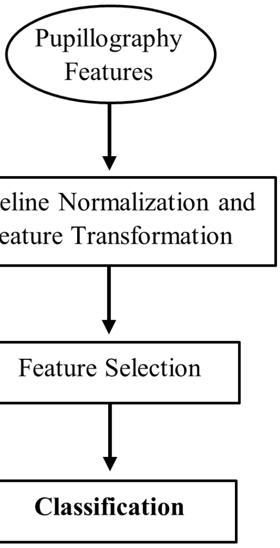


**Figure 6.4:** Flow chart of the steps followed for the classification using the pupillography features based on the frequency bands' limits previously defined extracted with a 180-second sliding window.

Furthermore, in order to test the existence of a complementary relationship between HRV and pupillography features, we simultaneously searched for the most discriminative features from both HRV and pupillography transformed features (both extracted using the 180-second time frame). The features selected through this search were then used to train another SVM classifier (figure 6.5).

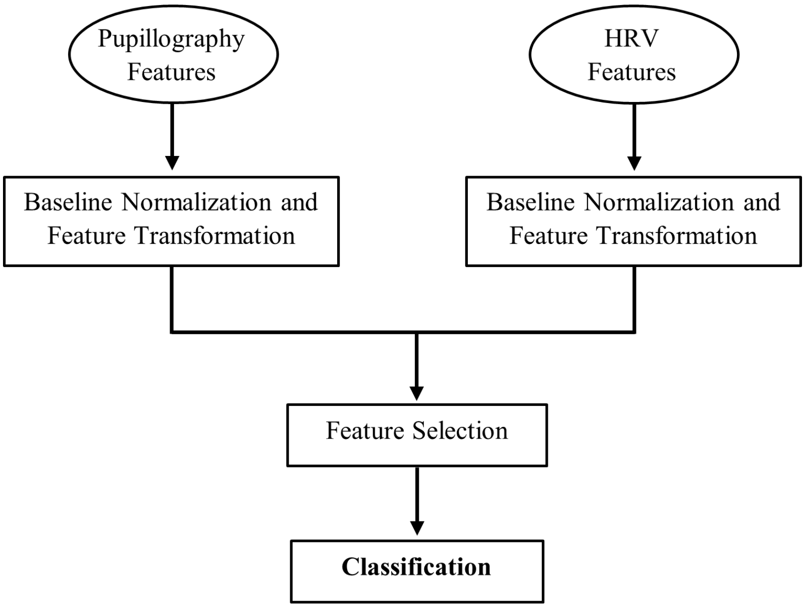

Pupillography Features
HRV Features
Baseline Normalization and Feature Transformation
Baseline Normalization and Feature Transformation
Feature Selection
Classification

**Figure 6.5:** Flow chart of the steps followed for the classification using both pupillography and HRV features extracted with a 180-second sliding window.

#### 6.1.4.1 Cross-validation scheme

The classification procedure was performed using a Nested Leave-One-Subject-Out cross-validation scheme and SVM classifiers with linear kernel. Using the same procedure described in section 5.1.4.1 of the HRV measurements chapter, we used the inner loop of the nested cross-validation scheme to conduct a grid search, which allows the optimization of several classification parameters. Among these parameters are the classifier's regularization parameter (C), the number of features selected that optimizes the classification results and the selected features. The classification performance assessment is conducted in the nested cross-validation outer loop.

#### 6.1.4.2 Feature Selection

At this point of the study, we have 228 transformed pupillography features obtained (section 6.1.2.3). In order to reduce the number of features, we selected the ones considered the most discriminative of the code sections' complexity (low vs high) using the Kruskal-Wallis test. Similarly to the process performed in the HRV classification approach (section 5.1.4.2), we conducted a grid search to find the number of features which optimized the classifier performance. The results indicated that five was the number of features which optimized the performance. This way, we selected the five most occurring transformed features as the most discriminative top 5 features in this search.

The grid search was then repeated using pupillography and HRV transformed features simultaneously, making up a total of 817 transformed features (228 pupillography + 589 HRV) present in this search. The conclusion was that five was again the number of features which optimized the classification performance. This way, the five most occurring transformed features as the most discriminative top 5 features in this search were selected for the SVM classifiers train and test.

#### 6.1.4.3 Classifier Train and Test

A SVM classifier was trained and tested using the top 5 most discriminative features selected in the grid search based on the pupillography transformed features. This classifier performance was compared to the performance achieved by the classifier based on the HRV transformed features resulting from the 180-second time frame (section 5.1.4). A second SVM classifier was trained and tested using the top 5 most discriminative features selected in the grid search based on simultaneously pupillography and HRV transformed features.

As mentioned in section 6.1.4.1, the classification procedure was performed using a Nested Leave-One-Subject-Out cross-validation scheme for each of the two described classifiers. In order to assess the performance of each classifier, the F-measure (or F1-score) was the evaluation metric selected (see section 5.1.4.3 to consult the F-measure metric description). The average and standard deviation of the F-measures obtained for the respective classifier were computed in the outer loop of the Nested Leave-One-Subject-Out cross-validation scheme.

## 6.2 Results

### 6.2.1 Frequency Bands definition

Figures 6.6 and 6.7 represent the results obtained following the section 6.1.3 methodology. Figure 6.6 corresponds to the top 10 mean correlation results based on the geometric mean between the LF and HF mean correlation across runs criteria (approach 1), obtained for each band combination tested. Figure 6.7 correspond to the top 10 mean correlation results base on the LF/HF mean correlation across runs criteria (approach 2) obtained for each band combination tested.

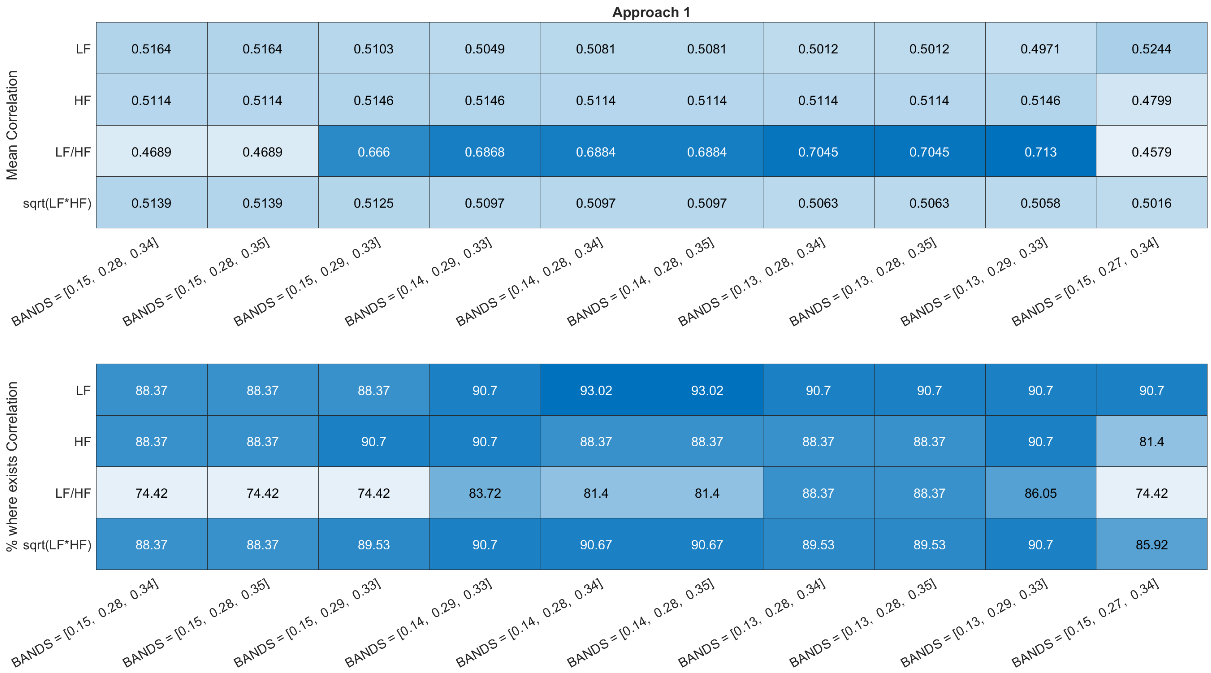
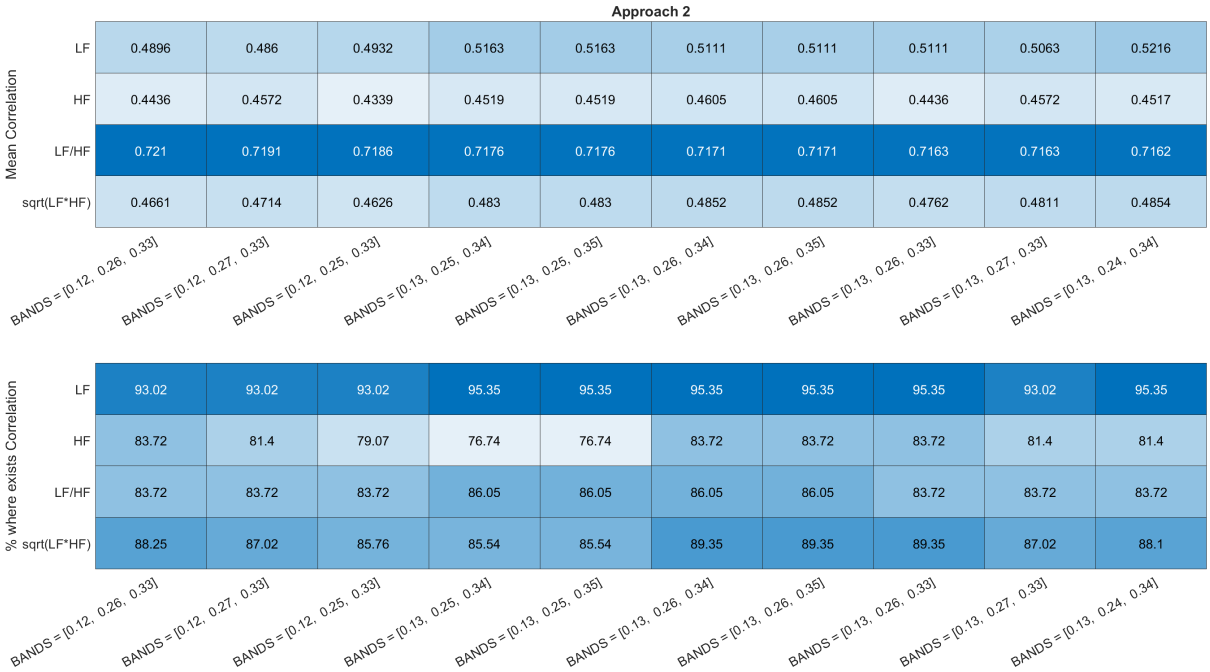


**Figure 6.6:** Correlation means (first heatmap) and percentage of runs with significant correlation (second heatmap) between the LF, HF, LF/HF pupillography and the HRV features, top 10 bands results following approach 1. The fourth row on the heatmaps corresponds to the geometric mean between the LF and HF results.



| | BANDS = [0.12, 0.26, 0.33] | BANDS = [0.12, 0.27, 0.33] | BANDS = [0.12, 0.25, 0.33] | BANDS = [0.13, 0.25, 0.34] | BANDS = [0.13, 0.25, 0.35] | BANDS = [0.13, 0.26, 0.34] | BANDS = [0.13, 0.26, 0.35] | BANDS = [0.13, 0.26, 0.33] | BANDS = [0.13, 0.27, 0.33] | BANDS = [0.13, 0.24, 0.34] |
|---|---|---|---|---|---|---|---|---|---|---|
| LF | 0.4896 | 0.486 | 0.4932 | 0.5163 | 0.5163 | 0.5111 | 0.5111 | 0.5111 | 0.5063 | 0.5216 |
| HF | 0.4436 | 0.4572 | 0.4339 | 0.4519 | 0.4519 | 0.4605 | 0.4605 | 0.4436 | 0.4572 | 0.4517 |
| LF/HF | 0.721 | 0.7191 | 0.7186 | 0.7176 | 0.7176 | 0.7171 | 0.7171 | 0.7163 | 0.7163 | 0.7162 |
| sqrt(LF*HF) | 0.4661 | 0.4714 | 0.4626 | 0.483 | 0.483 | 0.4852 | 0.4852 | 0.4762 | 0.4811 | 0.4854 |

| | BANDS = [0.12, 0.26, 0.33] | BANDS = [0.12, 0.27, 0.33] | BANDS = [0.12, 0.25, 0.33] | BANDS = [0.13, 0.25, 0.34] | BANDS = [0.13, 0.25, 0.35] | BANDS = [0.13, 0.26, 0.34] | BANDS = [0.13, 0.26, 0.35] | BANDS = [0.13, 0.26, 0.33] | BANDS = [0.13, 0.27, 0.33] | BANDS = [0.13, 0.24, 0.34] |
|---|---|---|---|---|---|---|---|---|---|---|
| LF | 93.02 | 93.02 | 93.02 | 95.35 | 95.35 | 95.35 | 95.35 | 95.35 | 93.02 | 95.35 |
| HF | 83.72 | 81.4 | 79.07 | 76.74 | 76.74 | 83.72 | 83.72 | 83.72 | 81.4 | 81.4 |
| LF/HF | 83.72 | 83.72 | 83.72 | 86.05 | 86.05 | 86.05 | 86.05 | 83.72 | 83.72 | 83.72 |
| sqrt(LF*HF) | 88.25 | 87.02 | 85.76 | 85.54 | 85.54 | 89.35 | 89.35 | 89.35 | 87.02 | 88.1 |

**Figure 6.7:** Correlation means (first heatmap) and percentage of runs with significant correlation (second heatmap) between the LF, HF, LF/HF pupillography and the HRV features, top 10 bands results following approach 2. The fourth row on the heatmaps corresponds to the geometric mean between the LF and HF results.

The first heatmap in each figure represents the mean correlation results of the LF, HF, LF/HF and the geometric mean between the LF and HF band results (sqrt(LF*HF)). The second heatmap in each figure is the percentage of runs having a significant correlation for the same LF, HF, LF/HF and the geometric mean between the LF and HF band results (sqrt(LF*HF)). The bands' combination used in the pupillography band power extraction are identified using three numbers: the first corresponds to the beginning of the LF band; the second to the end of the LF band, and the beginning of the HF band (immediately after the end of the LF band); and finally, the third number corresponds to the end of the HF band. For example: "BANDS = [0.15, 0.28, 0.34]" means that the tested LF band was from 0.15Hz to 0.28Hz (LF = [0.15, 0.28]Hz), and the HF from 0.28Hz to 0.34Hz (HF = ]0.28, 0.34]).

### 6.2.2 Classification

Table 6.1 contains the top 5 most discriminative pupillography features. The occurrence is the percentage of times the feature is present in the five most discriminative features set at the inner loop of the nested cross-validation scheme. The top 5 mean position corresponds to the average position the feature occupied when present in the five most discriminative features set, which varies between 1 and 5.

**Table 6.1:** Top 5 most discriminative Pupillography Features.

**180 seconds Time Frame**

| Pupillography Feature | Transformation | Occurrence (%) | Top 5 Mean Position |
|---|---|---|---|
| LF | peaks rate | 100.0 | 1.2 |
| HF | peaks rate | 100.0 | 1.9 |
| HFpeak | peaks rate | 80.22 | 3.5 |
| LFnu | peaks rate | 72.53 | 3.9 |
| LF/HF | peaks rate | 51.65 | 4.1 |

Table 6.2 presents the top 5 most discriminative features resultant from the grid search conducted using all HRV and Pupillography features.

**Table 6.2:** Top 5 most discriminative overall Features.

**180 seconds Time Frame**

| Signal | Feature | Transformation | Occurrence (%) | Top Mean 5 Position |
|---|---|---|---|---|
| Pupillography | LF | peaks rate | 100.0 | 1.5 |
| Pupillography | HF | peaks rate | 100.0 | 2.2 |
| Pupillography | LFnu | peaks rate | 91.21 | 3.1 |
| Pupillography | HFpeak | peaks rate | 86.81 | 3.4 |
| Pupillography | HFnu | peaks rate | 26.37 | 4.5 |

Table 6.3 presents the mean and standard deviation F1-Scores of the classifiers trained and tested using the top 5 most discriminative features of the HRV (section 5.1.4, appendix C), Pupillography (table 6.1) and overall (HRV + Pupillography;

table 6.2), extracted with the 180-second window.

**Table 6.3:** Performance obtained by the top 5 most discriminative features of the HRV, Pupillography and overall (HRV + Pupillography).

| | **HRV** | **Pupillography** | **HRV + Pupillograqhy** |
|---|---|---|---|
| Mean F1-scores | 0,74 | 0,76 | 0,76 |
| Standard Deviation of the F1-scores | 0,10 | 0,07 | 0,10 |

## 6.3 Discussion

### 6.3.1 Frequency Bands definition

Considering the top 10 results obtained following the two approaches (Figures 6.6 and 6.7), it is possible to verify consistent results for the bands' regions that maximize the correlation with the HRV bands. This consistency indicates that the real pupillography LF and HF bands are situated around the region from 0.12Hz to 0.35Hz. Regarding the first approach, we can verify the existence of bands that maximize the simultaneous correlation of both the LF and HF features. However, in some band combinations, although the simultaneous correlation of the LF and HF power features, the correlation of the LF/HF ratio is not optimized. On the other hand, considering the second approach, the opposite is verified, with some bands maximizing the LF/HF correlation but not the individual band combinations. This observation is especially true for the HF band power, in which the correlation is not above 0.50 in any of the top 10 combinations of this second approach. This fact can be related to the respiration effect on the HRV, which was not accounted for in the current HRV signal pre-processing.

In order to find the most satisfactory and generalizable result, both the individual LF and HF band and LF/HF correlation should be maximized without compromising the correlation of the other. Considering this, two band combinations of approach 1 obtained an individual LF and HF correlation above 0.50 and an LF/HF correlation above 0.70. These band combinations correspond to the LF band from 0.13Hz to 0.28Hz and the HF band from 0.28Hz to 0.34Hz for one combination and from 0.28Hz to 0.35Hz in the other. Both these combinations have generalizable results and high correlation results since, in addition to high mean correlation, these combinations also achieved a high percentage of runs having significant correlation. As the results of the two combinations were the same, we selected the LF band as being from 0.13Hz to 0.28Hz and the HF band from 0.28Hz to 0.35Hz since a slightly wider HF band may be more generalizable in different scenarios. The features described in section 6.1.2.1 were then extracted from these bands.

### 6.3.2 Classification

Starting with the analysis of the classifier based on the top 5 most discriminative features from the pupillography signal (table 6.1), it obtained a mean F1-score of 0.76 with a standard deviation of 0.07. This result is in line with the result obtained by the classifier based on the HRV features extracted using the 180 seconds time frame (0.74 ± 0.10). In fact, the pupillography-based classifier obtained the best performance in the overall study, having the highest mean F1-Score with the lowest variability.

Regarding the classification using the top 5 most discriminative features resultant from the grid search conducted using all HRV and Pupillography features (F1-score: 0.76 ± 0.10), the first noticeable result is that all the top 5 features resulted from the Pupillography signal (table 6.2). Another interesting observation is that the top 5 most discriminative features resultant from the grid search conducted using all HRV and Pupillography features have one different feature from the top 5 produced by the search using exclusively Pupillography features. This observation is explained precisely by the inclusion of the HRV features, which alter the occurrence of each feature, also explaining the different variability obtained. In this way, the feature set obtained through the search using exclusively Pupillography features produces more reliable results (i.e., with lower variability), being the best performance classifier in the overall study.

Considering the present results, we can conclude that it is possible to achieve similar performances using pupillography features to those achieved using HRV features. In fact, the pupillography features have proven to possess a higher discriminative power than the HRV features, leading to higher classification performances. However, the obtained results were unexpected since the pupillography frequency features were based on the frequency bands' limits defined using the HRV signal as the reference (section 6.1.3) and should be carefully analysed. Furthermore, the dataset used for the bands' definition was the same employed for the classifiers' train and test. In this way, we can be in the presence of an overfitting result, and the present results should be confirmed by testing the pupillography features in a different dataset.

For future work, the present results should also be tested in other time frames shorter than 180 seconds. A similar study to the one used to evaluate the HRV features reliability in smaller windows should also be conducted to assess these features' discrimination power in shorter time segments. Furthermore, since the frequency of interest of the pupillography PSD ranges from 0Hz to approximately 1Hz (as seen in figure 6.2), a study should be conducted to test the existence of different frequency bands besides the LF and HF bands, which may assess distinct ANS dynamics. Finally, additional features related to eye-tracking can be extracted outside the pupillography frequency domain, such as the blink duration, the blink rate or the fixed staring duration, among others.

# 7

# Threats to Validity and Future Work

In summary, overall promising results were achieved, indicating that both the HRV and the pupillography signals can be used to develop a non-invasive tool for identifying the higher complexity code sections using the programmer's biofeedback. However, some study limitations were present, which translated into threats to our conclusion's validity. First, it is essential to mention that the data collection study was designed with a broader goal and not specifically for HRV stability assessment. As such, several different biosignals and images were collected. Functional Magnetic Resonance Imaging (fMRI) was one of the exams performed. This exam forces the experiment to be conducted inside an fMRI scanner. The fMRI has an inherent noise effect on the ECG signal. This effect was mitigated through several ECG pre-processing and segmentation methods (section 5.1.1). The methods employed effectively mitigate the fMRI noise and are capable of detecting ECG peaks, which are necessary to compute a quality HRV signal. Regarding the pupillography time series, is essential to mention that obtaining a perfectly clean pupillography signal is nearly unachievable, even with the best pre-processing methodologies, due to blink-related and other artefacts

It is also important to underline that the subjects were alone in a quiet and isolated room when performing the tasks to control the experimental environment. Furthermore, the subjects were informed apriori about all the protocols and processes of the experiment and were instructed not to take anything that could stimulate/inhibit them the day before the experiment. Nevertheless, these external effects are minimized, given that the potential effect is blurred as we perform an intra-subject analysis in the HRV statistical analysis (section 5.1.3) and in the pupillography frequency bands definition (section 6.1.3), and the external effects are present in the different measurements compared. Regarding the classification analysis (sections 5.1.4 and 6.1.4), the extracted features were baseline normalized before the classification procedures, which also reduces these potential external effects.

Another limitation of our study was the time frame employed for the HRV features extraction (180 seconds) that were used as the golden standard for the different studies conducted. The measurements extracted with a 180-second window are already considered ultra-short-term HRV measurements. Ideally, a 5 minutes (300 seconds) window reference would be preferable since this is a well-known and consensual time frame in the scientific community. That being said, this was not possible due to our dataset constraints. From our original 21x4 (subjects x runs by subject) runs, we had a few middle run dropouts, which led to only 47 having more than 180 seconds, considering the HRV measurements study. For the pupillography frequency bands study, this number is further reduced to 43 runs due to time accordance between the data extracted with the ECG and the eye tracker (pupil diameter time series). If the chosen reference were 300 seconds measurements, the dataset would be substantially reduced, leading to lower statistical power. Furthermore, the study is performed during software code inspection tasks (i.e., bug detection), which is a highly complex, dynamic, and cognitively demanding task - in this study context, a 5 minutes window is a considerably large window. A window of this size would capture physiological data corresponding to more than one code section, where the subject could feel different difficulty levels, leading to inaccurate results since it would capture different ANS dynamics. Another relevant constraint in the dataset is the fact that all study subjects had the same gender (male). This fact is hard to counter once the software engineering and programming field are largely dominated by male subjects, which makes it challenging to balance the gender groups in the experiment.

Regarding the study context, most of the related work carried out until now was developed with the subjects at rest or performing elementary tasks in very controlled environments. In contrast, our study is done in a highly demanding task environment. Naturally, the dynamic characteristics of the higher cognitive function, resulting from our experiment context, will generate more dynamic signals. Also, the code sections inspected do not have all the same complexity. In this way, the transition from one code section to the next is expected to produce physiological signals with different characteristics and patterns, which are expected to present high variability in these periods, impacting the analysis performed. Another limitation regarding the experiment design is the code snippets used for code inspection. These snippets were developed to be representative of real-world software. However, due to time constraints and practical reasons (the study was conducted inside an MRI machine), the snippets used could not be as long as an actual software code. Furthermore, the fact that the present experimental design protocol did not account for daytime cognitive stress variations or more extended ANS dynamics is another limitation.

In order to complement the present work, additional future work should be conducted. First, regarding the HRV classification impact study, a classifier fed with statistical transformations of features extracted using different time frames should be tested since different features at different time resolutions may assess distinct ANS dynamics and patterns. Furthermore, additional classification algorithms should be tested, such as Decision Trees, Random Forest or K-Nearest Neighbours, to name a few, and compare the performances obtained to choose the most suitable technique. Another significant effect that should be studied is the influence of the breathing effect on the HRV, which was not accounted for in the current HRV signal preprocessing.

Regarding the pupillography features, a similar study to the ultra-short-term HRV analysis should be conducted to assess the adequate time resolution for these features and how the time frames impact the pupillography classification results. Moreover, additional eye-tracking-related features can be extracted outside the pupillography frequency domain, such as the blink duration, the blink rate or the fixed staring duration, among others.

# 8

# Conclusions

The present work assessed the quality and reliability of Heart Rate Variability (HRV) and Pupillography (Pupil Diameter time series) measurements for cognitive stress discrimination in a code inspection context. Regarding the HRV measurements, a statistical and a classification approach were conducted to investigate the impact of reducing the duration of time frames on the HRV feature extraction process and determining whether HRV-based tools can effectively be used in software development environments. Concerning the Pupillography measurements, the present work investigated the adequate LF and HF band limits for the feature extraction and the acute stress discrimination capacity of the features extracted from these bands to discern between low and high complexity code sections.

The HRV statistical approach studied the reliability of 31 ultra-short-term HRV features extracted using time frames of variable sizes (ranging from 3 minutes down to 10 seconds) in a code inspection context. In this analysis, we investigated the smallest time frame where each feature behaviour is still representative of the 180-second correspondent measurement (used as reference). From this approach, the main results show 13 features that presented at least 50% correlation when using 60-second windows. The HF and mean NN features achieved around 50% correlation using a 30-second window, and this window was the smallest time frame considered to have reliable measurements. From this investigation, a set of five features could be identified as the most reliable for the smallest time frames considering the present context: the mean NN, the HF, the LF, the LFpeak and the totPow features. Furthermore, the mean NN feature proved particularly robust to the time resolution reduction.

The HRV classification approach analyzed the impact of the extracting window in the complexity classification of software code sections, using Support Vector Machine (SVM) classifiers. The HRV features extracted with the different time frames were associated with the corresponding code section gazed at the extraction time, and statistical transformations of these features were computed. The F1-Scores obtained for the different classifications ranged from 0.62 to 0.75 across all windows. Furthermore, excluding the 10-second corresponding results, a window that proved to be too short of a time frame in the current context, the mean F1-scores obtained ranged between 0.66 and 0.75, indicating that it is possible to achieve similar classification performances using smaller time frames. However, it was observed a consistent increase in the performance results' variability with the time frame reduction.

Regarding the pupillography measurements, several pupillography frequency band combinations were tested to find the LF and HF bands that maximized the correlation with the HRV LF and HF bands. Following this procedure, we were capable of selecting adequate LF and HF band limits for the feature extraction: the LF band from 0.13Hz to 0.28Hz and the HF band from 0.28Hz to 0.35Hz. The features extracted from these bands were associated with the corresponding code section, and statistical transformations of these features were computed. An (SVM) classifier was trained using these transformed features, achieving a 0.76 F1-Score mean value, very similar to the HRV based classifiers' performances. In fact, the pupillography-based classifier obtained the best performance in the overall study, having the highest mean F1-Score with the lowest variability, and the pupillography features have proven to possess the highest code complexity discriminative ability among the features in the present study. Indicating that it could be possible to achieve an entire non-intrusive method using pupillography features for code complexity classification.

# Bibliography

[1] I. Sandu, A. Salceanu, and O. Bejenaru, "New approach of the customer defects per lines of code metric in automotive sw development applications," in *Journal of Physics: Conference Series*, vol. 1065, p. 052006, IOP Publishing, 2018.

[2] R. Minelli, A. Mocci, and M. Lanza, "I know what you did last summer-an investigation of how developers spend their time," in *2015 IEEE 23rd International Conference on Program Comprehension*, pp. 25–35, IEEE, 2015.

[3] S. Matteson, "TechRepublic report: Software failure caused $1.7 trillion in financial losses in 2017." `https://www.techrepublic.com/article/report-software-failure-caused-1-7-trillion-in-financial-losses-in-2017/`, 2018. Last Accessed: 2022-07-07.

[4] G. K. Saha, "Software fault avoidance issues," *Ubiquity*, vol. 2006, no. November, pp. 1–15, 2006.

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

[6] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[7] 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, 2011.

[8] A. Bernardes, R. Couceiro, J. Medeiros, J. Henriques, C. Teixeira, J. Durães, H. Madeira, and P. Carvalho, "Impact of ultra-short-term hrv features in software code sections complexity classification," in *2022 IEEE 21st Mediterranean*

*Electrotechnical Conference (MELECON)*, pp. 579–584, IEEE, 2022.

[9] A. Bernardes, R. Couceiro, J. Medeiros, J. Henriques, C. Teixeira, M. Simões, J. Durães, R. Barbosa, H. Madeira, and P. Carvalho, "How reliable are ultra-short-term hrv measurements during cognitively demanding tasks?," *Sensors*, vol. 22, no. 17, p. 6528, 2022.

[10] F. F. Evans-martin and D. A. Cooley, "Organization of the nervous system," in *Your Body. How It Works. The Nervous System*, ch. 3, pp. 31–51, Chelsea House Publishers, 2005.

[11] J. E. Hall and M. E. Hall, *Guyton and Hall textbook of medical physiology e-Book*. Elsevier Health Sciences, 2020.

[12] Heart and S. F. of Canada, "An illustration of the human brain with several regions labeled." `https://www.researchgate.net/figure/1-An-illustration-of-the-human-brain-with-several-regions-labeled-Source-Heart-and_fig5_291346848`, 2007. Last Accessed: 2022-10-07.

[13] D. A. Mandal and A. Cashin-Garbutt, "Medical Life Sciences, what is the nervous system?." `https://www.news-medical.net/health/What-is-the-Nervous-System.aspx`, 2022. Last Accessed: 2022-10-07.

[14] L. K. McCorry, "Physiology of the autonomic nervous system," *American journal of pharmaceutical education*, vol. 71, no. 4, 2007.

[15] 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`, 2021. Last Accessed: 2022-10-07.

[16] S. Betti, R. M. Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. Del Ry, and F. Cavallo, "Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1748–1758, 2017.

[17] J. He, K. Li, X. Liao, P. Zhang, and N. Jiang, "Real-time detection of acute cognitive stress using a convolutional neural network from electrocardiographic signal," *IEEE Access*, vol. 7, pp. 42710–42717, 2019.

[18] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, “Discriminating stress from cognitive load using a wearable eda device,” *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2009.

[19] D. McDuff, S. Gontarek, and R. Picard, “Remote measurement of cognitive stress via heart rate variability,” in *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, pp. 2957–2960, IEEE, 2014.

[20] P. A. Kirschner, F. Kirschner, and F. Paas, “Cognitive load theory,” 2009.

[21] S. Whittemore and D. A. Cooley, “Overview of the human circulatory system,” in *Your Body. How It Works. The Circulatory System*, ch. 2, pp. 16–21, Chelsea House Publishers, 2004.

[22] B. T. E. of Encyclopaedia, “”heart”. Encyclopedia Britannica.” `https://www.britannica.com/science/heart`, 2021. Last Accessed: 2022-10-07.

[23] M. AlGhatrif and J. Lindsay, “A brief review: history to understand fundamentals of electrocardiography,” *Journal of community hospital internal medicine perspectives*, vol. 2, no. 1, p. 14383, 2012.

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

[25] M. A. Serhani, H. T. El Kassabi, H. Ismail, and A. Nujum Navaz, “Ecg monitoring systems: Review, architecture, processes, and key challenges,” *Sensors*, vol. 20, no. 6, p. 1796, 2020.

[26] S. Whittemore and D. A. Cooley, “Pumping blood: How the heart works,” in *Your Body. How It Works. The Circulatory System*, ch. 6, pp. 62–73, Chelsea House Publishers, 2004.

[27] I. Tawakal, E. Suryana, A. Noviyanto, I. P. Satwika, S. Alvissalim, I. Hermawan, S. M. Isa, and W. Jatmiko, “Analysis of multi codebook glvq versus standard glvq in discriminating sleep stages,” pp. 197–202, 01 2012.

[28] J. Pumprla, K. Howorka, D. Groves, M. Chester, and J. Nolan, “Functional assessment of heart rate variability: physiological basis and practical applications,” *International journal of cardiology*, vol. 84, no. 1, pp. 1–14, 2002.

[29] H. ChuDuc, K. NguyenPhan, and D. NguyenViet, "A review of heart rate variability and its applications," *APCBEE procedia*, vol. 7, pp. 80–85, 2013.

[30] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[31] S. Sammito and I. Böckelmann, "Factors influencing heart rate variability," in *International Cardiovascular Forum Journal*, vol. 6, 2016.

[32] J. Choi and R. Gutierrez-Osuna, "Removal of respiratory influences from heart rate variability in stress monitoring," *IEEE Sensors Journal*, vol. 11, no. 11, pp. 2649–2656, 2011.

[33] J. S. Gasior, J. Sacha, P. J. Jeleń, J. Zieliński, and J. Przybylski, "Heart rate and respiratory rate influence on heart rate variability repeatability: effects of the correction for the prevailing heart rate," *Frontiers in physiology*, vol. 7, p. 356, 2016.

[34] A. Stys and T. Stys, "Current clinical applications of heart rate variability," *Clinical cardiology*, vol. 21, no. 10, pp. 719–724, 1998.

[35] L. Pecchia, R. Castaldo, L. Montesinos, and P. Melillo, "Are ultra-short heart rate variability features good surrogates of short-term ones? state-of-the-art review and recommendations," *Healthcare technology letters*, vol. 5, no. 3, pp. 94–100, 2018.

[36] A. Boardman, F. S. Schlindwein, and A. P. Rocha, "A study on the optimum order of autoregressive models for heart rate variability," *Physiological measurement*, vol. 23, no. 2, p. 325, 2002.

[37] G. Forte, F. Favieri, and M. Casagrande, "Heart rate variability and cognitive function: A systematic review," *Frontiers in neuroscience*, vol. 13, p. 710, 2019.

[38] F. Shaffer, Z. M. Meehan, and C. L. Zerr, "A critical review of ultra-short-term heart rate variability norms research," *Frontiers in neuroscience*, vol. 14, p. 594880, 2020.

[39] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," in *2007 29th annual international conference of the ieee engineering in medicine and biology society*, pp. 4656–4659, IEEE, 2007.

[40] E. S. Perkins and H. Davson, “”human eye”. Encyclopedia Britannica.” `https://www.britannica.com/science/human-eye`, 2021. Last Accessed: 2022-11-07.

[41] S. Sirois and J. Brisson, “Pupillometry,” *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 6, pp. 679–692, 2014.

[42] N. E. Institute, “How the eyes work.” `https://www.nei.nih.gov/learn-about-eye-health/healthy-vision/how-eyes-work`, 2022. Last Accessed: 2022-11-07.

[43] M. Płużyczka, “The first hundred years: A history of eye tracking as a research method,” *Applied Linguistics Papers*, no. 25/4, pp. 101–116, 2018.

[44] R. J. Jacob and K. S. Karn, “Commentary on section 4. eye tracking in human-computer interaction and usability research: Ready to deliver the promises,” *The mind’s eye*, vol. 2, no. 3, pp. 573–605, 2003.

[45] Z. Sharafi, Z. Soh, and Y.-G. Guéhéneuc, “A systematic literature review on the usage of eye-tracking in software engineering,” *Information and Software Technology*, vol. 67, pp. 79–107, 2015.

[46] A. Kar and P. Corcoran, “A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms,” *IEEE Access*, vol. 5, pp. 16495–16519, 2017.

[47] A. Murata and H. Iwase, “Evaluation of mental workload by fluctuation analysis of pupil area,” in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286)*, vol. 6, pp. 3094–3097, IEEE, 1998.

[48] P. van der Wel and H. van Steenbergen, “Pupil dilation as an index of effort in cognitive control tasks: A review,” *Psychonomic bulletin & review*, vol. 25, no. 6, pp. 2005–2015, 2018.

[49] 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.

[50] K. Sahu and R. Srivastava, “Revisiting software reliability,” *Data Management,*

*Analytics and Innovation*, pp. 221–235, 2019.

[51] 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.

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

[53] 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.

[54] R. Castaldo, W. Xu, P. Melillo, L. Pecchia, L. Santamaria, and C. James, "Detection of mental stress due to oral academic examination via ultra-short-term hrv analysis," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3805–3808, IEEE, 2016.

[55] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, 2019.

[56] F. Mokhayeri, M. Akbarzadeh-T, and S. Toosizadeh, "Mental stress detection using physiological signals based on soft computing techniques," in *2011 18th Iranian Conference of Biomedical Engineering (ICBME)*, pp. 232–237, IEEE, 2011.

[57] T. F. o. t. E. S. o. C. t. N. A. S. o. P. Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[58] M. Zubair and C. Yoon, "Multilevel mental stress detection using ultra-short pulse rate variability series," *Biomedical Signal Processing and Control*, vol. 57, p. 101736, 2020.

[59] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term hrv features as surrogates of short term hrv: A case study on mental stress

detection in real life," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–13, 2019.

[60] M. Nardelli, A. Greco, J. Bolea, G. Valenza, E. P. Scilingo, and R. Bailon, "Reliability of lagged poincaré plot parameters in ultrashort heart rate variability series: Application on affective sounds," *IEEE journal of biomedical and health informatics*, vol. 22, no. 3, pp. 741–749, 2017.

[61] F. Landreani, A. Faini, A. Martin-Yebra, M. Morri, G. Parati, and E. G. Caiani, "Assessment of ultra-short heart variability indices derived by smartphone accelerometers for stress detection," *Sensors*, vol. 19, no. 17, p. 3729, 2019.

[62] B. Weber, T. Fischer, and R. Riedl, "Brain and autonomic nervous system activity measurement in software engineering: A systematic literature review," *Journal of Systems and Software*, vol. 178, p. 110946, 2021.

[63] H. J. Baek, C.-H. Cho, J. Cho, and J.-M. Woo, "Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability," *Telemedicine and e-Health*, vol. 21, no. 5, pp. 404–414, 2015.

[64] K. Li, H. Rüdiger, and T. Ziemssen, "Spectral analysis of heart rate variability: time window matters," *Frontiers in neurology*, vol. 10, p. 545, 2019.

[65] L. Salahuddin, M. G. Jeong, and D. Kim, "Ultra short term analysis of heart rate variability using normal sinus rhythm and atrial fibrillation ecg data," in *2007 9th International Conference on e-Health Networking, Application and Services*, pp. 240–243, IEEE, 2007.

[66] U. Nussinovitch, K. P. Elishkevitz, K. Katz, M. Nussinovitch, S. Segev, B. Volovitz, and N. Nussinovitch, "Reliability of ultra-short ecg indices for heart rate variability," *Annals of Noninvasive Electrocardiology*, vol. 16, no. 2, pp. 117–122, 2011.

[67] J. McNames and M. Aboy, "Reliability and accuracy of heart rate variability metrics versus ecg segment duration," *Medical and Biological Engineering and Computing*, vol. 44, no. 9, pp. 747–756, 2006.

[68] H. Wilhelm and B. Wilhelm, "Clinical applications of pupillography," *Journal of Neuro-ophthalmology*, vol. 23, no. 1, pp. 42–49, 2003.

[69] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during

simple problem-solving," *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964.

[70] M. Pedrotti, M. A. Mirzaei, A. Tedesco, J.-R. Chardonnet, F. Mérienne, S. Benedetto, and T. Baccino, "Automatic stress classification with pupil diameter analysis," *International Journal of Human-Computer Interaction*, vol. 30, no. 3, pp. 220–236, 2014.

[71] H. Lüdtke, B. Wilhelm, M. Adler, F. Schaeffel, and H. Wilhelm, "Mathematical procedures in data recording and processing of pupillary fatigue waves," *Vision research*, vol. 38, no. 19, pp. 2889–2896, 1998.

[72] M. Nakayama and Y. Shimizu, "Frequency analysis of task evoked pupillary response and eye-movement," in *Proceedings of the 2004 symposium on Eye tracking research & applications*, pp. 71–76, 2004.

[73] V. Peysakhovich, M. Causse, S. Scannella, and F. Dehais, "Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort," *International Journal of Psychophysiology*, vol. 97, no. 1, pp. 30–37, 2015.

[74] J. Lee, J. Kim, K. Park, and G. Khang, "Evaluation of the methods for pupil size estimation: on the perspective of autonomic activity," in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, pp. 1501–1504, IEEE, 2004.

[75] H. Rehatschek and G. Kienast, "Vizard-an innovative tool for video navigation, retrieval, annotation and editing," in *Proceedings of the 23rd Workshop of PVA: Multimedia and Middleware*, p. 11, 2001.

[76] B. Stemmer and J. F. Connolly, "The eeg/erp technologies in linguistic research: An essay on the advantages they offer and a survey of their purveyors," *The Mental Lexicon*, vol. 6, no. 1, pp. 141–170, 2011.

[77] S. Research, "Eyelink 1000 plus." `https://www.sr-research.com/eyelink-1000-plus/`, 2022. Last Accessed: 2022-04-08.

[78] R. K. Niazy, C. F. Beckmann, G. D. Iannetti, J. M. Brady, and S. M. Smith, "Removal of fmri environment artifacts from eeg data using optimal basis sets," *Neuroimage*, vol. 28, no. 3, pp. 720–737, 2005.

[79] I. I. Christov, "Real time electrocardiogram qrs detection using combined adap-

tive threshold," *Biomedical engineering online*, vol. 3, no. 1, pp. 1–9, 2004.

[80] Kubios, "HRV analysis methods." `https://www.kubios.com/hrv-analysis-methods/`, 2022. Last Accessed: 2022-11-08.

[81] M. Yılmaz, H. Kayançiçek, and Y. Çekici, "Heart rate variability: highlights from hidden signals," *J. Integr. Cardiol*, vol. 4, pp. 1–8, 2018.

[82] M. Vollmer, "A robust, simple and reliable measure of heart rate variability using relative rr intervals," in *2015 Computing in Cardiology Conference (CinC)*, pp. 609–612, IEEE, 2015.

[83] T. K. Sahoo, A. Mahapatra, and N. Ruban, "Stress index calculation and analysis based on heart rate variability of ecg signal with arrhythmia," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, vol. 1, pp. 1–7, IEEE, 2019.

[84] P. Melillo, M. Bracale, and L. Pecchia, "Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination," *Biomedical engineering online*, vol. 10, no. 1, pp. 1–13, 2011.

[85] M. J. Katz, "Fractals and the analysis of waveforms," *Computers in biology and medicine*, vol. 18, no. 3, pp. 145–156, 1988.

[86] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, 1988.

[87] R. S. Gomolka, S. Kampusch, E. Kaniusas, F. Thürk, J. C. Széles, and W. Klonowski, "Higuchi fractal dimension of heart rate variability during percutaneous auricular vagus nerve stimulation in healthy and diabetic subjects," *Frontiers in physiology*, vol. 9, p. 1162, 2018.

[88] N. C. Silver and W. P. Dunlap, "Averaging correlation coefficients: should fisher's z transformation be used?," *Journal of applied psychology*, vol. 72, no. 1, p. 146, 1987.

[89] Y. Sasaki *et al.*, "The truth of the f-measure," *Teach tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[90] R. Couceiro, G. Duarte, J. Durães, J. Castelhano, C. Duarte, C. Teixeira, M. C. Branco, P. Carvalho, and H. Madeira, "Pupillography as indicator of programmers' mental effort and cognitive overload," in *2019 49th Annual*

*IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 638–644, IEEE, 2019.

[91] O. Salem, Y. Liu, and A. Mehaoua, “A lightweight anomaly detection framework for medical wireless sensor networks,” in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 4358–4363, IEEE, 2013.

[92] D. Kondrashov and M. Ghil, “Spatio-temporal filling of missing points in geophysical data sets,” *Nonlinear Processes in Geophysics*, vol. 13, no. 2, pp. 151–159, 2006.

[93] R. Sassi, V. D. Corino, and L. T. Mainardi, “Analysis of surface atrial signals: time series with missing data?,” *Annals of biomedical engineering*, vol. 37, no. 10, pp. 2082–2092, 2009.

[94] F. Onorati, M. Mauri, V. Russo, and L. Mainardi, “Reconstruction of pupil dilation signal during eye blinking events,” in *Proceeding of the 7th International Workshop on Biosignal Interpretation*, pp. 117–120, 2012.

[95] A. Eleuteri, A. C. Fisher, D. Groves, and C. J. Dewhurst, “An efficient time-varying filter for detrending and bandwidth limiting the heart rate variability tachogram without resampling: Matlab open-source code and internet web-based implementation,” *Computational and mathematical methods in medicine*, vol. 2012, 2012.

# Appendices

# A

# Wilcoxon Rank Sum Test - Linear Regressions of the Statistical Percentages (HRV)

**Table A.1:** Wilcoxon Rank Sum Test - Linear Regressions of the Statistical Percentages (Time Domain).

| | Time Domain | | | | | |
|---|---|---|---|---|---|---|
| | mNN | SDNN | SDSD | RMSSD | NN50 | pNN50 |
| A | -2.84 | -5.7 | -3.98 | -3.99 | -1.75 | -1.32 |
| B | 98.01 | 85.13 | 94.14 | 94.35 | 20.47 | 95.87 |
| RSQ | 0.97 | 0.94 | 0.94 | 0.94 | 0.16 | 0.50 |

**Table A.2:** Wilcoxon Rank Sum Test - Linear Regressions of the Statistical Percentages (Non-Linear Domain).

| | Non-Linear Domain | | | | | |
|---|---|---|---|---|---|---|
| | ApEn | SD1 | SD2 | KFD | HFD | PTM |
| A | -5.09 | -5.70 | -5.70 | -1.85 | -7.60 | -3.04 |
| B | 77.78 | 85.13 | 85.13 | 21.66 | 107.94 | 94.66 |
| RSQ | 0.87 | 0.94 | 0.94 | 0.18 | 0.85 | 0.93 |

**Table A.3:** Wilcoxon Rank Sum Test - Linear Regressions of the Statistical Percentages (Geometrical Domain).

| | Geometrical Domain | | |
|---|---|---|---|
| | SI | TI | TINN |
| A | -3.95 | -4.89 | -4.97 |
| B | 48.55 | 62.82 | 65.15 |
| RSQ | 0.51 | 0.70 | 0.73 |

**Table A.4:** Wilcoxon Rank Sum Test - Linear Regressions of the Statistical Percentages (Frequency Domain - part 1).

| | Frequency Domain | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | totPow | Peak | VLF | LF | HF | VLFpeak | LFpeak | HFpeak |
| A | -6.10 | -6.14 | -5.41 | -6.24 | -6.24 | -5.34 | -5.69 | -5.29 |
| B | 80.83 | 92.0 | 71.42 | 101.31 | 97.86 | 70.92 | 91.64 | 99.79 |
| RSQ | 0.80 | 0.93 | 0.76 | 0.98 | 0.95 | 0.78 | 0.93 | 0.96 |

**Table A.5:** Wilcoxon Rank Sum Test - Linear Regressions of the Statistical Percentages (Frequency Domain - part 2).

| | Frequency Domain | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | VLFnu | LFnu | HFnu | VLFpeak-nu | LFpeak-nu | HFpeak-nu | LF/HF | LFpeak/HFpeak |
| A | -5.66 | -5.37 | -5.76 | -5.76 | -5.07 | -5.14 | -4.71 | -4.71 |
| B | 77.63 | 86.31 | 83.59 | 79.61 | 92.25 | 92.17 | 90.37 | 89.9 |
| RSQ | 0.84 | 0.87 | 0.87 | 0.85 | 0.89 | 0.94 | 0.85 | 0.89 |

# B

# Spearman's Correlation Test - Linear Regressions of the Mean Correlations (HRV)

**Table B.1:** Spearman's Correlation Test - Linear Regressions of the Mean Correlations (Time Domain).

**Time Domain**

| | mNN | SDNN | SDSD | RMSSD | NN50 | pNN50 |
|---|---|---|---|---|---|---|
| A | -0.033 | -0.043 | -0.037 | -0.037 | -0.038 | -0.038 |
| B | 0.979 | 0.936 | 0.909 | 0.909 | 0.772 | 0.772 |
| RSQ | 0.99 | 0.98 | 0.97 | 0.97 | 0.90 | 0.90 |

**Table B.2:** Spearman's Correlation Test - Linear Regressions of the Mean Correlations (Non-Linear Domain).

| | **Non-Linear Domain** | | | | | |
|---|---|---|---|---|---|---|
| | ApEn | SD1 | SD2 | KFD | HFD | PTM |
| A | -0.042 | -0.043 | -0.043 | -0.041 | -0.043 | -0.037 |
| B | 0.888 | 0.936 | 0.936 | 0.917 | 0.934 | 0.902 |
| RSQ | 0.96 | 0.98 | 0.98 | 0.97 | 0.97 | 0.96 |

**Table B.3:** Spearman's Correlation Test - Linear Regressions of the Mean Correlations (Geometrical Domain).

| | **Geometrical Domain** | | |
|---|---|---|---|
| | SI | TI | TINN |
| A | -0.043 | -0.045 | -0.041 |
| B | 0.905 | 0.906 | 0.823 |
| RSQ | 0.97 | 0.97 | 0.93 |

**Table B.4:** Spearman's Correlation Test - Linear Regressions of the Mean Correlations (Frequency Domain - part 1).

| | **Frequency Domain** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | totPow | Peak | VLF | LF | HF | VLFpeak | LFpeak | HFpeak |
| A | -0.04 | -0.043 | -0.052 | -0.039 | -0.038 | -0.052 | -0.04 | -0.043 |
| B | 1.094 | 1.066 | 1.112 | 1.091 | 1.099 | 1.105 | 1.064 | 1.065 |
| RSQ | 0.91 | 0.97 | 0.94 | 0.95 | 0.9 | 0.94 | 0.97 | 0.97 |

**Table B.5:** Spearman's Correlation Test - Linear Regressions of the Mean Correlations (Frequency Domain - part 2).

| | **Frequency Domain** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | VLFnu | LFnu | HFnu | VLFpeak-nu | LFpeak-nu | HFpeak-nu | LF/HF | LFpeak/HFpeak |
| A | -0.054 | -0.052 | -0.046 | -0.055 | -0.049 | -0.048 | -0.046 | -0.048 |
| B | 1.11 | 1.103 | 1.099 | 1.048 | 0.88 | 1.031 | 1.101 | 1.051 |
| RSQ | 0.95 | 0.97 | 0.94 | 0.99 | 0.98 | 0.98 | 0.96 | 0.99 |

# C

# 5 most discriminative transformed HRV features by time frame (HRV)

**Table C.1:** 5 most discriminative transformed HRV features by time frame (selected using the Kruskal Wallis test at each Nested Leave-One-Subject-Out inner iteration).

**180 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 100.0 | 1.0 |
| mNN | peaks rate | 82.42 | 3.2 |
| HFD | peaks rate | 78.02 | 3.2 |
| LF/HF | peaks rate | 76.92 | 2.9 |
| TI | peaks rate | 56.04 | 4.0 |

**170 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 93.41 | 2.2 |
| HFnu | peaks rate | 86.81 | 2.2 |
| LF/HF | peaks rate | 78.02 | 3.2 |
| LFnu | peaks rate | 65.93 | 3.1 |
| HFpeak-nu | peaks rate | 57.14 | 3.5 |

**160 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| LF/HF | peaks rate | 98.90 | 1.6 |
| HF | peaks rate | 95.60 | 2.3 |
| TI | peaks rate | 82.42 | 3.4 |
| LFpeak/HFpeak | peaks rate | 45.05 | 3.9 |
| totPow | peaks rate | 43.96 | 3.7 |

**150 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 100.0 | 1.0 |
| mNN | peaks rate | 100.0 | 2.5 |
| HFnu | peaks rate | 80.00 | 3.4 |
| HFpeak | peaks rate | 75.24 | 3.4 |
| Peak | peaks rate | 45.71 | 4.1 |

**140 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| LF/HF | peaks rate | 100.0 | 1.3 |
| HF | peaks rate | 100.0 | 1.7 |
| totPow | peaks rate | 59.05 | 4.0 |
| LFpeak | peaks rate | 49.52 | 3.7 |
| VLF | peaks rate | 46.67 | 3.9 |

**130 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| totPow | peaks rate | 100.0 | 1.2 |
| HF | peaks rate | 100.0 | 1.8 |
| HFpeak | peaks rate | 93.33 | 3.5 |
| LF | peaks rate | 72.38 | 3.7 |
| HFpeak-nu | peaks rate | 29.52 | 4.6 |

**120 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 98.10 | 1.8 |
| LF/HF | peaks rate | 93.33 | 1.6 |
| VLFnu | peaks rate | 66.67 | 3.6 |
| HFD | peaks rate | 56.19 | 4.1 |
| Peak | peaks rate | 53.33 | 3.6 |

**110 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| LF/HF | peaks rate | 94.29 | 1.6 |
| HF | peaks rate | 78.10 | 2.8 |
| totPow | peaks rate | 67.62 | 2.9 |
| VLF | peaks rate | 67.62 | 3.1 |
| LFpeak | peaks rate | 67.62 | 3.5 |

**100 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| VLFpeak | peaks rate | 99.05 | 2.3 |
| totPow | peaks rate | 88.57 | 2.2 |
| VLFnu | peaks rate | 83.81 | 3.0 |
| TI | peaks rate | 79.05 | 3.3 |
| HF | peaks rate | 60.95 | 3.3 |

**90 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 100.0 | 1.1 |
| TI | peaks rate | 96.19 | 3.4 |
| HFnu | peaks rate | 84.76 | 3.3 |
| VLFpeak | peaks rate | 77.14 | 3.2 |
| VLF | peaks rate | 75.24 | 3.7 |

**80 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 99.05 | 1.2 |
| VLF | peaks rate | 80.95 | 3.2 |
| HFnu | peaks rate | 74.29 | 3.4 |
| VLFpeak | peaks rate | 69.52 | 3.3 |
| LF/HF | peaks rate | 68.57 | 3.0 |

**70 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| VLFnu | peaks rate | 100.0 | 1.0 |
| VLFpeak | peaks rate | 100.0 | 2.4 |
| LF/HF | peaks rate | 85.71 | 3.3 |
| HFnu | peaks rate | 73.33 | 3.7 |
| VLF | peaks rate | 73.33 | 4.1 |

**60 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HF | peaks rate | 100.0 | 1.0 |
| LF/HF | peaks rate | 91.43 | 2.4 |
| LFnu | peaks rate | 85.71 | 3.3 |
| HFnu | peaks rate | 72.38 | 3.8 |
| LF | peaks rate | 28.57 | 4.6 |

**50 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| LFnu | peaks rate | 100.0 | 1.5 |
| HFnu | peaks rate | 85.71 | 2.1 |
| totPow | peaks rate | 77.14 | 3.2 |
| HFpeak | peaks rate | 66.67 | 3.8 |
| VLF | peaks rate | 61.90 | 3.7 |

**40 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HFnu | peaks rate | 100.0 | 1.6 |
| totPow | peaks rate | 100.0 | 2.1 |
| LF/HF | peaks rate | 100.0 | 2.4 |
| LFnu | peaks rate | 57.14 | 4.4 |
| HF | peaks rate | 56.19 | 4.1 |

**30 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| SDNN | peaks rate | 100.0 | 1.2 |
| SD1 | peaks rate | 100.0 | 2.2 |
| SD2 | peaks rate | 98.10 | 3.2 |
| LF | peaks rate | 43.81 | 4.3 |
| HFnu | peaks rate | 37.14 | 4.1 |

**20 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| totPow | peaks min | 89.52 | 2.1 |
| VLFnu | peaks min | 85.71 | 2.1 |
| HF | peaks rate | 85.71 | 2.5 |
| LFpeak | peaks rate | 44.76 | 4.0 |
| VLFpeak-nu | peaks min | 37.14 | 3.5 |

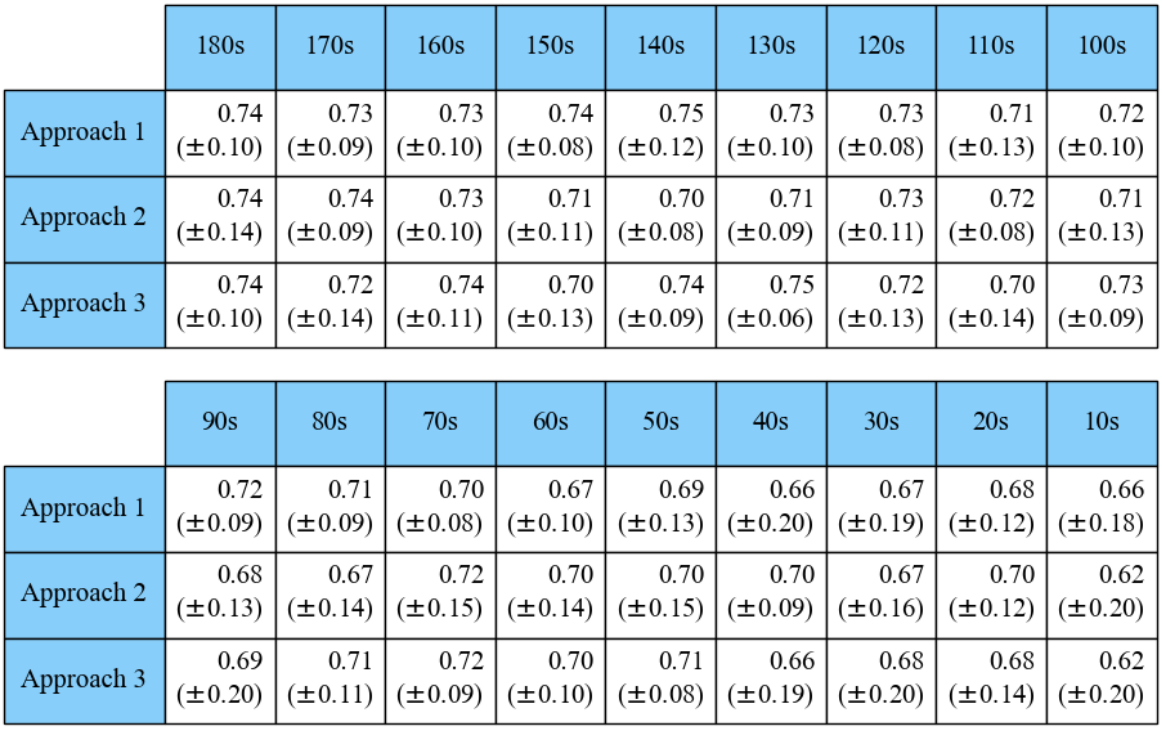
**10 seconds Time Frame**

| HRV Feature | Transformation | Occurrence (%) | Top 5<br>Mean Position |
|---|---|---|---|
| HFpeak | peaks min | 94.29 | 2.0 |
| HFpeak-nu | min | 75.24 | 2.3 |
| KFD | peaks rate | 61.90 | 3.0 |
| HFD | min | 60.00 | 3.1 |
| ApEn | peaks min | 54.29 | 2.9 |

# D

# Mean and Standard Deviation of the F1-Scores values obtained in the classification process (HRV)

**Table D.1:** Mean and Standard Deviation of the F1-Scores values obtained using the datasets based on the different time frames and approaches.



| | 180s | 170s | 160s | 150s | 140s | 130s | 120s | 110s | 100s |
|---|---|---|---|---|---|---|---|---|---|
| Approach 1 | 0.74<br>(±0.10) | 0.73<br>(±0.09) | 0.73<br>(±0.10) | 0.74<br>(±0.08) | 0.75<br>(±0.12) | 0.73<br>(±0.10) | 0.73<br>(±0.08) | 0.71<br>(±0.13) | 0.72<br>(±0.10) |
| Approach 2 | 0.74<br>(±0.14) | 0.74<br>(±0.09) | 0.73<br>(±0.10) | 0.71<br>(±0.11) | 0.70<br>(±0.08) | 0.71<br>(±0.09) | 0.73<br>(±0.11) | 0.72<br>(±0.08) | 0.71<br>(±0.13) |
| Approach 3 | 0.74<br>(±0.10) | 0.72<br>(±0.14) | 0.74<br>(±0.11) | 0.70<br>(±0.13) | 0.74<br>(±0.09) | 0.75<br>(±0.06) | 0.72<br>(±0.13) | 0.70<br>(±0.14) | 0.73<br>(±0.09) |

| | 90s | 80s | 70s | 60s | 50s | 40s | 30s | 20s | 10s |
|---|---|---|---|---|---|---|---|---|---|
| Approach 1 | 0.72<br>(±0.09) | 0.71<br>(±0.09) | 0.70<br>(±0.08) | 0.67<br>(±0.10) | 0.69<br>(±0.13) | 0.66<br>(±0.20) | 0.67<br>(±0.19) | 0.68<br>(±0.12) | 0.66<br>(±0.18) |
| Approach 2 | 0.68<br>(±0.13) | 0.67<br>(±0.14) | 0.72<br>(±0.15) | 0.70<br>(±0.14) | 0.70<br>(±0.15) | 0.70<br>(±0.09) | 0.67<br>(±0.16) | 0.70<br>(±0.12) | 0.62<br>(±0.20) |
| Approach 3 | 0.69<br>(±0.20) | 0.71<br>(±0.11) | 0.72<br>(±0.09) | 0.70<br>(±0.10) | 0.71<br>(±0.08) | 0.66<br>(±0.19) | 0.68<br>(±0.20) | 0.68<br>(±0.14) | 0.62<br>(±0.20) |

# E

# Scientific Articles produced during the present thesis

sensors

*Article*

# How Reliable Are Ultra-Short-Term HRV Measurements during Cognitively Demanding Tasks?

**André Bernardes *, Ricardo Couceiro, Júlio Medeiros, Jorge Henriques, César Teixeira, Marco Simões, João Durães, Raul Barbosa, Henrique Madeira and Paulo Carvalho**

Centre for Informatics and Systems of the University of Coimbra (CISUC), 3030-290 Coimbra, Portugal
* Correspondence: ambernardes11@gmail.com

**Abstract:** Ultra-short-term HRV features assess minor autonomous nervous system variations such as variations resulting from cognitive stress peaks during demanding tasks. Several studies compare ultra-short-term and short-term HRV measurements to investigate their reliability. However, existing experiments are conducted in low cognitively demanding environments. In this paper, we propose to evaluate these measurements' reliability under cognitively demanding tasks using a near real-life setting. For this purpose, we selected 31 HRV features, extracted from data collected from 21 programmers performing code comprehension, and compared them across 18 different time frames, ranging from 3 min to 10 s. Statistical significance and correlation tests were performed between the features extracted using the larger window (3 min) and the same features extracted with the other 17 time frames. We paired these analyses with Bland–Altman plots to inspect how the extraction window size affects the HRV features. The main results show 13 features that presented at least 50% correlation when using 60-second windows. The HF and mNN features achieved around 50% correlation using a 30-second window. The 30-second window was the smallest time frame considered to have reliable measurements. Furthermore, the mNN feature proved to be quite robust to the shortening of the time resolution.

**Keywords:** ultra-short-term HRV features; statistical significance; correlation; cognitively demanding tasks; code comprehension

**Citation:** Bernardes, A.; Couceiro, R.; Medeiros, J.; Henriques, J.; Teixeira, C.; Simões, M.; Durães, J.; Barbosa, R.; Madeira, H.; Carvalho, P. How Reliable Are Ultra-Short-Term HRV Measurements during Cognitively Demanding Tasks? *Sensors* **2022**, *22*, 6528. https://doi.org/10.3390/s22176528

Academic Editors: Andrea Cataldo and Maysam Abbod

Received: 22 June 2022
Accepted: 24 August 2022
Published: 30 August 2022

## 1. Introduction

Heart rate (HR) is defined as "the number of heartbeats per minute" [1], and does not provide direct information about autonomic nervous system (ANS) dynamics since it is a static index of autonomic input to the sinoatrial node [2]. On the other hand, heart rate variability (HRV) is described as "the fluctuation in the time intervals between adjacent heartbeats" (RR intervals or NN intervals) [1]. RR intervals are measured in milliseconds (ms) and "result mostly from the dynamic interaction between the parasympathetic and the sympathetic inputs to the heart through the sinoatrial node" [3]. Unlike HR, HRV analysis is useful for providing a "quantitative assessment of cardiac autonomic regulation" [2].

HRV is often used as a non-invasive marker of ANS activity, and its spectrum analysis can measure the sympathovagal balance [4]. Thereby, many studies point to the potential of HRV for diagnosis and prognosis of health problems [5] and other areas such as the measurement of a subject's cognitive load [6–8].

From the physiological point of view, our nervous system has two major subdivisions: the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS is formed by the spinal cord and the brain and is responsible for receiving signals from different body components, processing these signals and producing responses as new signals to be delivered across these body components. The PNS is the subdivision accountable for carrying messages exchanged between the CNS and the organs, glands, muscles and senses [9]. Belonging to the PNS, we have a further two different constituents: the somatic nervous system (SNS), which is the constituent responsible (essentially) for voluntary and conscious actions, and the autonomic nervous system (ANS), where our study focuses. The ANS has two subdivisions: the sympathetic and the parasympathetic nervous systems. The sympathetic nervous system is more associated with stressful situations that need an emergency response, the so-called fight-or-flight mode.

In contrast, parasympathetic nervous system activity is associated with more conservative and restoring processes, bringing the body back to a stable state [6]. In this way, the CNS has the power to influence every other system in the body, which, in theory, makes it possible to access data, such as the cognitive load of the brain through physiological signals controlled by the ANS, that might be captured using HRV. In fact, HRV is considered an index of autonomic control of the heart and has been pointed out as having a good physiological correlation with cognitive functioning [3]. However, despite several studies in the field, this is still a controversial subject across the scientific community.

Taking a leading role in the regulation of cardiac function, the ANS controls the constriction and relaxation of blood vessels, which allows it to regulate blood pressure and, in this way, is capable of adjusting heart rate and heart contractility. Using an ECG signal, it is possible to compute HRV through time fluctuations of the intervals between the consecutive R-peaks. Then, from the HRV measurement, different features can be extracted in time, geometric, non-linear and frequency domains, which can be used to access the different ANS sympathetic and parasympathetic systems' dynamics. Across different papers that approach HRV analysis, the most common features referenced in the time domain are mNN, SDNN, SDSD, RMSSD, NN50 and pNN50 (see terminology in Table 1). Regarding power spectrum density analysis, the frequency domain is divided into three bands: the very-low-frequency band (VLF: under 0.04 HZ), the low-frequency band (LF: 0.04 to 0.15 HZ) and the high-frequency band (0.15 to 0.4 HZ) [10]. The features extracted from each band most referenced in the literature are the total power and the peak. The ratio between the LF power and the HF power is also frequently mentioned.

From the features previously mentioned, some features have already been linked to physiological dynamics. Starting with the VLF band, this band is mentioned to be a heart's intrinsic nervous system consequence. The SDNN, as mentioned in [10], is influenced by every cyclic component responsible for variability in the recording period. This feature is highly correlated with the LF band, and the two are associated with both the sympathetic and parasympathetic systems' dynamics. The LF band is as well-linked to blood pressure regulation via baroreceptors. The features RMSSD, pNN50 and the HF band are also correlated and are closely influenced by the parasympathetic system. Thus, the ratio between the LF power and the HF power is believed to be a good measure of the balance between the sympathetic and parasympathetic systems. Although this belief is not consensual, and this relationship is not as straightforward as once believed, we can still look at this ratio as a metric of one system's predominance over another [7].

In addition to time and frequency domains, several authors have also pursued the extraction of measures in the non-linear space in order to unveil non-linear HRV patterns. Based on the studies present in the literature, several measurements have been selected, focusing on their consistency when extracted using small time frames (e.g., 5 min) [2,6,11,12], which are: approximate entropy, Poincare's plot parameters (SD1 and SD2), point transition measure, Katz fractal dimension and Higuchi fractal dimension from the non-linear domain, stress index, HRV triangular index and TINN from the geometric domain.

Short-term time frames (5 min in length) are already a standard and are currently well-accepted as suitable for extracting accurate HRV measurements [10]. However, the need to extract HRV measurements using time frames shorter than 1 min (ultra-short HRV features) has grown for several reasons [5,11,13]. Among these are the need to reduce the time spent and costs in the extraction of these indexes, the fact that they are incompatible with the dynamics of the physiological mechanisms to be captured (e.g., cognitive load spikes), or the need to extract these features in new environments using modern wearable devices [13]. In fact, these brought a wide range of new applications that can benefit from the advances in ultra-short-term measurements fields.

The current study is integrated into the Biofeedback Augmented Software Engineering (BASE) project, which aims to develop a solution capable of using biofeedback from the programmer to detect code areas more prone to error and save time in the bug detection process. Code comprehension and bug detection tasks consume up to 70% of the programmers working time, representing millions spent every year trying to avoid these software faults [14]. The BASE project aims to solve this significant socioeconomic problem. The interest in using HRV in software engineering is growing very fast, and applications such as the identification of problematic code areas (that may have bugs and need revision) require very fast response in assessing programmers' cognitive loads using HRV [15]. In order to ensure a real-time response and to detect acute cognitive stress changes, we need time analysis windows as short as possible to achieve the required time resolution, which is the motivation behind the current study.

In order to find out which features are adequate to be used in real-life applications, such as the mentioned code inspection context, and to understand their time frame limitations, this study aims to analyze the stability of 31 existing HRV features. To this end, these features were extracted using time frames of variable sizes (e.g., as low as 10 s) in a code inspection context, an environment very demanding from the intellectual perspective and, therefore, with individuals subjected to high cognitive loads. Our current work investigates the smallest time frame where each feature is reliable, performing an intra-subject and intra-run analysis to avoid biasing the results.

## 2. Background and Related Work

Long-term (recordings lasting 24 h) and short-term (recordings lasting 5 min) HRV measurements are well-documented and conventionally accepted as valid HRV measurements, having multiple clinical applications [10]. However, as mentioned earlier, there is a growing need to use shorter segments. Some studies focus on investigating the reliability of these ultra-short-term measurements when compared to short-term HRV measurements.

In order to evaluate ultra-short-term HRV measurements' reliability as a surrogate of the short-term HRV, different analyses can be performed. A procedure proposed by Pechia et al. [2] included a correlation analysis to test the existence of a significant association between features. If the correlation was significant and the correlation coefficient was above 0.07, the researchers performed a Bland–Altman plot to analyze the degree of bias. In case the data dispersion remained within the 95% line of agreement, the final step was to perform an effect size statistic (Cohen's d statistic to parametric data or Cliff's delta statistic to non-parametric data). The feature was then considered a good surrogate if the effect size statistic test only detected minor differences. The mentioned procedure agrees with Shaffer et al. [5], which recommend using correlation/regression analyses paired with a Bland–Altman plot. Both works agree that only a correlation analysis is not enough to determine if an ultra-short-term HRV feature is a good surrogate of short-term HRV. In fact, the two compared measurements can be highly correlated but still have significantly different values.

A 2017 study carried out by Castaldo et al. [11] used Bland–Altman plots and Spearman's rank correlation analysis to assess which ultra-short-term HRV features are a valid surrogate of short-term HRV. The study also built a machine learning model using ultra-short-term HRV features to discriminate between stress and rest states. The conclusions were that mean HR, the standard deviation of HR, mNN, SDNN, HF and SD2 are appropriate short-term HRV surrogates for mental stress assessment. The paper also highlighted a machine learning model obtained using the mNN, the standard deviation of HR and the HF features, which achieved an accuracy above 88%.

In an article by Salahuddin et al. [8], the authors used mobile-derived ECG recording to extract several HRV measurements and the Kruskal–Wallis test to analyze the reliability of these measurements. It was "assumed that short-term analysis was not significantly different to the 150-second analysis if the $p$-value was greater than 0.05", and the goal was to find until which window span a feature is a good estimative of the 150-second window. The authors concluded that mean RR and RMSSD extracted using 10-second windows were not significantly different from the estimates using 150-second windows. This finding was also confirmed when using 20-second windows for extracting pNN50, HF, LF/HF, LFnu and HFnu features, 30-second windows for LF features and 50-second windows for VLF features. As for the remaining features studied by the authors, a minimum time frame of 60 s was necessary for extracting features that were not significantly different from the 150-second reference features. This study's data were recorded during the subject's day-to-day activities such as regular daily work, study, physical activities and sleep.

In the work of Baek et al. [16], a similar approach has been used to evaluate the reliability of ultra-short-term HRV measurements as short-term (5 min) HRV surrogates. The data were acquired in 5-minute recordings while the subjects were "sitting at rest in a comfortable chair". In order to accomplish the proposed goal, the authors computed the $p$-value by the Kruskal–Wallis test, the Pearson correlation r and Bland–Altman plot analysis comparing 5-minute short-term measurements with ultra-short-term ones with different time frames (270, 240,210, 180, 150, 120, 90, 60, 30, 20 and 10 s). The highlighted features with the best results in this study were the mean HR, where 10-second windows were used to obtain results comparable to the 5-minute analysis, the HF, which required 20-second windows, and the RMSSD, which required 30-second windows.

Following similar approaches, other works, such as the publications by Landreani et al. [13], Li et al. [17], Salahuddin et al. [18], Nussinovitch et al. [19] and McNames et al. [20], were able to converge on a common set of conclusions, where mean HR, mean RR, SDSD, RMSSD, pNN50, HF, LF/HF, LFnu and HFnu were shown to be reliable under the 60-second recordings.

From the reported works, it is possible to conclude that ultra-short-term measurements are far from being consensual. Due to their extraction, only some features keep their stability under small window constraints. Additionally, it is still unclear what the time frame limit is for each HRV feature that can be applied to compute a reliable surrogate of its counterpart extracted from 5-minute recordings. Furthermore, the studies found related to this topic were developed with the subjects at rest or performing elementary tasks in controlled environments. In this work, we aim to elucidate these aspects and validate them under stressful and intellectually demanding environments; more precisely, with the subjects performing software code inspection tasks (i.e., bug detection), which is a highly complex, dynamic, and cognitively demanding task. The main goal of our work is precisely to investigate ultra-short-term HRV features to determine whether HRV-based tools can effectively be used in software development environments. To this extent, our present study investigates the smallest time frame, i.e., the finest time resolution, where each feature is reliable, i.e., the smallest time frame where each feature behavior is still representative of the corresponding 180-second measurement, under our experiment context.

Another relevant aspect that is worth mentioning is that the existing studies perform an inter-subject analysis of the features, i.e., perform the correlation or statistical analysis after concatenating the features collected from different subjects. This fact can lead to biased correlation values since it captures the inter-subject feature tendencies that may overwhelm the actual feature tendencies. In order to avoid this kind of bias, our study performs an intra-subject and intra-run feature analysis.

## 3. Methods

### *3.1. Participants*

The data used in the current work were collected in the scope of the BASE project and aimed at the research of error making and error discovery during software inspection tasks, using functional magnetic resonance imaging (fMRI) and other non-invasive sensors such as the ECG. In order to collect the data used in the study, we opened a call for participation in the experiment. Through this process, we obtained 49 candidates consisting of a mixture of students (pursuing PhDs and MSs in different computer science fields), academic professors and professional specialists in the software sector (code reviewers). The candidates were then interviewed and screened to guarantee their fitment to the study objectives. During the interview, demographic and biometric characteristics (e.g., age), professional status, programming experience, availability and motivation were collected. Subsequently, each candidate's proficiency level was also assessed based on the score provided by two questionnaires: (1) a programming experience questionnaire and (2) a technical questionnaire (see Appendix A). The first questionnaire aimed to assess the candidate's programming experience based on the candidate's coding volume in the last three years. The second questionnaire's goal, composed of 10 questions, was to assess the candidate's coding skills. The programming experience gave us an overall idea of the experience in the past years from the candidate: (1) experience in SW programming (number of years); (2) lines programmed in any language in the last 3 years (approximate number); (3) lines programmed in C in the last 3 years (approximate number); and (4) lines written in the biggest C program written (approximate number). On the other hand, the technical questionnaire was used for candidate characterization regarding present knowledge and coding skills, which is, therefore, more helpful in selecting and classifying the candidates. Based on the results obtained in these questionnaires, the candidates with a score below 3 (out of 10) were considered not eligible since they were not representative of software industry professionals. The remaining ones were characterized as non-experienced (score between 4 and 7) and experienced (score between 8 and 10). In summary, 21 male subjects, ranging from 19 to 40 years, with a median of 22 years, were selected for the experiments after the screening process.

All subjects provided written informed consent, and all the data were anonymized. This study was approved by the Ethical Committee of the Faculty of Medicine of the University of Coimbra, following the Declaration of Helsinki and the standard procedures for studies involving human subjects.

### *3.2. Experimental Protocol and Setup*

The selected candidates were submitted to 4 different runs of code inspection tasks using 4 code snippets written in C code language (selected randomly at each run). Each run started with a fixation cross in the middle of the screen for 30 s. Subsequently, three tasks were presented to the subject: a natural language reading (literary excerpt) task, a neutral (bug-free and straightforward code) code reading task, and one code inspection (code with bugs) task. The order of the presentation was randomly selected to avoid biasing the results, following a randomized control crossover design. Between each task and at the end of each run, a fixation cross was presented to the subject for 30 s. The description of each task is provided as follows:

1. **Natural language reading**: In this task, a text in natural language is presented to the subject (selected randomly from the set of 4 different texts) for 60 s. The presented texts were selected in order to have neutral characteristics and avoid measurement fluctuations induced by narrative-triggered emotions;
2. **Simple code snippet reading**: In this task, the subject is presented with a simple and iterative code snippet (selected randomly from the set of 4 different neutral code snippets) for 300 s. The presented code snippets were selected with the objective of inducing the subject into a state of low cognitive effort which will be used as a reference state during the posterior analysis;
3. **Code inspection**: In this task, a code snippet in C language is displayed to the subject (selected randomly from a set of 4 different code snippets of different complexities) for a maximum of 600 s. In this task, the subject is asked to analyze and inspect the code, aiming for bug detection.

The schematic representation of each run is provided below (see Figure 1).

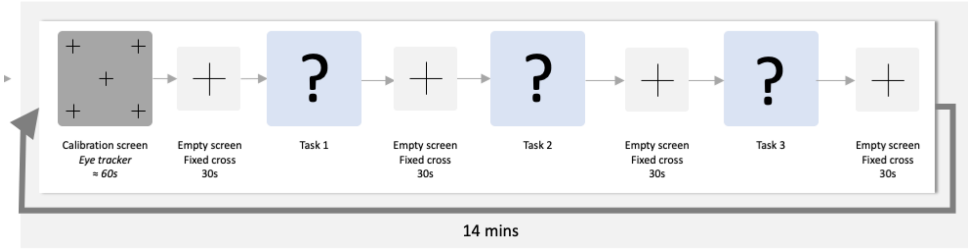


**Figure 1.** Schematic representation of an experiment run.

Each run lasted about 21 min, meaning the whole protocol lasted about 1 h and 20 min. During the experiment, the subjects were alone in a quiet, isolated room when performing the tasks. Furthermore, the subjects were informed a priori about all the protocol and processes of the experiment, and they were also instructed not to take anything that could stimulate/inhibit them the day before the experiment. The code inspection tasks were presented to participants using the Vizard software [21].

The equipment used to collect the electrocardiogram (ECG) signal was the Maglink RT (Neuroscan) with a sampling frequency of 10 kHz [22] (see equipment set-up in Figure 2). For ECG signal acquisition, electrodes from the Neuroscan equipment were positioned in the V1 and V2 locations. The electroencephalogram (EEG) was also collected using an EEG cap with 64 channels, although the measured biosignal was not used in the current analysis.

**Figure 2.** Equipment set-up used in the experiment.

### 3.3. Pre-Processing and ECG Segmentation

Given the nature of the experiment, an initial pre-processing was necessary to remove the gradient artifact (GA) induced by the MRI scanner on the ECG signals. To this end, an average artifact subtraction (AAS) technique based on the algorithm from Niazy et al. [23] was performed to reduce this artifact on ECG data. In addition to the GA correction, the ECG presents some changes in its morphology due to the magnetic field produced by the MRI machine. Therefore, the ECG signal tends to present a T-wave larger than the QRS complex and an R-wave with reduced amplitude. Thus, traditional QRS detection algorithms tend to fail and lead to incorrect RR interval calculation.

Nevertheless, the R-peak detection algorithm proposed by Christov et al. [24] is commonly used in these scenarios, given its robustness and high performance in R-peak detection on ECG signals recorded inside an MRI scanner. The data were visually inspected to assess the quality of the R-peak detection process. After having the R-peaks detected, we proceeded to the RR interval computation to obtain the HRV time series.

### 3.4. Feature Extraction

In order to carry out HRV analysis, following pre-processing and ECG segmentation, we proceeded with the feature extraction from the code inspection data collected during each subject run. A total of 31 features across time, geometrical, non-linear and frequency domains were extracted using a sliding window of variable size and a jumping step of 1 s.

The sliding window size used ranged from 3 min to 10 s, being iteratively reduced by 10 s, amounting to a total of 18 different windows (see Figure 3). All of the 31 features were extracted applying the 18 different sliding windows.

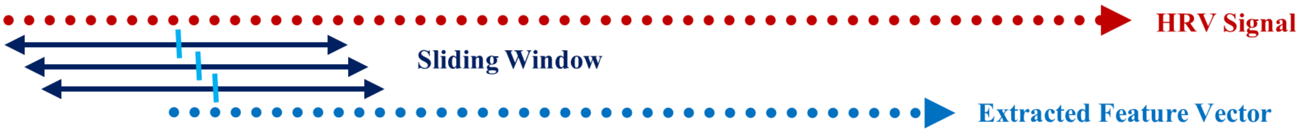


**Figure 3.** Schematic representation of the extraction of a feature using one of the sliding windows. In the end, we obtained a total of 558 feature vectors, corresponding to the 31 features times 18 window sizes for each experiment run.

The described procedure produced vectors of individual measurements from the HRV data collected during the code inspection task (to facilitate referencing, we will call these the “extracted feature vectors”). Each individual measurement was computed based on a RR signal portion with the size of the sliding window employed. The individual measurements were then associated with the time instant corresponding to the center of the RR signal portion used to compute the respective individual measurement.

It is important to mention that the same RR signal produces “extracted feature vectors” of different lengths according to the time frame applied in the extraction process. The vector obtained with the 180-second sliding window is the one with fewer individual measurements, while the vector extracted with the 10-second sliding window is the larger, having more individual measurements. In this study, the “extracted feature vectors” using the 180-second sliding window were used as the gold standard in the statistical and correlation analyses.

The 31 features explored in this study included six features from the time domain, three from the geometrical domain, six from the non-linear domain and 16 from the frequency domain. The different features were selected based on the current literature on ultra-short-term HRV measurements and are the result of a search conducted for the most reliable features extracted using small time frames (see Table 1).

**Table 1.** Set of Features used in the current study presenting the designation used across the document, the units of measurement, a description of the feature and the papers reporting that feature for the analysis of HRV.

| HRV Features Initials | Units | HRV Features Description | References |
|---|---|---|---|
| | | **Time Domain** | |
| mNN | [ms] | mean of NN (or RR) intervals | [2] |
| SDNN | [ms] | standard deviation of NN (or RR) intervals | [2,10] |
| SDSD | [ms] | standard deviation of the differences between heartbeats | [2,10] |
| RMSSD | [ms] | the root mean square of the differences between heartbeats | [2,10] |
| NN50 | – | number of consecutive RR intervals differing by more than 50 milliseconds | [2,10] |
| pNN50 | [%] | proportion of consecutive RR intervals differing by more than 50 milliseconds | [2,10] |
| | | **Geometrical Domain** | |
| TI | – | HRV Triangular Index–integral of the NN interval histogram divided by the height of the histogram | [2,10,25,26] |
| TINN | – | Triangular Interpolation of RR (or NN interval) Histogram—baseline width of the NN interval histogram | [2,10,25,26] |
| SI | – | Baevsky’s Stress Index | [27] |
| | | **Non-Linear Domain** | |
| ApEn | – | Approximate Entropy—measures the complexity or irregularity of the RR series | [28] |
| SD1 | [ms] | Standard Deviation of the Poincare plot perpendicular to the line of identity | [2,6] |
| SD2 | [ms] | Standard Deviation of the Poincare plot along the line of identity | [2,6] |

**Table 1.** *Cont.*

| HRV Features Initials | Units | HRV Features Description | References |
|---|---|---|---|
| PTM | – | Point Transition Measure—quantifies the temporal variation at the point-to-point level of the Poincare plot | [6] |
| KFD | – | Katz Fractal Dimension | [29] |
| HFD | – | Higuchi Fractal Dimension | [30] |
| | | **Frequency Domain** | |
| VLF | [ms$^2$] | Very-Low-Frequency band power ($\leq$0.04 Hz) | [2,10] |
| LF | [ms$^2$] | Low-Frequency band power (0.04–0.15 Hz) | [2,10] |
| HF | [ms$^2$] | High-Frequency band power (0.15–0.4 Hz) | [2,10] |
| VLFnu | n.u. | VLF power normalized | [2,10] |
| Lfnu | n.u. | LF power normalized | [2,10] |
| HFnu | n.u. | HF power normalized | [2,10] |
| VLFpeak | [ms$^2$] | VLF power frequency peak | [2,10] |
| LFpeak | [ms$^2$] | LF power frequency peak | [2,10] |
| HFpeak | [ms$^2$] | HF power frequency peak | [2,10] |
| VLFpeak-nu | n.u. | VLF power frequency peak normalized | [2,10] |
| LFpeak-nu | n.u. | LF power frequency peak normalized | [2,10] |
| HFpeak-nu | n.u. | HF power frequency peak normalized | [2,10] |
| totPow | [ms$^2$] | Total Power | [2,10] |
| Peak | [ms$^2$] | Overall frequency power peak | [2,10] |
| LF/HF | – | Ratio of LF and HF band powers | [2,10] |
| LFpeak/HFpeak | – | Ratio of LF and HF band power frequency peak | [2,10] |

### *3.5. Statistical Analysis*

In order to determine if the features extracted follow a normal distribution, the Kolmogorov–Smirnov test was performed individually by measurement in each experiment run. The test's null hypothesis was that the data follow a standard normal distribution. At a 5% significance level, we obtained the rejection of the null hypothesis for every measurement in all runs. The conclusion was that our data do not follow a standard normal distribution, so the statistical significance and correlation tests applied must be non-parametric. Figure 4 represents the general flow chart of the experimental steps followed to evaluate the ultra-short-term HRV measurements' reliability.

#### 3.5.1. Statistical Significance Test

To assess the sliding window size stability limit for each feature, i.e., to assess the smallest time frame that enables feature stability (when compared to the chosen reference), the Wilcoxon rank sum test was performed. In this test, the measurements extracted using the different time frames were placed against the measurements obtained using the 180-second sliding window. The test was performed independently for every experimental run and to all 31 features in the study. With the explained procedure, we were able to inspect how the variation of the window size in the feature extraction process affected the different measurements, assuming the 180-second window as a reference.

From this process, a $p$-value was obtained for all 31 features extracted using the 18 different sliding windows, totalizing a matrix of 31 × 18 $p$-values for each experimental run of the different subjects. To analyze the global extension of this test across the different runs, we computed the percentage of runs where each feature extracted with specific sliding window size and the same feature extracted using the 180-second sliding window do not present significant statistical differences. These percentages were arranged in tables where the effect of reducing the sliding window size on the extraction can be observed (see Figures 5 and 6).

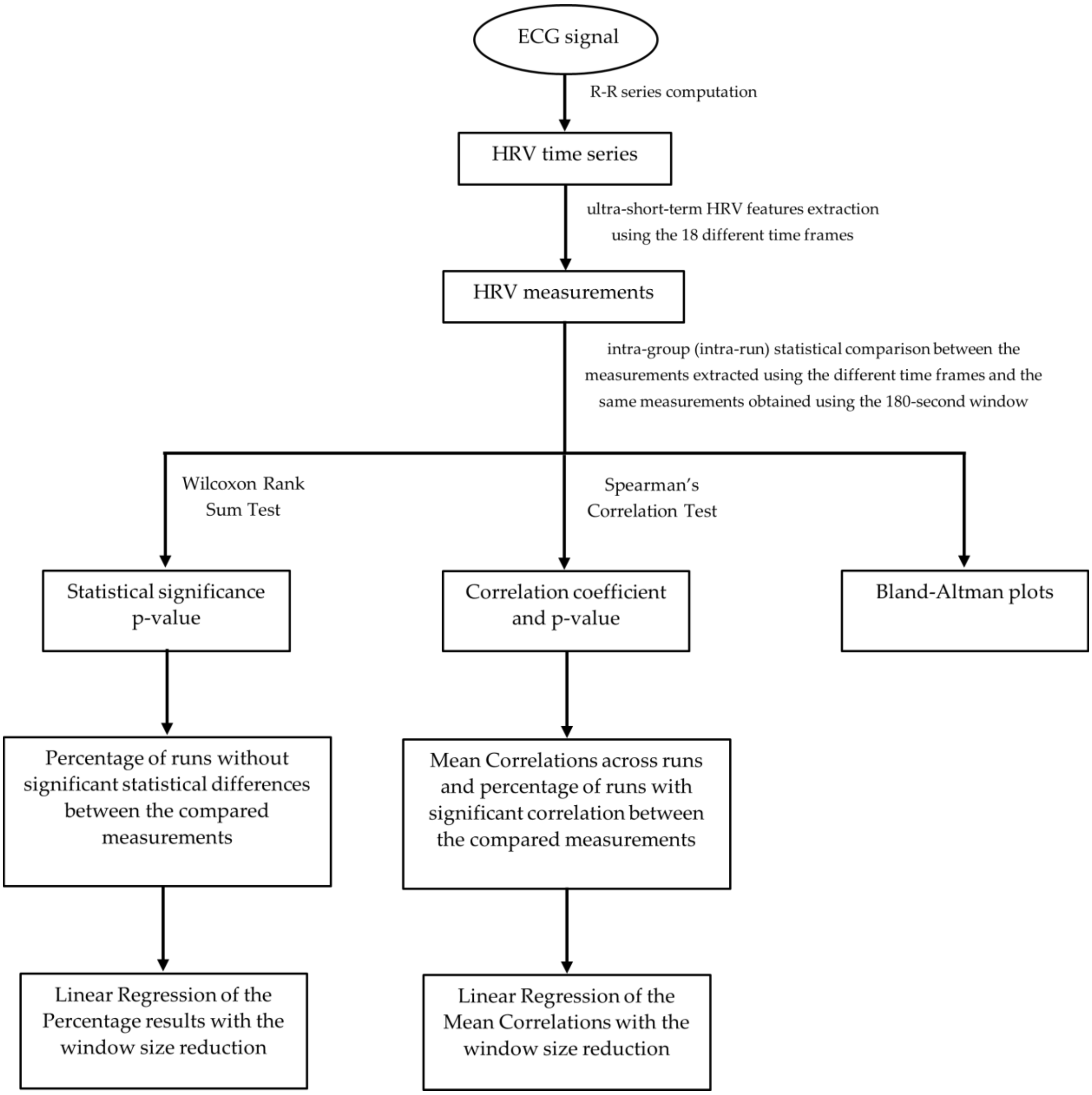
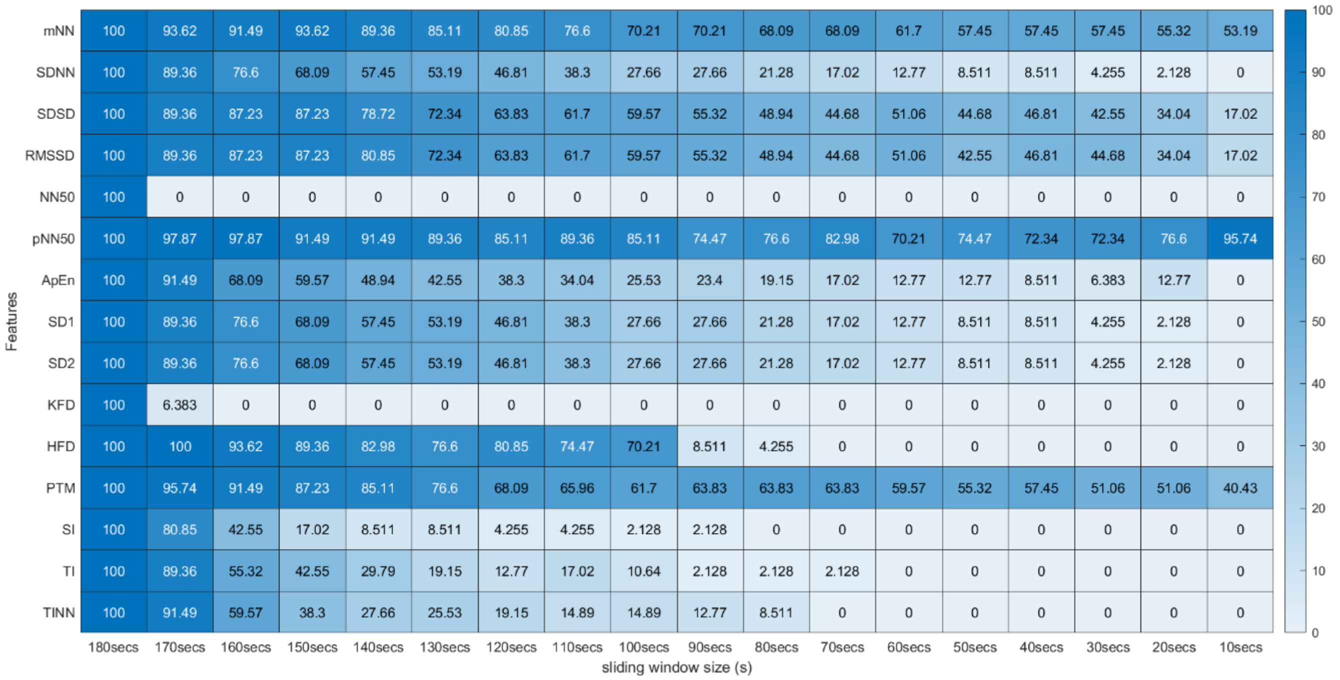

ECG signal
R-R series computation
HRV time series
ultra-short-term HRV features extraction using the 18 different time frames
HRV measurements
intra-group (intra-run) statistical comparison between the measurements extracted using the different time frames and the same measurements obtained using the 180-second window
Wilcoxon Rank Sum Test
Spearman's Correlation Test
Statistical significance p-value
Correlation coefficient and p-value
Bland-Altman plots
Percentage of runs without significant statistical differences between the compared measurements
Mean Correlations across runs and percentage of runs with significant correlation between the compared measurements
Linear Regression of the Percentage results with the window size reduction
Linear Regression of the Mean Correlations with the window size reduction

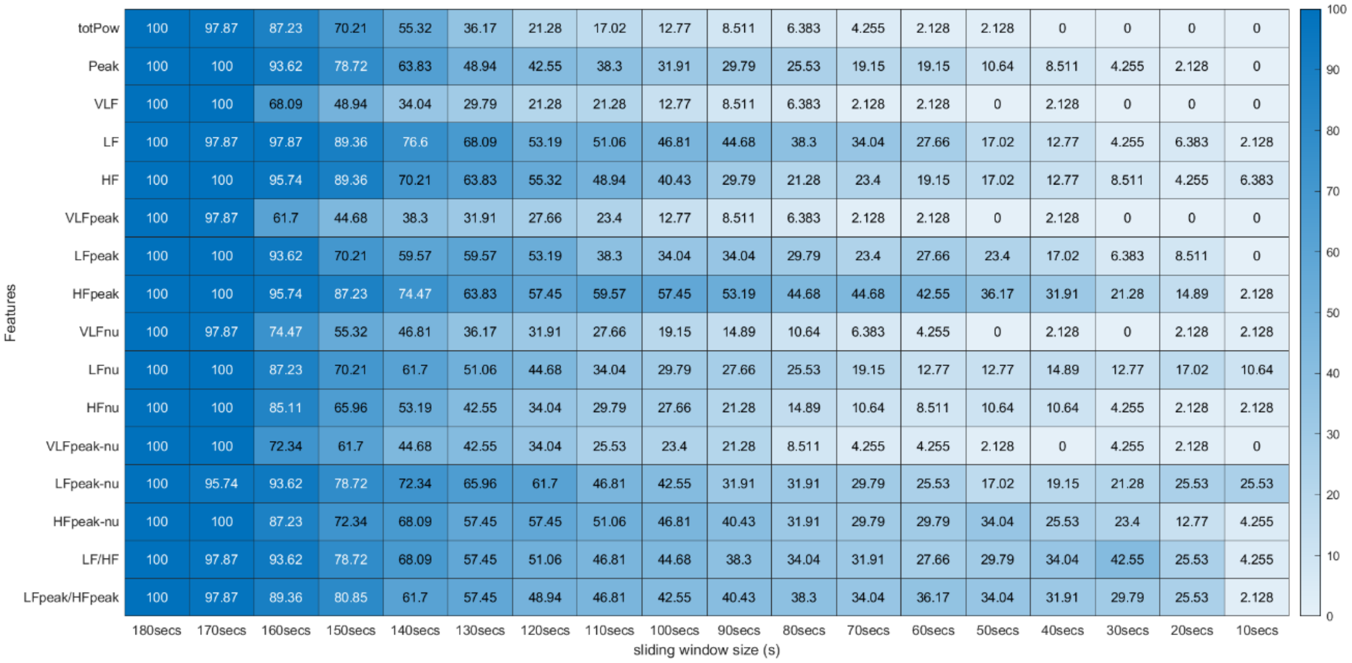
**Figure 4.** General Flow Chart of the experimental steps followed to evaluate the ultra-short-term HRV measurements' reliability.

In order to further investigate the time frame reduction effect, a graph was created for the results obtained from each feature. The 18 different time frames (window sizes) were placed as the independent variable on the *xx* axis, and the percentage of runs without significant statistical differences (results in Figures 5 and 6) as the dependent variable on the *yy* axis. It is essential to mention that the 180-second time frame is positioned at the origin of the *xx* axis, and each unit of this axis corresponds to a reduction of 10 s in the time frame used, the minimum (10-second time frame) being the last result on the *xx* axis. The *yy* axis ranges between 0 and 100%. A linear regression was performed for the results of each feature, and the respective coefficients of determination ($R^2$) were computed for each linear regression (see Figures 7 and 8).



| | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mNN | 100 | 93.62 | 91.49 | 93.62 | 89.36 | 85.11 | 80.85 | 76.6 | 70.21 | 70.21 | 68.09 | 68.09 | 61.7 | 57.45 | 57.45 | 57.45 | 55.32 | 53.19 |
| SDNN | 100 | 89.36 | 76.6 | 68.09 | 57.45 | 53.19 | 46.81 | 38.3 | 27.66 | 27.66 | 21.28 | 17.02 | 12.77 | 8.511 | 8.511 | 4.255 | 2.128 | 0 |
| SDSD | 100 | 89.36 | 87.23 | 87.23 | 78.72 | 72.34 | 63.83 | 61.7 | 59.57 | 55.32 | 48.94 | 44.68 | 51.06 | 44.68 | 46.81 | 42.55 | 34.04 | 17.02 |
| RMSSD | 100 | 89.36 | 87.23 | 87.23 | 80.85 | 72.34 | 63.83 | 61.7 | 59.57 | 55.32 | 48.94 | 44.68 | 51.06 | 42.55 | 46.81 | 44.68 | 34.04 | 17.02 |
| NN50 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pNN50 | 100 | 97.87 | 97.87 | 91.49 | 91.49 | 89.36 | 85.11 | 89.36 | 85.11 | 74.47 | 76.6 | 82.98 | 70.21 | 74.47 | 72.34 | 72.34 | 76.6 | 95.74 |
| ApEn | 100 | 91.49 | 68.09 | 59.57 | 48.94 | 42.55 | 38.3 | 34.04 | 25.53 | 23.4 | 19.15 | 17.02 | 12.77 | 12.77 | 8.511 | 6.383 | 12.77 | 0 |
| SD1 | 100 | 89.36 | 76.6 | 68.09 | 57.45 | 53.19 | 46.81 | 38.3 | 27.66 | 27.66 | 21.28 | 17.02 | 12.77 | 8.511 | 8.511 | 4.255 | 2.128 | 0 |
| SD2 | 100 | 89.36 | 76.6 | 68.09 | 57.45 | 53.19 | 46.81 | 38.3 | 27.66 | 27.66 | 21.28 | 17.02 | 12.77 | 8.511 | 8.511 | 4.255 | 2.128 | 0 |
| KFD | 100 | 6.383 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HFD | 100 | 100 | 93.62 | 89.36 | 82.98 | 76.6 | 80.85 | 74.47 | 70.21 | 8.511 | 4.255 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PTM | 100 | 95.74 | 91.49 | 87.23 | 85.11 | 76.6 | 68.09 | 65.96 | 61.7 | 63.83 | 63.83 | 63.83 | 59.57 | 55.32 | 57.45 | 51.06 | 51.06 | 40.43 |
| SI | 100 | 80.85 | 42.55 | 17.02 | 8.511 | 8.511 | 4.255 | 4.255 | 2.128 | 2.128 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| TI | 100 | 89.36 | 55.32 | 42.55 | 29.79 | 19.15 | 12.77 | 17.02 | 10.64 | 2.128 | 2.128 | 2.128 | 0 | 0 | 0 | 0 | 0 | 0 |
| TINN | 100 | 91.49 | 59.57 | 38.3 | 27.66 | 25.53 | 19.15 | 14.89 | 14.89 | 12.77 | 8.511 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Figure 5.** Wilcoxon Rank Sum Test (Time, Non-Linear and Geometrical Domain) * Percentage of runs where the feature (line) extracted with a respective window size (column) did not present significant statistical differences compared to the same feature extracted using the 180-second window.



| | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| totPow | 100 | 97.87 | 87.23 | 70.21 | 55.32 | 36.17 | 21.28 | 17.02 | 12.77 | 8.511 | 6.383 | 4.255 | 2.128 | 2.128 | 0 | 0 | 0 | 0 |
| Peak | 100 | 100 | 93.62 | 78.72 | 63.83 | 48.94 | 42.55 | 38.3 | 31.91 | 29.79 | 25.53 | 19.15 | 19.15 | 10.64 | 8.511 | 4.255 | 2.128 | 0 |
| VLF | 100 | 100 | 68.09 | 48.94 | 34.04 | 29.79 | 21.28 | 21.28 | 12.77 | 8.511 | 6.383 | 2.128 | 2.128 | 0 | 2.128 | 0 | 0 | 0 |
| LF | 100 | 97.87 | 97.87 | 89.36 | 76.6 | 68.09 | 53.19 | 51.06 | 46.81 | 44.68 | 38.3 | 34.04 | 27.66 | 17.02 | 12.77 | 4.255 | 6.383 | 2.128 |
| HF | 100 | 100 | 95.74 | 89.36 | 70.21 | 63.83 | 55.32 | 48.94 | 40.43 | 29.79 | 21.28 | 23.4 | 19.15 | 17.02 | 12.77 | 8.511 | 4.255 | 6.383 |
| VLFpeak | 100 | 97.87 | 61.7 | 44.68 | 38.3 | 31.91 | 27.66 | 23.4 | 12.77 | 8.511 | 6.383 | 2.128 | 2.128 | 0 | 2.128 | 0 | 0 | 0 |
| LFpeak | 100 | 100 | 93.62 | 70.21 | 59.57 | 59.57 | 53.19 | 38.3 | 34.04 | 34.04 | 29.79 | 23.4 | 27.66 | 23.4 | 17.02 | 6.383 | 8.511 | 0 |
| HFpeak | 100 | 100 | 95.74 | 87.23 | 74.47 | 63.83 | 57.45 | 59.57 | 57.45 | 53.19 | 44.68 | 44.68 | 42.55 | 36.17 | 31.91 | 21.28 | 14.89 | 2.128 |
| VLFnu | 100 | 97.87 | 74.47 | 55.32 | 46.81 | 36.17 | 31.91 | 27.66 | 19.15 | 14.89 | 10.64 | 6.383 | 4.255 | 0 | 2.128 | 0 | 2.128 | 2.128 |
| LFnu | 100 | 100 | 87.23 | 70.21 | 61.7 | 51.06 | 44.68 | 34.04 | 29.79 | 27.66 | 25.53 | 19.15 | 12.77 | 12.77 | 14.89 | 12.77 | 17.02 | 10.64 |
| HFnu | 100 | 100 | 85.11 | 65.96 | 53.19 | 42.55 | 34.04 | 29.79 | 27.66 | 21.28 | 14.89 | 10.64 | 8.511 | 10.64 | 10.64 | 4.255 | 2.128 | 2.128 |
| VLFpeak-nu | 100 | 100 | 72.34 | 61.7 | 44.68 | 42.55 | 34.04 | 25.53 | 23.4 | 21.28 | 8.511 | 4.255 | 4.255 | 2.128 | 0 | 4.255 | 2.128 | 0 |
| LFpeak-nu | 100 | 95.74 | 93.62 | 78.72 | 72.34 | 65.96 | 61.7 | 46.81 | 42.55 | 31.91 | 31.91 | 29.79 | 25.53 | 17.02 | 19.15 | 21.28 | 25.53 | 25.53 |
| HFpeak-nu | 100 | 100 | 87.23 | 72.34 | 68.09 | 57.45 | 57.45 | 51.06 | 46.81 | 40.43 | 31.91 | 29.79 | 29.79 | 34.04 | 25.53 | 23.4 | 12.77 | 4.255 |
| LF/HF | 100 | 97.87 | 93.62 | 78.72 | 68.09 | 57.45 | 51.06 | 46.81 | 44.68 | 38.3 | 34.04 | 31.91 | 27.66 | 29.79 | 34.04 | 42.55 | 25.53 | 4.255 |
| LFpeak/HFpeak | 100 | 97.87 | 89.36 | 80.85 | 61.7 | 57.45 | 48.94 | 46.81 | 42.55 | 40.43 | 38.3 | 34.04 | 36.17 | 34.04 | 31.91 | 29.79 | 25.53 | 2.128 |

**Figure 6.** Wilcoxon Rank Sum Test (Frequency Domain) *. * Percentage of runs where the feature (line) extracted with a respective window size (column) did not present significant statistical differences compared to the same feature extracted using the 180-second window.

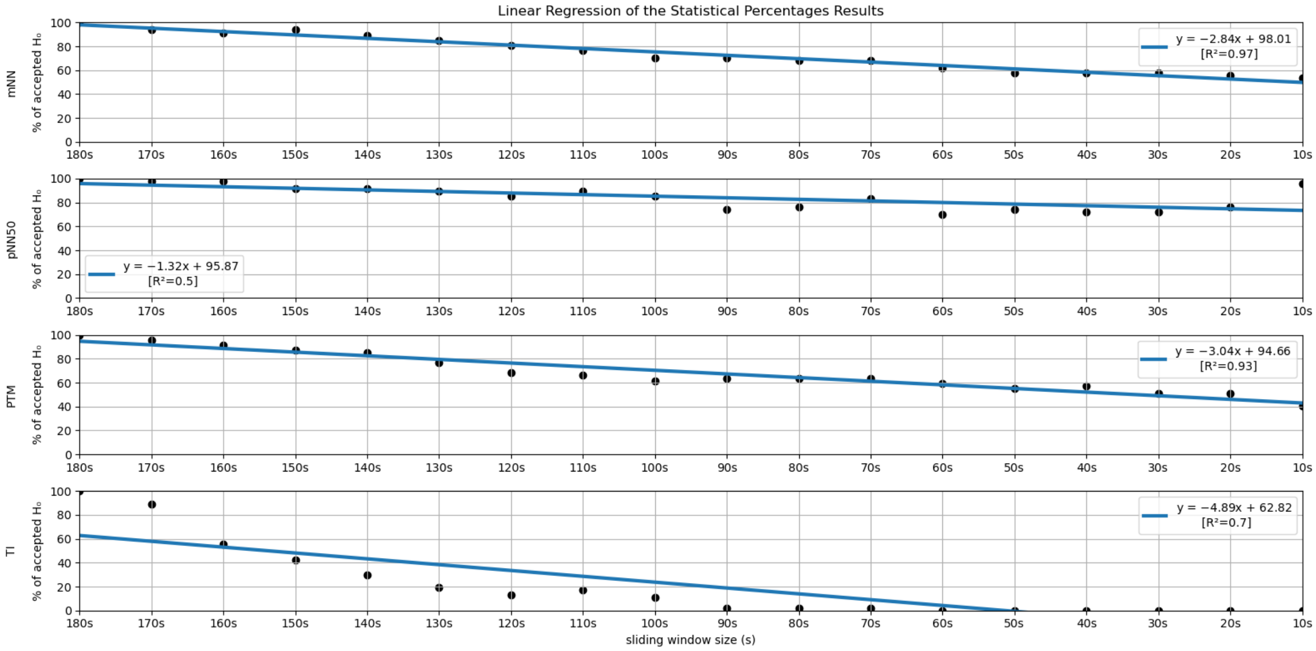


**Figure 7.** Linear Regressions of the Statistical Percentages obtained for the features mNN, pNN50, PTM and TI.

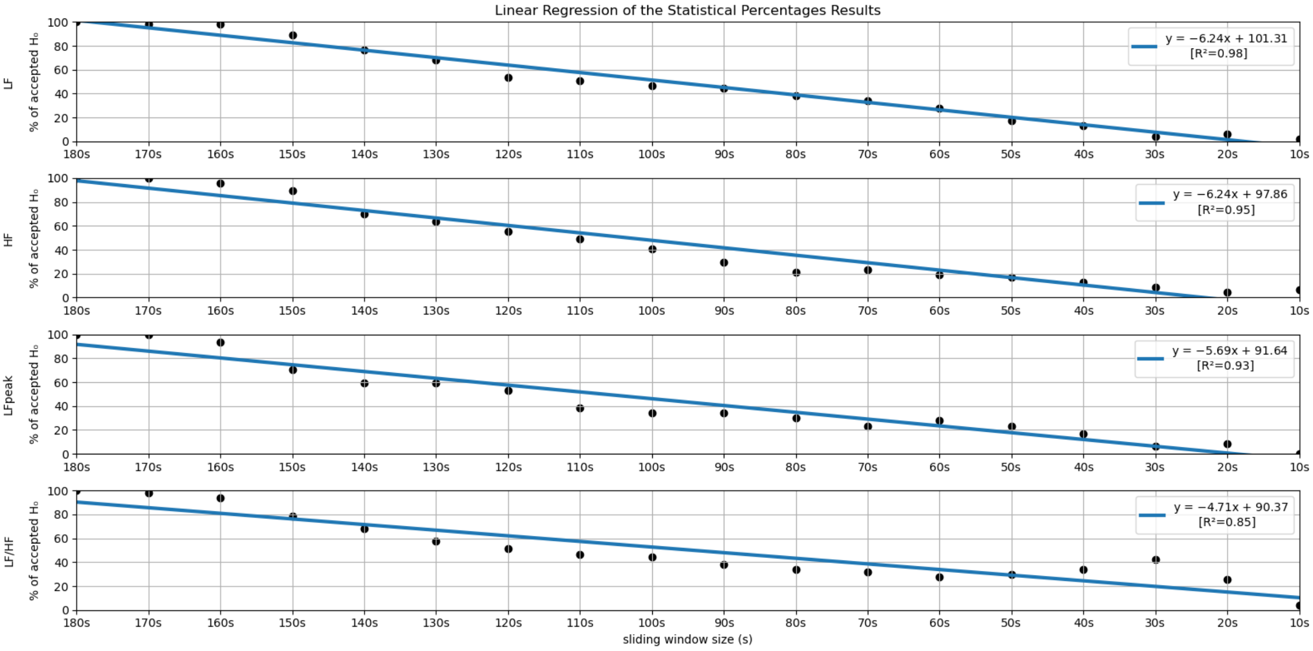


**Figure 8.** Linear Regressions of the Statistical Percentages obtained for the features LF, HF, LFpeak and LF/HF.

### 3.5.2. Correlation Test

In order to complement the insight obtained with the significance test, Spearman's correlation test was also performed. Through the application of the Spearman's correlation test, both a *p*-value and a correlation coefficient were obtained. The *p*-value was used to determine if a significant correlation exists between the data compared, while the correlation coefficient is a measure of how correlated they are. With this test, the measurements obtained with the different sliding windows were compared to the measurements acquired with the 180-second reference time frame. Again, the procedure was conducted independently for each run.

An important difference between the Spearman's correlation test and the significance test of the previous section is that, in this correlation test, we must compare two vectors with the same length. As explained in the feature extraction section, extracting a measurement from an RR signal with a sliding window of 180 s produces a vector inferior in length when compared to extracting the same measurement using a sliding window of an inferior time span. Hence, with this test, we used the portion of the extracted feature vectors, computed using the smaller time frames, corresponding to the instances of the measurements in the vector resulting from the 180-second extraction (see Figure 9).

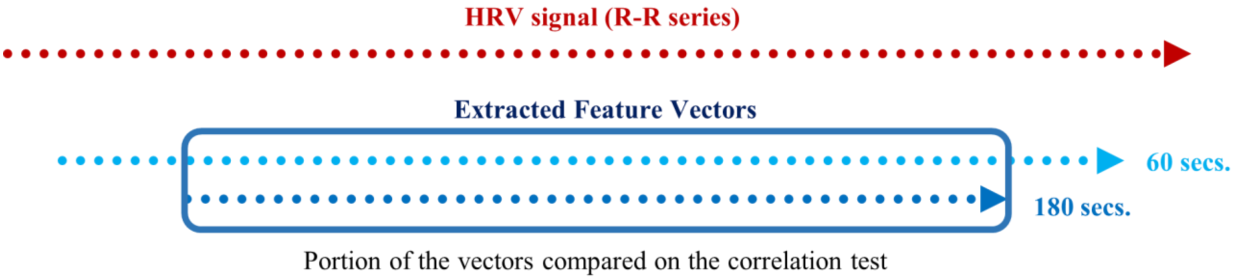
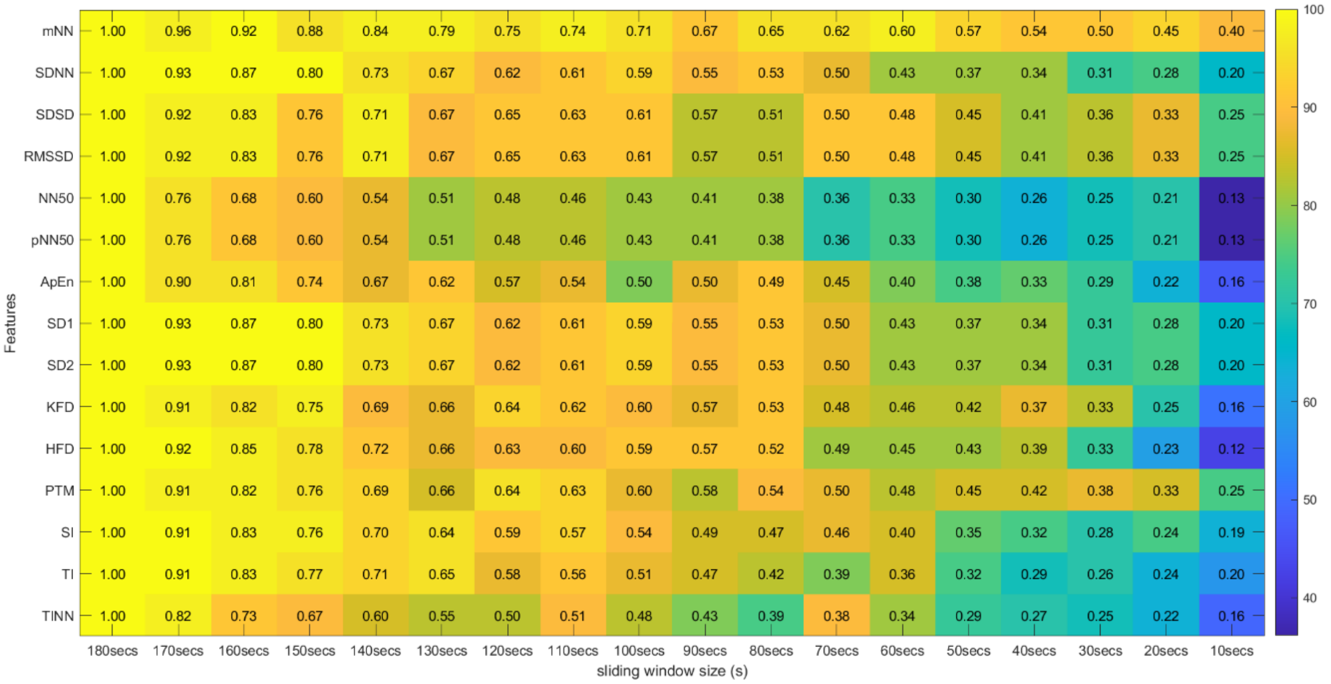

HRV signal (R-R series)
Extracted Feature Vectors
60 secs.
180 secs.
Portion of the vectors compared on the correlation test

**Figure 9.** Schematic of the portion of two feature vectors compared in the correlation test, extracted using 180- and 60-second sliding windows with 1-second steps.

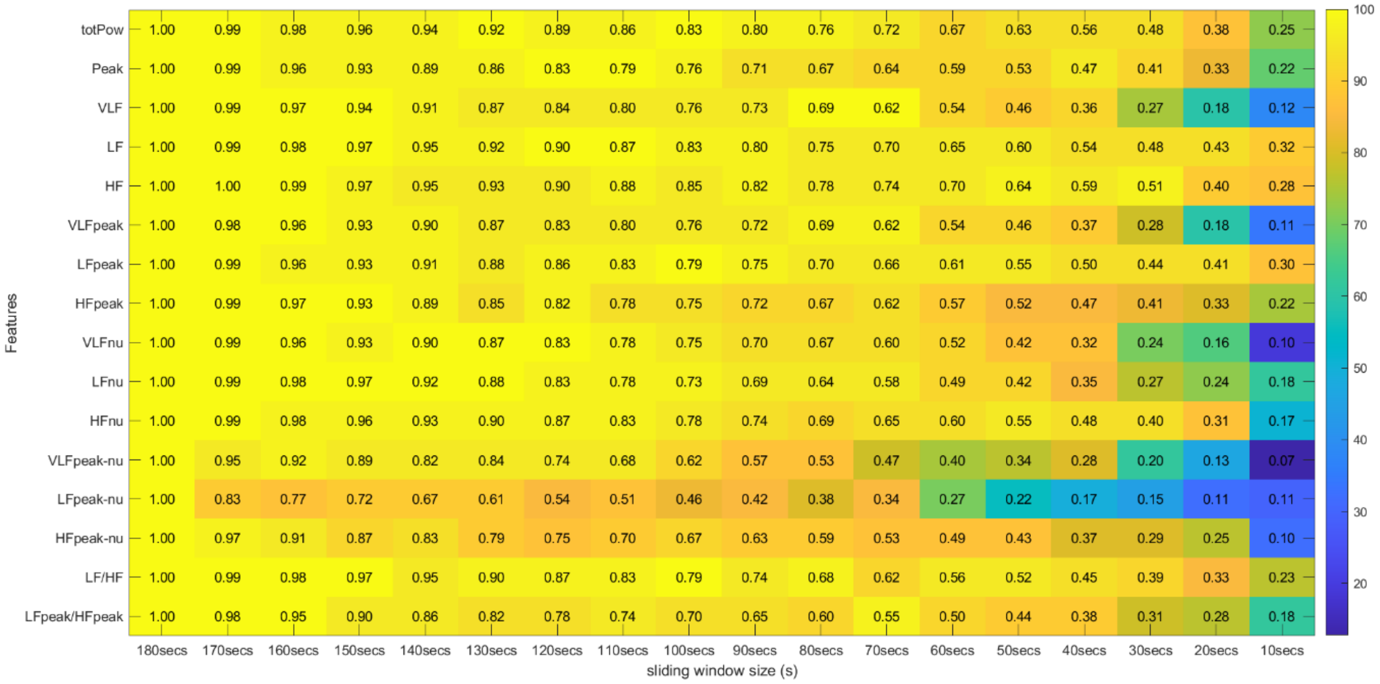
After the completion of the correlation test, for each experimental run of the different subjects, a *p*-value and a correlation coefficient were computed for all 31 features extracted using the 18 different sliding windows in the study (31 × 18 matrix of *p*-values and 31 × 18 matrix of correlation coefficients). After this step, we computed the percentages of runs where significant correlation existed, and these percentages were arranged in a matrix. The matrix lines correspond to the features and the columns to the sliding window size used in their extraction.

Regarding the correlation coefficients obtained through this procedure, we calculated the means of these values across the different runs. Due to existing runs with different sizes, the means were computed using Fisher's z transformation. This method allowed us to give more weight to the features extracted in runs with a larger time length [31]. With this step, the average correlation across runs was obtained between the 31 features extracted with the 180-second window and the same 31 features extracted with the other time frames in the study. The Fisher's mean values were placed in tables where the lines correspond to the features and the columns to the sliding window sizes used in their extraction. In these tables we can efficiently observe the overall effect of reducing the sliding window time span on the correlation values (see Figures 10 and 11).



| | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mNN | 1.00 | 0.96 | 0.92 | 0.88 | 0.84 | 0.79 | 0.75 | 0.74 | 0.71 | 0.67 | 0.65 | 0.62 | 0.60 | 0.57 | 0.54 | 0.50 | 0.45 | 0.40 |
| SDNN | 1.00 | 0.93 | 0.87 | 0.80 | 0.73 | 0.67 | 0.62 | 0.61 | 0.59 | 0.55 | 0.53 | 0.50 | 0.43 | 0.37 | 0.34 | 0.31 | 0.28 | 0.20 |
| SDSD | 1.00 | 0.92 | 0.83 | 0.76 | 0.71 | 0.67 | 0.65 | 0.63 | 0.61 | 0.57 | 0.51 | 0.50 | 0.48 | 0.45 | 0.41 | 0.36 | 0.33 | 0.25 |
| RMSSD | 1.00 | 0.92 | 0.83 | 0.76 | 0.71 | 0.67 | 0.65 | 0.63 | 0.61 | 0.57 | 0.51 | 0.50 | 0.48 | 0.45 | 0.41 | 0.36 | 0.33 | 0.25 |
| NN50 | 1.00 | 0.76 | 0.68 | 0.60 | 0.54 | 0.51 | 0.48 | 0.46 | 0.43 | 0.41 | 0.38 | 0.36 | 0.33 | 0.30 | 0.26 | 0.25 | 0.21 | 0.13 |
| pNN50 | 1.00 | 0.76 | 0.68 | 0.60 | 0.54 | 0.51 | 0.48 | 0.46 | 0.43 | 0.41 | 0.38 | 0.36 | 0.33 | 0.30 | 0.26 | 0.25 | 0.21 | 0.13 |
| ApEn | 1.00 | 0.90 | 0.81 | 0.74 | 0.67 | 0.62 | 0.57 | 0.54 | 0.50 | 0.50 | 0.49 | 0.45 | 0.40 | 0.38 | 0.33 | 0.29 | 0.22 | 0.16 |
| SD1 | 1.00 | 0.93 | 0.87 | 0.80 | 0.73 | 0.67 | 0.62 | 0.61 | 0.59 | 0.55 | 0.53 | 0.50 | 0.43 | 0.37 | 0.34 | 0.31 | 0.28 | 0.20 |
| SD2 | 1.00 | 0.93 | 0.87 | 0.80 | 0.73 | 0.67 | 0.62 | 0.61 | 0.59 | 0.55 | 0.53 | 0.50 | 0.43 | 0.37 | 0.34 | 0.31 | 0.28 | 0.20 |
| KFD | 1.00 | 0.91 | 0.82 | 0.75 | 0.69 | 0.66 | 0.64 | 0.62 | 0.60 | 0.57 | 0.53 | 0.48 | 0.46 | 0.42 | 0.37 | 0.33 | 0.25 | 0.16 |
| HFD | 1.00 | 0.92 | 0.85 | 0.78 | 0.72 | 0.66 | 0.63 | 0.60 | 0.59 | 0.57 | 0.52 | 0.49 | 0.45 | 0.43 | 0.39 | 0.33 | 0.23 | 0.12 |
| PTM | 1.00 | 0.91 | 0.82 | 0.76 | 0.69 | 0.66 | 0.64 | 0.63 | 0.60 | 0.58 | 0.54 | 0.50 | 0.48 | 0.45 | 0.42 | 0.38 | 0.33 | 0.25 |
| SI | 1.00 | 0.91 | 0.83 | 0.76 | 0.70 | 0.64 | 0.59 | 0.57 | 0.54 | 0.49 | 0.47 | 0.46 | 0.40 | 0.35 | 0.32 | 0.28 | 0.24 | 0.19 |
| TI | 1.00 | 0.91 | 0.83 | 0.77 | 0.71 | 0.65 | 0.58 | 0.56 | 0.51 | 0.47 | 0.42 | 0.39 | 0.36 | 0.32 | 0.29 | 0.26 | 0.24 | 0.20 |
| TINN | 1.00 | 0.82 | 0.73 | 0.67 | 0.60 | 0.55 | 0.50 | 0.51 | 0.48 | 0.43 | 0.39 | 0.38 | 0.34 | 0.29 | 0.27 | 0.25 | 0.22 | 0.16 |

**Figure 10.** Spearman's Correlation Test (Time, Non-Linear and Geometrical Domains) ** Heatmap Colors: Percentage of runs where there exists significant correlation between the feature (row) extracted using the respective window size (column) and the same feature obtained using the 180-second sliding window. Cell Values: Means, across the different runs, of the correlation coefficients between the feature (row) extracted using the respective window size (column) and the same feature obtained using the 180-second sliding window.



| | 180secs | 170secs | 160secs | 150secs | 140secs | 130secs | 120secs | 110secs | 100secs | 90secs | 80secs | 70secs | 60secs | 50secs | 40secs | 30secs | 20secs | 10secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| totPow | 1.00 | 0.99 | 0.98 | 0.96 | 0.94 | 0.92 | 0.89 | 0.86 | 0.83 | 0.80 | 0.76 | 0.72 | 0.67 | 0.63 | 0.56 | 0.48 | 0.38 | 0.25 |
| Peak | 1.00 | 0.99 | 0.96 | 0.93 | 0.89 | 0.86 | 0.83 | 0.79 | 0.76 | 0.71 | 0.67 | 0.64 | 0.59 | 0.53 | 0.47 | 0.41 | 0.33 | 0.22 |
| VLF | 1.00 | 0.99 | 0.97 | 0.94 | 0.91 | 0.87 | 0.84 | 0.80 | 0.76 | 0.73 | 0.69 | 0.62 | 0.54 | 0.46 | 0.36 | 0.27 | 0.18 | 0.12 |
| LF | 1.00 | 0.99 | 0.98 | 0.97 | 0.95 | 0.92 | 0.90 | 0.87 | 0.83 | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.54 | 0.48 | 0.43 | 0.32 |
| HF | 1.00 | 1.00 | 0.99 | 0.97 | 0.95 | 0.93 | 0.90 | 0.88 | 0.85 | 0.82 | 0.78 | 0.74 | 0.70 | 0.64 | 0.59 | 0.51 | 0.40 | 0.28 |
| VLFpeak | 1.00 | 0.98 | 0.96 | 0.93 | 0.90 | 0.87 | 0.83 | 0.80 | 0.76 | 0.72 | 0.69 | 0.62 | 0.54 | 0.46 | 0.37 | 0.28 | 0.18 | 0.11 |
| LFpeak | 1.00 | 0.99 | 0.96 | 0.93 | 0.91 | 0.88 | 0.86 | 0.83 | 0.79 | 0.75 | 0.70 | 0.66 | 0.61 | 0.55 | 0.50 | 0.44 | 0.41 | 0.30 |
| HFpeak | 1.00 | 0.99 | 0.97 | 0.93 | 0.89 | 0.85 | 0.82 | 0.78 | 0.75 | 0.72 | 0.67 | 0.62 | 0.57 | 0.52 | 0.47 | 0.41 | 0.33 | 0.22 |
| VLFnu | 1.00 | 0.99 | 0.96 | 0.93 | 0.90 | 0.87 | 0.83 | 0.78 | 0.75 | 0.70 | 0.67 | 0.60 | 0.52 | 0.42 | 0.32 | 0.24 | 0.16 | 0.10 |
| LFnu | 1.00 | 0.99 | 0.98 | 0.97 | 0.92 | 0.88 | 0.83 | 0.78 | 0.73 | 0.69 | 0.64 | 0.58 | 0.49 | 0.42 | 0.35 | 0.27 | 0.24 | 0.18 |
| HFnu | 1.00 | 0.99 | 0.98 | 0.96 | 0.93 | 0.90 | 0.87 | 0.83 | 0.78 | 0.74 | 0.69 | 0.65 | 0.60 | 0.55 | 0.48 | 0.40 | 0.31 | 0.17 |
| VLFpeak-nu | 1.00 | 0.95 | 0.92 | 0.89 | 0.82 | 0.84 | 0.74 | 0.68 | 0.62 | 0.57 | 0.53 | 0.47 | 0.40 | 0.34 | 0.28 | 0.20 | 0.13 | 0.07 |
| LFpeak-nu | 1.00 | 0.83 | 0.77 | 0.72 | 0.67 | 0.61 | 0.54 | 0.51 | 0.46 | 0.42 | 0.38 | 0.34 | 0.27 | 0.22 | 0.17 | 0.15 | 0.11 | 0.11 |
| HFpeak-nu | 1.00 | 0.97 | 0.91 | 0.87 | 0.83 | 0.79 | 0.75 | 0.70 | 0.67 | 0.63 | 0.59 | 0.53 | 0.49 | 0.43 | 0.37 | 0.29 | 0.25 | 0.10 |
| LF/HF | 1.00 | 0.99 | 0.98 | 0.97 | 0.95 | 0.90 | 0.87 | 0.83 | 0.79 | 0.74 | 0.68 | 0.62 | 0.56 | 0.52 | 0.45 | 0.39 | 0.33 | 0.23 |
| LFpeak/HFpeak | 1.00 | 0.98 | 0.95 | 0.90 | 0.86 | 0.82 | 0.78 | 0.74 | 0.70 | 0.65 | 0.60 | 0.55 | 0.50 | 0.44 | 0.38 | 0.31 | 0.28 | 0.18 |

**Figure 11.** Spearman's Correlation Test (Frequency Domain) ** Heatmap Colors: Percentage of runs where there exists significant correlation between the feature (row) extracted using the respective window size (column) and the same feature obtained using the 180-second sliding window. Cell Values: Means, across the different runs, of the correlation coefficients between the feature (row) extracted using the respective window size (column) and the same feature obtained using the 180-second sliding window.

Similar to the process performed in the previous subsection, a linear regression was performed with the mean correlation results obtained for each feature. The coefficients of determination ($R^2$) associated with the linear regression were also computed (see Figures 12 and 13).

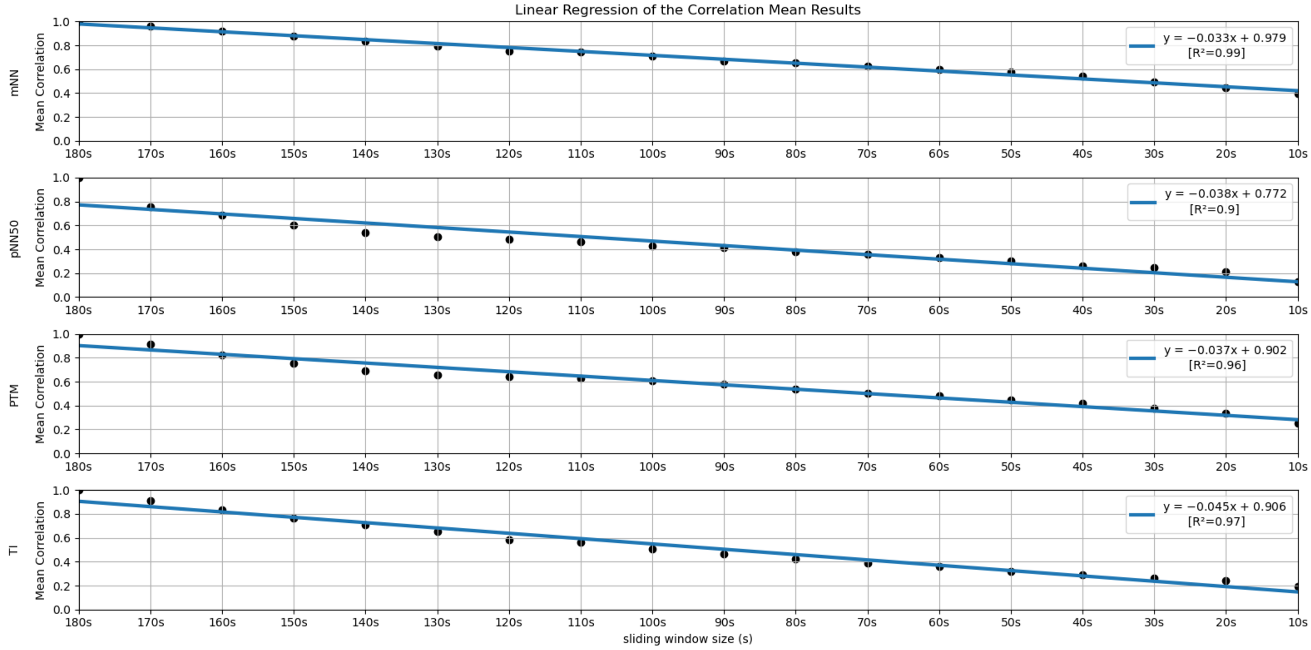


**Figure 12.** Linear Regressions of the Mean Correlations across runs obtained for the features mNN, pNN50, PTM and TI.

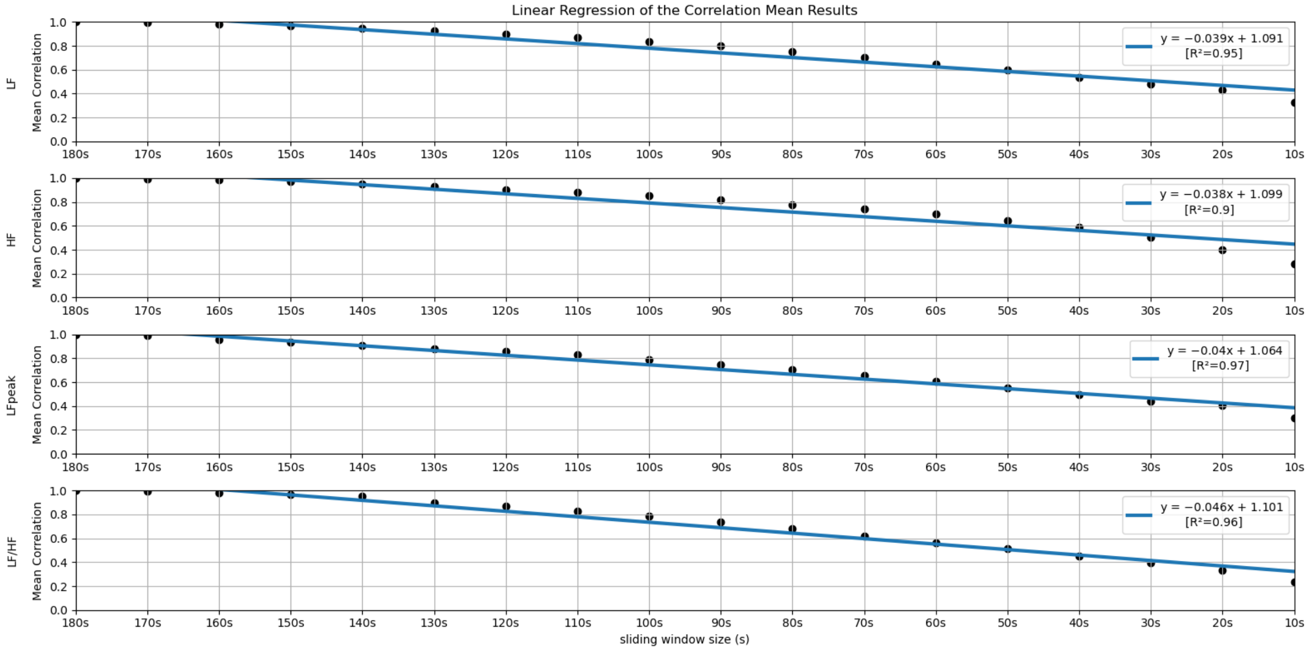


**Figure 13.** Linear Regressions of the Mean Correlations across runs obtained for the features LF, HF, LFpeak and LF/HF.

### 3.5.3. Bland–Altman Plots

Following the recommendations of Pechia et al. [2] and Shaffer et al. [5], we proceeded with Bland–Altman plot analysis to evaluate the features' degree of bias. A relevant difference between our approach and that of existing studies was that we performed an intra-subject and intra-run feature analysis, i.e., we tested the features' correlation and statistical differences using different time frames within the same experimental run. In order to maintain the same intra-group analysis approach, we performed a Bland–Altman plot analysis for non-parametric data for each feature extracted with the different extracting window sizes compared with the same feature extracted with the 180-second window. This procedure was repeated for every experimental run. Figure 14 illustrates 5 Bland–Altman plot examples where it is possible to observe the level of agreement between the compared measurements.

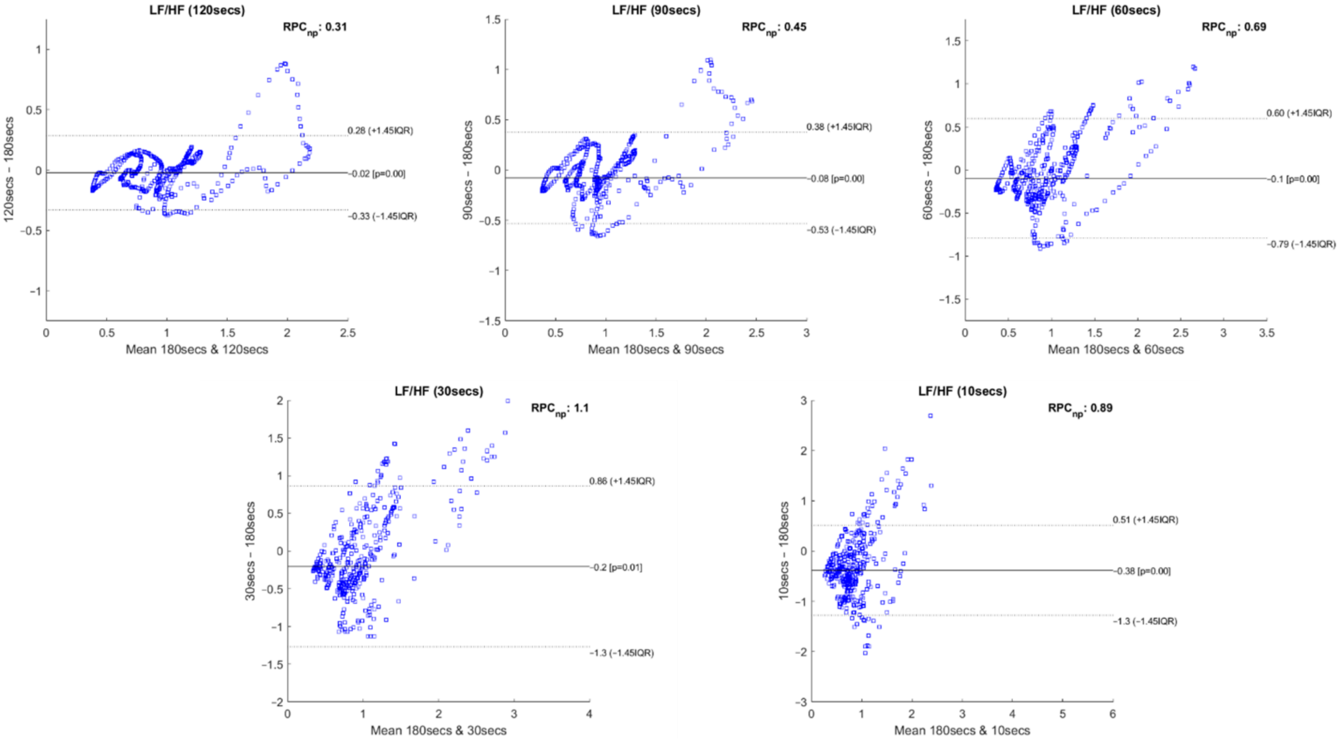


**Figure 14.** Bland–Altman plots of the LF/HF feature extracted with 120-, 90-, 60-, 30- and 10-second time frames compared to the LF/HF extracted with 180-second time frame, regarding a single experimental run of an individual subject.

## 4. Results

### *4.1. Statistical Significance Test Results*

Figures 5 and 6 summarize the results obtained using the procedure introduced in Section 3.5.1. In particular, Figure 5 summarizes the results related to the time, geometrical and non-linear domain features, whereas Figure 6 presents the results achieved using the frequency domain features. The values of each cell correspond to the percentage of runs where there is no significant difference between the feature (row) extracted using a respective window size (column), and the same feature obtained using the 180-second reference sliding window. Figures 7 and 8 are graphical representations of the linear regressions obtained for the time, geometrical and non-linear domain features results and for the frequency domain features results, respectively. The feature lines chosen to be presented were the ones considered representative of the overall results. In Appendix B, it is possible to consult the slopes, the yy interceptions and the coefficients of determination obtained for every feature in the study.

From both tables, it is possible to observe that reducing the sliding window size has a great impact on the results of the significance test in almost every feature. This drop represents a large decrease, through the time frame reduction, in the percentage of runs where there is no significant difference between extracting features using that window and the 180-second reference window. The linear regressions obtained provide quantitative and visual support for this claim.

### 4.2. Correlation Test Results

Figures 10 and 11 introduce the correlation testing results described in Section 3.5.2, i.e., Figure 10 corresponds to the correlation analysis of the time, geometrical and non-linear domain features, whereas Figure 11 corresponds to the correlation analysis of the frequency domain features. Each cell value corresponds to the Fisher's means, across the different experimental runs, of the correlations between the feature (row) extracted using a window size (column) and the same feature obtained using the 180-second sliding window. The heatmap colors correspond to the percentage of runs where a significant correlation ($\alpha = 0.05$) exists between the feature (line) extracted using a given window size (column) and the same feature obtained using the 180-second sliding window.

Figures 12 and 13 allow a visual inspection of the linear regressions obtained for the time, geometrical and non-linear domain features' correlation means and for the frequency domain features' correlation means, respectively. Following the same scheme as Section 4.1 (statistical significance), we selected a few representative linear regression examples to be graphically presented. Appendix C contains the values obtained for the slopes, the yy interceptions and the coefficients of determination.

### 4.3. Bland–Altman Plots Results

Figure 14 depicts the Bland–Altman plots achieved for the feature LF/HF, extracted with the 120-, 90-, 60-, 30- and 10-second time frames compared to the respective features extracted using a 180-second window. The data used to perform these plots correspond to a single experimental run of an individual subject. The Bland–Altman plots allow us to observe the degree of bias present between the compared features and if the data dispersion remains within the 95% line of agreement.

## 5. Discussion

Regarding the statistical analysis and the time domain features, it is observed that four features have significance levels remaining relatively stable throughout variation in the sliding window duration, having a yy interception value close to 100 and a relatively smooth slope. The mentioned features are the SDSD, the RMSSD (basically the normalized version of the SDSD, which explains the similar percentages obtained in both measurements), the pNN50 and the mean NN (mNN). The latter two correspond to the features that exhibit the highest overall stability in the significance study. It is important to note that the linear regression obtained from the pNN50 significance results has a low $R^2$ value (0.50). However, this low value results from a clear outlier in the 10-second time frame.

Keeping our attention on the non-linear and geometrical domain features, these groups have the lowest percentages of runs without significant differences between the compared features. In some cases, linear regressions with yy interception values are much further away from 100 (ApEn, KFD, SI, TI, TINN); in other cases, with very sharp slopes (SD1, SD2, HFD); or with both of these characteristics. This was an expected observation, considering the literature regarding similar studies on ultra-short-term HRV measurements. However, the point transition measure (PTM) shows promising results, since the yy interception is 94.66% and the slope is −3.04, which is a relatively soft slope in the overall context. The fact that this feature proposed in the work by Zubair et al. [6] attempts to quantify the temporal variation in the Poincare plot's point-to-point level may help to explain the much better significance results when compared to other non-linear measurements.

Lastly, regarding the features of the frequency domain, we notice that the features corresponding to the very-low-frequency band have the worst performance in this test. All these features have linear regression yy interception values between 70 and 80% and slope values below −5, which is a relatively sharp slope, considering other features. This result is expected, considering the current literature. It may be explained by the fact that the

VLF band includes waves with 25-second periods and above (frequencies under 0.04 Hz), which means that with a sliding window of fewer than 25 s, we cannot capture a full wave, which increases uncertainty. This complication may also be extended to the low-frequency Band. Considering the scope of the LF band, whose periods range from 6.7 s to 25 s, with a 10-second or a 20-second time frame, it is not possible to capture a complete oscillation period. Even the LF band may not be the best method regarding ultra-short-term HRV measurements. This is well-reflected in the obtained results. According to Pechia et al. [2], it is recommended that spectral analyses are performed on stationary recordings lasting at least ten times longer than the slower significant signal oscillation period. This may help to explain the quick drop in the acceptance percentage results. In fact, from the 110-second window, we soon observe that most features do not have even 50% of the runs without significant differences compared to the 180-second reference extraction window.

The results obtained based on the Wilcoxon rank sum test should be carefully analyzed due to some features' properties and the test's characteristics, which compare the sample distribution and its medians. In fact, a few features in the study rely on the window size used on the extraction to be computed, directly or indirectly, which affects its median values across the time frame reduction, and significant statistical differences will be found comparing these features extracted with two differently sized windows. One example is the NN50, where a larger window will naturally catch a greater number of consecutive RR intervals differing by more than 50 milliseconds for the same cognitive state. The features from the frequency domain, which compute the total power, are other examples (more oversized windows will, expectedly, have higher total power values for the same cognitive state). Furthermore, the HFD relies on the parameter "kmax" for its computation, dependent on the window size employed. On the other hand, features which are normalized values, such as pNN50, tend to have more consistent acceptance percentages through window size reduction. In this way, a comparison using an isolated statistical significance test such as the two-sided Wilcoxon rank sum, which compares the features' medians, may lead to biased results in these features. Furthermore, we believe that having different medians does not mean the feature is not suitable for extraction with smaller windows, considering our current goal, i.e., finding the smallest time frame where each feature behavior is still representative of the corresponding 180-second measurement, for cognitive stress level discrimination purposes.

Another problem with the isolated use of the significance analysis was that, in many features, the acceptance percentage decreases to zero very early as the window size decreases. This fact gives the false impression that, for instance, in the TI features, using a 60-second or a 10-second window essentially produces equivalent results. In this way, the correlation results corroborate some considerations made previously during the analysis of the significance tables. In addition, the linear regression obtained for the correlation results has more solid fits, having no $R^2$ values under 0.90, allowing more accurate conclusions. The correlation analysis may give us more insight into how a feature changes with the reduction of the window size and, in this way, may help us to evaluate each window size until a particular feature remains reliable in our study conditions.

Regarding the correlation analysis, starting with the time domain HRV feature set, the mNN is the only feature where the correlation mean remains above 0.50 until the 60-second window (more precisely, its correlation mean remains above 0.50 until the 30-second time frame). This feature achieved the highest correlation in the smaller time frame in the study (0.40). From the literature, some studies concluded that the mNN is reliable until the 10-second time frame, such as the study by Salahuddin et al. [8], so the expectation was to see higher correlation values until smaller extraction time frames. The same can be said regarding the RMSSD and the pNN50 features. In the literature, these features are often mentioned as being reliable using 60-second time frames and lower [11,16], yet, in the current experiment, the correlation means obtained for these features using the 60-second window were already below 0.50. However, in the study performed by Salahuddin et al. [8], some recommendations by Pechia et al. [2] and by Shaffer et al. [5] are not adopted, such as the recommendations regarding the features' bias quantification, which may increase with time frame reduction. Further, the mentioned study used a 150-second reference for the statistical analysis, while we used a 180-second time frame as a reference. Furthermore, the existing investigations, such as the study by Baek et al. [16], perform an inter-subject analysis of the features. This fact can lead to biased correlation values since it captures the inter-subject feature tendencies that may overwhelm the actual feature tendencies and increase the correlation of the features. In our present study, we perform an intra-subject and intra-run feature analysis, avoiding this kind of bias. This analysis difference explains the lower correlations obtained in the present study. It is also important to underline that, contrary to the most existing literature on the topic, our study is projected in a real-life, non-controlled environment, emulating contexts for real applications, such as bug detection algorithms based on the features under study. Accounting for these considerations, we cannot expect as high correlation values or as clean and clear results as those obtained in more controlled and resting environments, requiring lower cognitive effort.

Regarding the non-linear and geometrical domain HRV features, some features would probably be overlooked considering only the significance results. Let us take, as an example, the significance values of the KFD, the SI, the TI and the TINN. The acceptance percentages in the significance test drop to very low values in the 170-, 140-, 130- and 120-second windows, respectively. However, in the correlation results, we can observe the existence of correlation until smaller windows. Actually, in the KFD feature, more than 50% correlation is observed until 80 s, and this feature has correlation values similar to HFD. This similarity is expected since both features compute the fractal dimension. If the significance test results were the only ones taken into consideration, we could have erroneously concluded that these features are very distinct. From the geometrical and non-linear domains, the PTM was the feature that had a higher correlation within the smaller windows and with the softer slope ($-0.037$) of these two groups, which corroborates previous considerations.

Globally speaking, the features from the frequency domain are the ones that exhibit higher consistent correlation values for smaller windows. Several features from this domain have mean correlation values above 50% until windows of 40 s, with the HF obtaining more than 50% correlation when using the 30-second sliding window. This observation is substantially different from the significance results, which could biasedly suggest that the time domain's features are more reliable in smaller time frames. The set of features HF, LF, LFpeak and totPow contains the features with the most promising correlation results from this domain, having correlation mean values greater than 25% at the 10-second time frame. Analyzing the slopes of the linear regressions achieved using the correlation means, it is observable that the frequency domain features exhibit higher yy interception values, maintaining a relatively softer slope. These facts indicate that their tendencies are less impacted by sliding window size reduction. Once more, as expected, the VLF band had the poorest results from the frequency domain set of features, with the steepest linear regression slopes, despite the correlation results not being as low as the literature would suggest until the 60-second window, compared to the other measurements. The set of features with yy interception values of at least 0.95 and with softer slopes of the overall study were: the mNN ($a = -0.033$), the HF ($a = -0.038$), the LF ($a = -0.039$), the LFpeak ($a = -0.040$) and the totPow ($a = -0.040$). These features are also the those with the higher correlation means in the 10-second time frame. Both these indicators can mean that this set of measurements is adequate to perform the intended analysis in a code inspection context.

Table 2 contains the summarized top five features by sliding window, according to the correlation mean values obtained, where one observes that the features from the frequency domain clearly stand out. In fact, only one feature from the time domain reached these top five features: the mNN when the extracting sliding window was 60 s or under, being the feature with the highest correlation when using the 10-second sliding window. The HF is the most consistent feature with the highest correlation values until the 30-second time frame.

**Table 2.** Top 5 features by time frame regarding the correlation means.

| 120 s | | 90 s | | 60 s | | 30 s | | 10 s | |
|---|---|---|---|---|---|---|---|---|---|
| HF | 90% | HF | 82% | HF | 70% | HF | 51% | mNN | 40% |
| LF | 90% | totPow | 80% | totPow | 67% | mNN | 50% | LF | 32% |
| totPow | 89% | LF | 80% | LF | 65% | LF | 48% | LFpeak | 30% |
| LF/HF | 87% | LFpeak | 75% | LFpeak | 61% | totPow | 48% | HF | 28% |
| HFnu | 87% | HFnu | 74% | mNN | 60% | LFpeak | 44% | totPow | 25% |

From the significance and correlation analyses performed, it is observable that every HRV measurement present in the current study is affected by the time frame size used in their extraction. The Bland–Altman plots further corroborate this statement. These plots allow us to observe the generalized increase in the lines of agreement values of the features with decreasing extracting window size. Figure 14 corresponds to the Bland–Altman plots of the LF/HF feature extracted with 120-, 90-, 60-, 30- and 10-second time frames compared to the same measurements extracted with the 180-second time frame. In these illustrative plots, it is possible to observe this increase in the lines of agreement values with the decrease in the sliding window duration. The number of measurements that fall out of the lines of agreement also increases with reduction in time frame duration. In the LF/HF feature, this effect is very clearly observable. In this way, we can conclude that the degree of bias increases with a reduction in the analysis window size compared to the 180-second measurements. However, in this study context, it is observed that the variability which occurs in some features might be due to the fact that the samples extracted from the 180-second window capture a more overall picture of the ANS dynamics, i.e., during a window of 180-second duration, a higher degree of variability of the ANS activity might exist due to a higher degree of variability in cognitive stress during that period, in comparison to the samples extracted from the shorter windows, where a lower degree of the variability of cognitive stress is observed. This remark is in accordance with the increased variability observed as the time window duration is decreased. Therefore, given the task's nature and application, the existence of variability on some of the Bland–Altman plots might be seen as a concern but not as a limitation for application in software engineering, given the high correlations between the two time series of comparison (180-second vs. shorter windows). These results show that the prevalent cognitive state in both windows is similar but not necessarily equal, since larger windows will capture higher cognitive state fluctuations compared to shorter windows. Furthermore, these differences can be readily captured and compensated by current machine learning and statistical techniques used to model risk scores based on HRV.

Considering the results obtained, it is possible to observe that a set of features remains stable with the reduction in the window analysis size and is reliable for time frames of reduced duration. However, we also believe that some features that have underperformed results should not be excluded just yet, as they might contain complementary information, e.g., for class discrimination, which could be exploited when using machine learning algorithms. This should be assessed based on each problem at hand.

*Threats to Validity*

In this experiment, some limitations were present, which translated into threats to our conclusion's validity, which should also be discussed. First, it is essential to mention that the data collection study was designed with a broader goal and not specifically for HRV stability assessment, and, as such, several different biosignals and images were collected. Functional magnetic resonance imaging (fMRI) was one of the examinations performed. This examination forces the entirety of the experiment to be performed inside an fMRI scanner. The fMRI has an inherent noise effect on the ECG signal. This effect was mitigated through several ECG pre-processing and segmentation methods (Section 3.2: Pre-processing and ECG segmentation). The methods employed effectively mitigate the fMRI noise and are capable of detecting ECG peaks, which is necessary to compute a quality HRV signal. Furthermore, the subjects were alone in a quiet and isolated room when performing the tasks to control the experimental environment. Furthermore, the subjects were informed a priori about all the protocol and processes of the experiment and were instructed not to take anything that could stimulate/inhibit them the day before the experiment. Nevertheless, these external effects are minimized, given that the potential effect is blurred as we perform an intra-subject analysis, and the external effects are present in the measurements extracted with the different windows compared.

Another limitation of our study was the time frame used as a reference (180 s), which is already considered an ultra-short-term HRV. Ideally, a 5-minute (300 s) reference window would be preferable since this is a well-known and consensual time frame in the scientific community. That being said, this was not possible due to our dataset constraints. From our original 21 × 4 (subjects × runs by subject) runs, we had a few middle-run dropouts, which led to only 47 runs having more than 180 s. If the chosen reference were 300 s measurements, the dataset would be substantially reduced, leading to lower statistical power. Furthermore, the study is performed during software code inspection tasks (i.e., bug detection), which is a highly complex, dynamic and cognitively demanding task—in this study context, a 5-minute window is a considerably large window. A window of this size would capture physiological data corresponding to more than one code section, where the subject could feel different difficulty levels, leading to inaccurate results since it would capture different ANS dynamics. Another relevant constraint in the dataset is the fact that all study subjects had the same gender (male). This fact is hard to counter because the software engineering and programming fields are largely dominated by male subjects, which makes it challenging to balance the gender groups in the experiment.

Regarding the study context, most of the related work carried out until now was developed with the subjects at rest or performing elementary tasks in very controlled environments. In contrast, our study is done in a highly demanding task environment. Naturally, the dynamic characteristics of the higher cognitive function, resulting from our experimental context, will generate more dynamic signals. Additionally, the code sections inspected do not all have the same complexity. In this way, the transition from one code section to the next is expected to produce physiological signals with different characteristics and patterns, which are expected to present high variability in these periods, impacting the statistical and correlation analyses.

## 6. Conclusions

This paper studied the impact of reducing the duration of time frames on the HRV feature extraction process. The main goal of the present work was to investigate ultra-short-term HRV features to determine whether HRV-based tools can effectively be used in software development environments. To this extent, our present study investigated the smallest time frame, i.e., the finest time resolution, where each feature is reliable, i.e., the smallest time frame where each feature behavior is still representative of the corresponding 180-second measurement, under our experimental context.

Considering the results obtained, it is observed that the chosen time frame significantly impacts every feature in the study. The features from the frequency domain are those that maintain higher correlation levels until the smaller extraction window durations. From the set of the considered HRV features in this analysis, 13 features had at least 50% correlation when using the 60-second time frame (12 from the frequency domain and only 1, the mNN, from the time domain). The lower statistical significance results can be explained by the fact that features such as HF or LF compute the total power of the respective band. Using a window with a larger size will, expectedly, have higher total power values for the same cognitive state. Despite this fact, these features accurately represent the corresponding 180-second measurements' behavior, as observable in the correlation results. Furthermore, for cognitive stress level discrimination purposes, we do not need an exact surrogate of the short-term measurements, and the feature behavior and tendencies resultant from autonomous nervous system changes can be used to evaluate different cognitive stress levels.

Regarding the smaller window size in the study (10 s), only three features exhibited at least 30% correlation: the mNN, the LF and the LFpeak. Thus, a 10-second time frame is too optimistic in our study context (high cognitive stress). The 30-second time frame is the smallest window with features with at least 50% correlation, and only two fulfilled this criterion: the HF and the mNN. The mNN, the HF, the LF, the LFpeak and the totPow features presented softer linear regression slopes of the overall correlation analysis, with a yy interception value above 0.95, meaning they are less impacted by a reduction in the time frame duration. In this way, this set of five features has shown to be the most reliable for the smallest time frames considering the present context. The mNN feature has proven to be particularly robust to the reduction of the extracting window duration. This feature had a correlation mean of 50% using a 30-second window and showed no significant statistical differences in more than 50% of the experimental runs using all the sliding windows under study, while maintaining a low degree of bias compared to the 180-second reference.

Considering all the results, in a cognitively demanding task context, a classifier built with features extracted using time frames under 30 s might lead to inconsistent results, with potentially low scores and high deviations. However, further study is required to assess whether to discard features extracted using smaller time frames in machine learning contexts, since these features may catch some shorter cognitive patterns that larger time frames may not be able to discriminate. An approach using classifiers trained with datasets, each composed of features extracted with a different time frame, may offer more extensive insight and help to answer the raised hypothesis.

**Author Contributions:** Conceptualization, A.B., R.C., J.M., J.H., C.T., M.S., H.M. and P.C.; data curation, A.B., R.C., J.M., J.D. and R.B.; formal analysis, A.B.; funding acquisition, H.M.; investigation, A.B., R.C., J.M., J.D. and R.B.; methodology, A.B.; project administration, H.M. and P.C.; supervision, R.C., J.M. and P.C.; visualization, A.B.; writing—original draft, A.B.; writing—review and editing, A.B., R.C., J.M., J.H., C.T., M.S., H.M. and P.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially funded by the BASE project under Grant POCI-01-0145-FEDER-031581, in part by the Centro de Informática e Sistemas da Universidade de Coimbra (CISUC-funded by the FCT—Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit—UIDB/00326/2020 or project code UIDP/00326/2020), and also in part by Coimbra Institute for Biomedical Imaging and Translational Re-search (CIBIT), Institute of Nuclear Sciences Applied to Health (ICNAS), the University of Coimbra under Grant PTDC/PSI-GER/30852/2017 | CONNECT-BCI.

**Institutional Review Board Statement:** This study was approved by the Ethical Committee of the Faculty of Medicine of the University of Coimbra, following the Declaration of Helsinki and the standard procedures for studies involving human subjects.

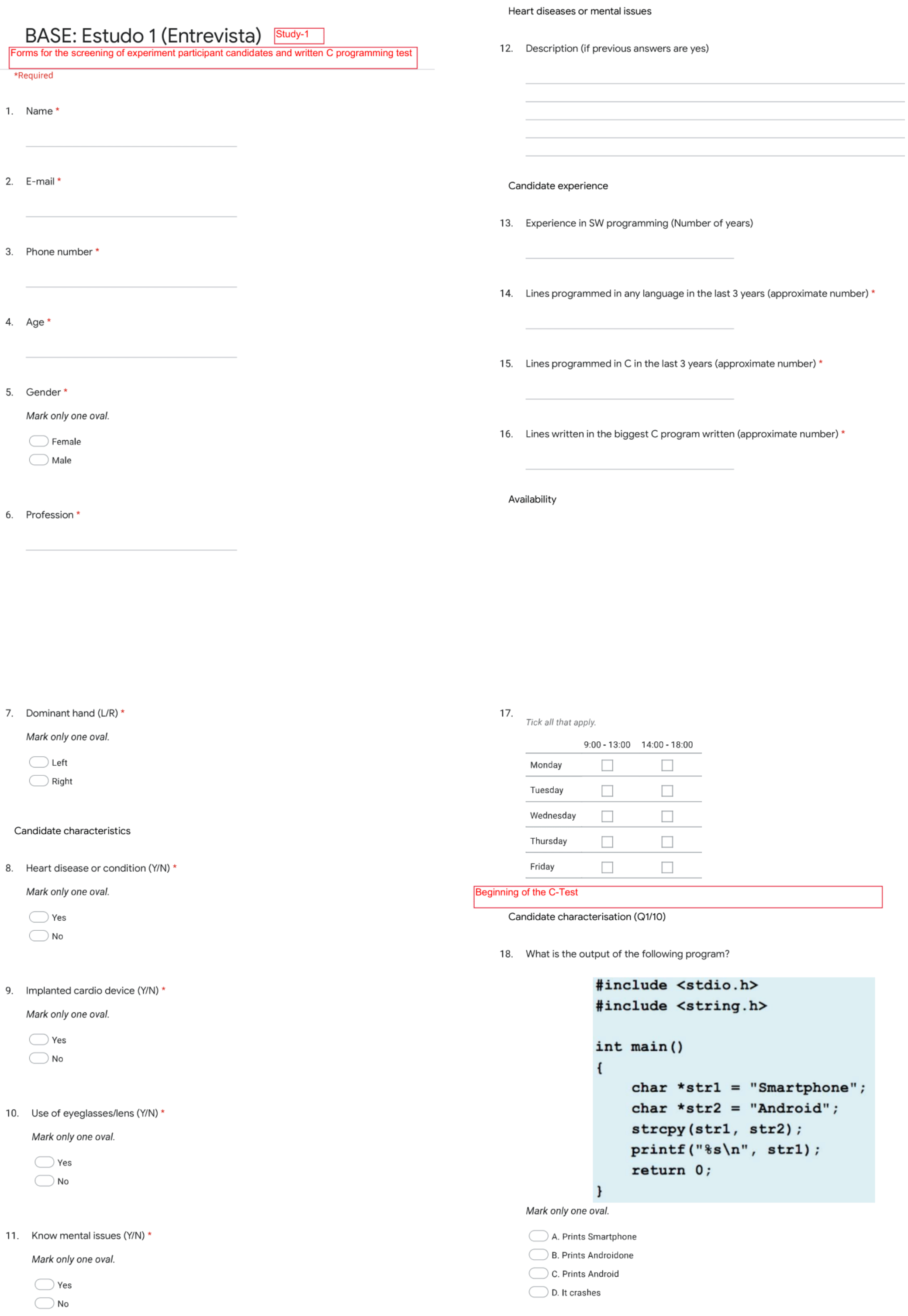
**Informed Consent Statement:** All subjects provided written informed consent and all data have been anonymized.

**Data Availability Statement:** The repository of the dataset used in this study is publicly available in the AI4EI platform (A European AI On Demand Platform and Ecosystem) and can be accessed in following link: https://ai4eu.dei.uc.pt/base-mental-effort-monitoring-dataset (accessed on 24 August 2022) Any other request of data can be submitted to the article authors.

**Acknowledgments:** The authors thank all the volunteers who took part in the experiment.

**Conflicts of Interest:** All the authors of this paper declare no conflict of interest. The funders did not take part in the study design, collection and analyses of data or in the writing of the manuscript. The decision to publish the results was made by the authors without the interference of the funders.

## Appendix A. Screening Questions and C-Test



BASE: Estudo 1 (Entrevista) Study-1

Forms for the screening of experiment participant candidates and written C programming test

*Required

1. Name *

2. E-mail *

3. Phone number *

4. Age *

5. Gender *

*Mark only one oval.*

○ Female
○ Male

6. Profession *

7. Dominant hand (L/R) *

*Mark only one oval.*

○ Left
○ Right

Candidate characteristics

8. Heart disease or condition (Y/N) *

*Mark only one oval.*

○ Yes
○ No

9. Implanted cardio device (Y/N) *

*Mark only one oval.*

○ Yes
○ No

10. Use of eyeglasses/lens (Y/N) *

*Mark only one oval.*

○ Yes
○ No

11. Know mental issues (Y/N) *

*Mark only one oval.*

○ Yes
○ No

Heart diseases or mental issues

12. Description (if previous answers are yes)

Candidate experience

13. Experience in SW programming (Number of years)

14. Lines programmed in any language in the last 3 years (approximate number) *

15. Lines programmed in C in the last 3 years (approximate number) *

16. Lines written in the biggest C program written (approximate number) *

Availability

17.

*Tick all that apply.*

| | 9:00 - 13:00 | 14:00 - 18:00 |
|---|---|---|
| Monday | □ | □ |
| Tuesday | □ | □ |
| Wednesday | □ | □ |
| Thursday | □ | □ |
| Friday | □ | □ |

Beginning of the C-Test

Candidate characterisation (Q1/10)

18. What is the output of the following program?

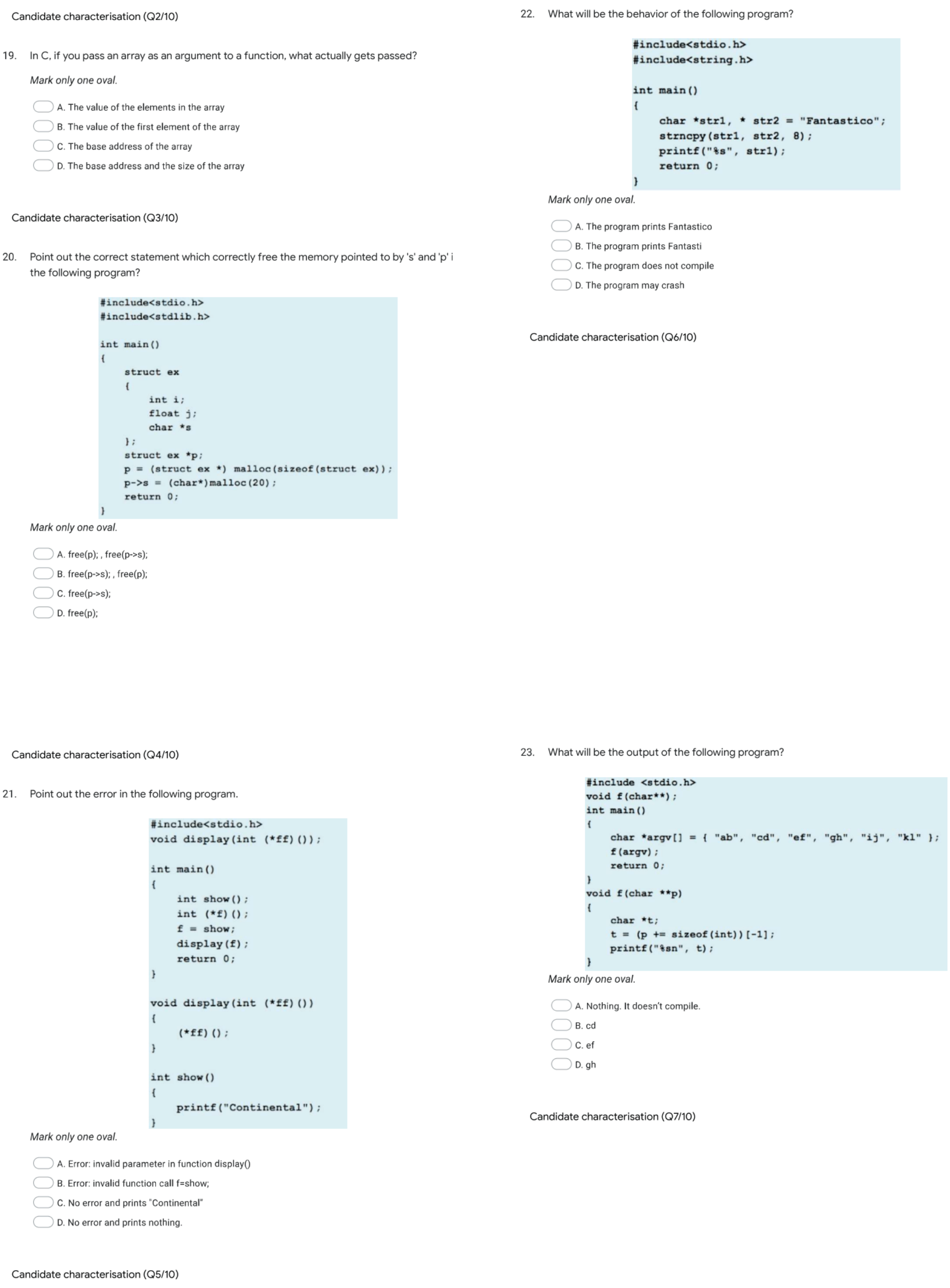
```
#include <stdio.h>
#include <string.h>

int main()
{
    char *str1 = "Smartphone";
    char *str2 = "Android";
    strcpy(str1, str2);
    printf("%s\n", str1);
    return 0;
}
```

*Mark only one oval.*

○ A. Prints Smartphone
○ B. Prints Androidone
○ C. Prints Android
○ D. It crashes

**Figure A1.** Programming experience questionnaire (questions 1 to 17) and technical questionnaire subpart (question 18).



Candidate characterisation (Q2/10)

19. In C, if you pass an array as an argument to a function, what actually gets passed?

*Mark only one oval.*

○ A. The value of the elements in the array
○ B. The value of the first element of the array
○ C. The base address of the array
○ D. The base address and the size of the array

Candidate characterisation (Q3/10)

20. Point out the correct statement which correctly free the memory pointed to by 's' and 'p' i the following program?

```
#include<stdio.h>
#include<stdlib.h>

int main()
{
    struct ex
    {
        int i;
        float j;
        char *s
    };
    struct ex *p;
    p = (struct ex *) malloc(sizeof(struct ex));
    p->s = (char*)malloc(20);
    return 0;
}
```

*Mark only one oval.*

○ A. free(p); , free(p->s);
○ B. free(p->s); , free(p);
○ C. free(p->s);
○ D. free(p);

Candidate characterisation (Q4/10)

21. Point out the error in the following program.

```
#include<stdio.h>
void display(int (*ff)());

int main()
{
    int show();
    int (*f)();
    f = show;
    display(f);
    return 0;
}

void display(int (*ff)())
{
    (*ff)();
}

int show()
{
    printf("Continental");
}
```

*Mark only one oval.*

○ A. Error: invalid parameter in function display()
○ B. Error: invalid function call f=show;
○ C. No error and prints "Continental"
○ D. No error and prints nothing.

Candidate characterisation (Q5/10)

22. What will be the behavior of the following program?

```
#include<stdio.h>
#include<string.h>

int main()
{
    char *str1, * str2 = "Fantastico";
    strncpy(str1, str2, 8);
    printf("%s", str1);
    return 0;
}
```

*Mark only one oval.*

○ A. The program prints Fantastico
○ B. The program prints Fantasti
○ C. The program does not compile
○ D. The program may crash

Candidate characterisation (Q6/10)

23. What will be the output of the following program?

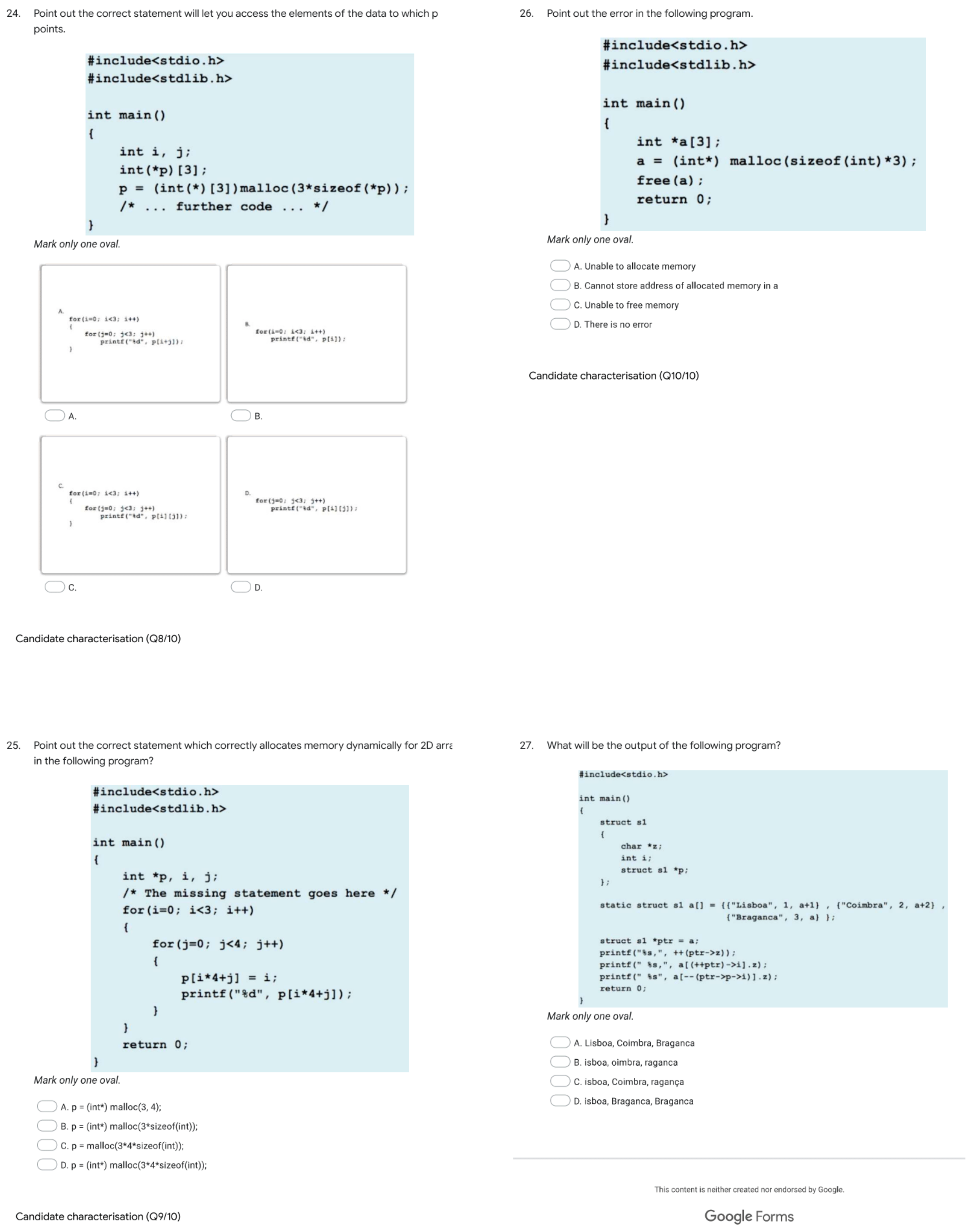
```
#include <stdio.h>
void f(char**);
int main()
{
    char *argv[] = { "ab", "cd", "ef", "gh", "ij", "kl" };
    f(argv);
    return 0;
}
void f(char **p)
{
    char *t;
    t = (p += sizeof(int))[-1];
    printf("%sn", t);
}
```

*Mark only one oval.*

○ A. Nothing. It doesn't compile.
○ B. cd
○ C. ef
○ D. gh

Candidate characterisation (Q7/10)

**Figure A2.** Technical questionnaire subpart (question 19 to 23).



24. Point out the correct statement will let you access the elements of the data to which p points.

```
#include<stdio.h>
#include<stdlib.h>

int main()
{
    int i, j;
    int(*p)[3];
    p = (int(*)[3])malloc(3*sizeof(*p));
    /* ... further code ... */
}
```

*Mark only one oval.*

A.
```
for(i=0; i<3; i++)
{
    for(j=0; j<3; j++)
        printf("%d", p[i+j]);
}
```

B.
```
for(i=0; i<3; i++)
    printf("%d", p[i]);
```

C.
```
for(i=0; i<3; i++)
{
    for(j=0; j<3; j++)
        printf("%d", p[i][j]);
}
```

D.
```
for(j=0; j<3; j++)
    printf("%d", p[i][j]);
```

○ A.
○ B.
○ C.
○ D.

Candidate characterisation (Q8/10)

25. Point out the correct statement which correctly allocates memory dynamically for 2D arra in the following program?

```
#include<stdio.h>
#include<stdlib.h>

int main()
{
    int *p, i, j;
    /* The missing statement goes here */
    for(i=0; i<3; i++)
    {
        for(j=0; j<4; j++)
        {
            p[i*4+j] = i;
            printf("%d", p[i*4+j]);
        }
    }
    return 0;
}
```

*Mark only one oval.*

○ A. p = (int*) malloc(3, 4);
○ B. p = (int*) malloc(3*sizeof(int));
○ C. p = malloc(3*4*sizeof(int));
○ D. p = (int*) malloc(3*4*sizeof(int));

Candidate characterisation (Q9/10)

26. Point out the error in the following program.

```
#include<stdio.h>
#include<stdlib.h>

int main()
{
    int *a[3];
    a = (int*) malloc(sizeof(int)*3);
    free(a);
    return 0;
}
```

*Mark only one oval.*

○ A. Unable to allocate memory
○ B. Cannot store address of allocated memory in a
○ C. Unable to free memory
○ D. There is no error

Candidate characterisation (Q10/10)

27. What will be the output of the following program?

```
#include<stdio.h>

int main()
{
    struct s1
    {
        char *z;
        int i;
        struct s1 *p;
    };

    static struct s1 a[] = {{"Lisboa", 1, a+1} , {"Coimbra", 2, a+2} ,
                                  {"Braganca", 3, a} };

    struct s1 *ptr = a;
    printf("%s,", ++(ptr->z));
    printf(" %s,", a[(++ptr)->i].z);
    printf(" %s", a[--(ptr->p->i)].z);
    return 0;
}
```

*Mark only one oval.*

○ A. Lisboa, Coimbra, Braganca
○ B. isboa, oimbra, raganca
○ C. isboa, Coimbra, ragança
○ D. isboa, Braganca, Braganca

This content is neither created nor endorsed by Google.

**Figure A3.** Technical questionnaire subpart (question 24 to 27).

## Appendix B

**Table A1.** Wilcoxon Rank Sum Test—Linear Regressions of the Statistical Percentages (subpart 1).

| | Time Domain | | | | | | Non-Linear Domain | | | | | | Geometrical Domain | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mNN | SDNN | SDSD | RMSSD | NN50 | pNN50 | ApEn | SD1 | SD2 | KFD | HFD | PTM | SI | TI | TINN |
| a | −2.84 | −5.70 | −3.98 | −3.99 | −1.75 | −1.32 | −5.09 | −5.70 | −5.70 | −1.85 | −7.60 | −3.04 | −3.95 | −4.89 | −4.97 |
| b | 98.01 | 85.13 | 94.14 | 94.35 | 20.47 | 95.87 | 77.78 | 85.13 | 85.13 | 21.66 | 107.94 | 94.66 | 48.55 | 62.82 | 65.15 |
| $R^2$ | 0.97 | 0.94 | 0.94 | 0.94 | 0.16 | 0.50 | 0.87 | 0.94 | 0.94 | 0.18 | 0.85 | 0.93 | 0.51 | 0.70 | 0.73 |

**Table A2.** Wilcoxon Rank Sum Test—Linear Regressions of the Statistical Percentages (subpart 2).

| | Frequency Domain | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **totPow** | **Peak** | **VLF** | **LF** | **HF** | **VLFpeak** | **LFpeak** | **HFpeak** |
| a | −6.1 | −6.14 | −5.41 | −6.24 | −6.24 | −5.34 | −5.69 | −5.29 |
| b | 80.83 | 92.0 | 71.42 | 101.31 | 97.86 | 70.92 | 91.64 | 99.79 |
| $R^2$ | 0.8 | 0.93 | 0.76 | 0.98 | 0.95 | 0.78 | 0.93 | 0.96 |

**Table A3.** Wilcoxon Rank Sum Test—Linear Regressions of the Statistical Percentages (subpart 3).

| | Frequency Domain | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **VLFnu** | **LFnu** | **HFnu** | **VLFpeak-nu** | **LFpeak-nu** | **HFpeak-nu** | **LF/HF** | **LFpeak/HFpeak** |
| a | −5.66 | −5.37 | −5.76 | −5.76 | −5.07 | −5.14 | −4.71 | −4.71 |
| b | 77.63 | 86.31 | 83.59 | 79.61 | 92.25 | 92.17 | 90.37 | 89.90 |
| $R^2$ | 0.84 | 0.87 | 0.87 | 0.85 | 0.89 | 0.94 | 0.85 | 0.89 |

## Appendix C

**Table A4.** Spearman's Correlation Test—Linear Regressions of the Mean Correlations (subpart 1).

| | Time Domain | | | | | | Non-Linear Domain | | | | | | Geometrical Domain | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **mNN** | **SDNN** | **SDSD** | **RMSSD** | **NN50** | **pNN50** | **ApEn** | **SD1** | **SD2** | **KFD** | **HFD** | **PTM** | **SI** | **TI** | **TINN** |
| a | −0.033 | −0.043 | −0.037 | −0.037 | −0.038 | −0.038 | −0.042 | −0.043 | −0.043 | −0.041 | −0.043 | −0.037 | −0.043 | −0.045 | −0.041 |
| b | 0.979 | 0.936 | 0.909 | 0.909 | 0.772 | 0.772 | 0.888 | 0.936 | 0.936 | 0.917 | 0.934 | 0.902 | 0.905 | 0.906 | 0.823 |
| $R^2$ | 0.99 | 0.98 | 0.97 | 0.97 | 0.90 | 0.90 | 0.96 | 0.98 | 0.98 | 0.97 | 0.97 | 0.96 | 0.97 | 0.97 | 0.93 |

**Table A5.** Spearman's Correlation Test—Linear Regressions of the Mean Correlations (subpart 2).

| | Frequency Domain | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **totPow** | **Peak** | **VLF** | **LF** | **HF** | **VLFpeak** | **LFpeak** | **HFpeak** |
| a | −0.040 | −0.043 | −0.052 | −0.039 | −0.038 | −0.052 | −0.040 | −0.043 |
| b | 1.094 | 1.066 | 1.112 | 1.091 | 1.099 | 1.105 | 1.064 | 1.065 |
| $R^2$ | 0.91 | 0.97 | 0.94 | 0.95 | 0.9 | 0.94 | 0.97 | 0.97 |

**Table A6.** Spearman's Correlation Test—Linear Regressions of the Mean Correlations (subpart 3).

| | Frequency Domain | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **VLFnu** | **LFnu** | **HFnu** | **VLFpeak-nu** | **LFpeak-nu** | **HFpeak-nu** | **LF/HF** | **LFpeak/HFpeak** |
| a | −0.054 | −0.052 | −0.046 | −0.055 | −0.049 | −0.048 | −0.046 | −0.048 |
| b | 1.11 | 1.103 | 1.099 | 1.048 | 0.88 | 1.031 | 1.101 | 1.051 |
| $R^2$ | 0.95 | 0.97 | 0.94 | 0.99 | 0.98 | 0.98 | 0.96 | 0.99 |

## References

1. Shaffer, F.; Ginsberg, J.P. Ginsberg. An overview of heart rate variability metrics and norms. *Front. Public Health* **2017**, *5*, 258. [CrossRef] [PubMed]
2. Pecchia, L.; Castaldo, R.; Montesinos, L.; Melillo, P. Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations. *Healthc. Technol. Lett.* **2018**, *5*, 94–100. [CrossRef] [PubMed]
3. Forte, G.; Favieri, F.; Casagrande, M. Heart rate variability and cognitive function: A systematic review. *Front. Neurosci.* **2019**, *13*, 710. [CrossRef] [PubMed]

4. Boardman, A.; Schlindwein, F.S.; Rocha, A.P.; Leite, A. A study on the optimum order of autoregressive models for heart rate variability. *Physiol. Meas.* **2002**, *23*, 325. [CrossRef]
5. Shaffer, F.; Meehan, Z.M.; Zerr, C.L. A critical review of ultra-short-term heart rate variability norms research. *Front. Neurosci.* **2020**, *14*, 594880. [CrossRef]
6. Zubair, M.; Yoon, C. Multilevel mental stress detection using ultra-short pulse rate variability series. *Biomed. Signal Process. Control* **2020**, *57*, 101736. [CrossRef]
7. Castaldo, R.; Melillo, P.; Bracale, U.; Caserta, M.; Triassi, M.; Pecchia, L. Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. *Biomed. Signal Process. Control* **2015**, *18*, 370–377. [CrossRef]
8. Salahuddin, L.; Cho, J.; Jeong, M.G.; Kim, D. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 22–26 August 2007.
9. Hall, J.E.; Hall, M.E. *Guyton and Hall Textbook of Medical Physiology e-Book*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2020.
10. Anonymous. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology the North American Society of Pacing and Electrophysiology. *Circulation* **1996**, *93*, 1043–1065. [CrossRef]
11. Castaldo, R.; Montesinos, L.; Melillo, P.; James, C.; Pecchia, L. Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 12. [CrossRef]
12. Nardelli, M.; Greco, A.; Bolea, J.; Valenza, G.; Scilingo, E.P.; Bailon, R. Reliability of lagged poincaré plot parameters in ultrashort heart rate variability series: Application on affective sounds. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 741–749. [CrossRef]
13. Landreani, F.; Faini, A.; Martin-Yebra, A.; Morri, M.; Parati, G.; Caiani, E.G. Assessment of ultra-short heart variability indices derived by smartphone accelerometers for stress detection. *Sensors* **2019**, *19*, 3729. [CrossRef] [PubMed]
14. Minelli, R.; Mocci, A.; Lanza, M. I know what you did last summer-an investigation of how developers spend their time. In Proceedings of the 2015 IEEE 23rd International Conference on Program Comprehension, Florence, Italy, 18–19 May 2015.
15. Weber, B.; Fischer, T.; Riedl, R. Brain and autonomic nervous system activity measurement in software engineering: A systematic literature review. *J. Syst. Softw.* **2021**, *178*, 110946. [CrossRef]
16. Baek, H.J.; Cho, C.-H.; Cho, J.; Woo, J.-M. Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. *Telemed. e-Health* **2015**, *21*, 404–414. [CrossRef] [PubMed]
17. Li, K.; Rüdiger, H.; Ziemssen, T. Spectral analysis of heart rate variability: Time window matters. *Front. Neurol.* **2019**, *10*, 545. [CrossRef] [PubMed]
18. Salahuddin, L.; Jeong, M.G.; Kim, D. Ultra short term analysis of heart rate variability using normal sinus rhythm and atrial fibrillation ECG data. In Proceedings of the 2007 9th International Conference on e-Health Networking, Application and Services, Taipei, Taiwan, 19–22 June 2007.
19. Nussinovitch, U.; Elishkevitz, K.P.; Katz, K.; Nussinovitch, M.; Segev, S.; Volovitz, B.; Nussinovitch, N. Reliability of ultra-short ECG indices for heart rate variability. *Ann. Noninvasive Electrocardiol.* **2011**, *16*, 117–122. [CrossRef]
20. McNames, J.; Aboy, M. Reliability and accuracy of heart rate variability metrics versus ECG segment duration. *Med. Biol. Eng. Comput.* **2006**, *44*, 747–756. [CrossRef]
21. Rehatschek, H.; Kienast, G. Vizard-an innovative tool for video navigation, retrieval, annotation and editing. In Proceedings of the 23rd Workshop of PVA: Multimedia and Middleware, Vienna, Austria, May 2001.
22. Stemmer, B.; Connolly, J.F. The EEG/ERP technologies in linguistic research: An essay on the advantages they offer and a survey of their purveyors. *Ment. Lex.* **2011**, *6*, 141–170. [CrossRef]
23. Niazy, R.; Beckmann, C.; Iannetti, G.; Brady, J.; Smith, S. Removal of FMRI environment artifacts from EEG data using optimal basis sets. *Neuroimage* **2005**, *28*, 720–737. [CrossRef]
24. Christov, I.I. Real time electrocardiogram QRS detection using combined adaptive threshold. *Biomed. Eng. Online* **2004**, *3*, 28. [CrossRef]
25. Yilmaz, M.; Kayancicek, H.; Cekici, Y. Heart rate variability: Highlights from hidden signals. *J. Integr. Cardiol.* **2018**, *4*, 1–8. [CrossRef]
26. Vollmer, M. A robust, simple and reliable measure of heart rate variability using relative RR intervals. In Proceedings of the 2015 Computing in Cardiology Conference (CinC), Nice, France, 6–9 September 2015.
27. Sahoo, T.K.; Mahapatra, A.; Ruban, N. Stress index calculation and analysis based on heart rate variability of ECG signal with arrhythmia. In Proceedings of the 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 22–23 March 2019; Volume 1.
28. Melillo, P.; Bracale, M.; Pecchia, L. Nonlinear Heart Rate Variability features for real-life stress detection. Case study: Students under stress due to university examination. *Biomed. Eng. Online* **2011**, *10*, 96. [CrossRef] [PubMed]
29. Katz, M.J. Fractals and the analysis of waveforms. *Comput. Biol. Med.* **1988**, *18*, 145–156. [CrossRef]
30. Higuchi, T. Approach to an irregular time series on the basis of the fractal theory. *Phys. D Nonlinear Phenom.* **1988**, *31*, 277–283. [CrossRef]
31. Silver, N.C.; Dunlap, W.P. Averaging correlation coefficients: Should Fisher's z transformation be used? *J. Appl. Psychol.* **1987**, *72*, 146. [CrossRef]

2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON) | 978-1-6654-4280-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/MELECON53508.2022.9842890

# Impact of Ultra-short-term HRV Features in Software Code Sections Complexity Classification

André Bernardes[1], Ricardo Couceiro[1], Júlio Medeiros[1], Jorge Henriques[1],
César Teixeira[1], João Durães[1,2], Henrique Madeira[1], Paulo Carvalho[1]

*Abstract*—**Ultra-short term HRV features are becoming increasingly popular due to the fact that they do not need long time periods for their assessment and, therefore, can be used in nearly real-time cognitive load assessment scenarios where the standard 1-min to 5-min time frames are not applicable. Several authors focused on the assessment of the validity of these features by comparing them to the accepted 5-min short term features, showing that the accuracy of these features decreases with the decrease of the analysis window length. However, there is one question that, to the best of our knowledge, has not been fully addressed yet. How does the reduction of the analysis window affect the classification process during cognitive demanding tasks? In this paper we propose the use of 18 different time frames, ranging from 3 minutes to 10 seconds, to extract HRV features from data collected out of 21 subjects during code comprehension tasks. The HRV features are then associated with a code section, gazed during an experiment run, and statistical transformations are computed to built the several datasets, where each section gazed is a sample. A Support Vector Machine (SVM) classifier was trained for each different dataset using a Leave-One-Subject-Out cross-validation procedure, following 3 distinct approaches. The classifier's goal is to discriminate between low and high complexity code sections analysed by the subjects during the experiment. The F1-Scores ranged from 0.79 to 0.64, indicating that it's possible to achieve similar, but lower classification results using smaller time frames, with a consistent increase of the variability in the performance evidenced by a higher standard deviation of F1-Scores in the smaller time frames.**

*Index Terms*—**time frames, sliding window, feature extraction procedure, HRV features, features statistical transformations, SVM classifier**

## I. Introduction

One of the most demanding tasks in software engineering is related to the detection of S/W bugs, either during software programming, or during code inspection tasks. These bugs typically occur in the more complex sections of the code, which demands a higher cognitive state to their comprehension. Programmers' difficulties in comprehending specific code sections do not necessarily map to classic software complexity metrics [1] [2], which justifies the quest for a neural link between programmers' cognitive states and potential code comprehension difficulties. Possible tools capable of identifying programmers' code comprehension difficulties in real-time (i.e., while the programmer is developing the code or doing a code review) would be essential to assist programmers during the code analysis/inspection process, warning them about code sections with high probability of having bugs. Knowing that code comprehension and bug detection occupies around 70% of the programmer's working time [3], it is easy to understand the relevance of identifying code programmers' comprehension difficulties to improve code reliability and reduce software development cost.

The Central Nervous System (CNS), composed by the brain and the spinal cord, is constantly exchanging information with all the body parts, and producing responses to every stimulus received. The Peripheral Nervous System (PNS) is the system responsible for sustaining these signal exchanges. A subpart of the PNS is the Autonomic Nervous System (ANS) which is divided in the sympathetic nervous system, accountable for being in control during stressful events, and the parasympathetic nervous system, which is responsible for the restoration processes to bring the body towards a stable state [4]. That said, the CNS is capable of influencing all the physiological systems, including the ANS, and therefore it is possible to measure the cognitive manifestations through the analysis of bio-signals that capture the activity of the ANS.

One of these bio signals is the Heart Rate Variability (HRV). HRV consists of the variation in the duration of the intervals between successive heartbeats (R-R intervals) [5], measured in milliseconds (ms). These intervals are believed to behave like an index of the autonomic control [6], since they are influenced by the dynamic interaction between the parasympathetic and the sympathetic systems signals delivered to the heart (via the sinoatrial node) [7] and have been referenced has having a good correlation with the cognitive load [7].

In this study we propose to analyse the performance of multiple classifiers fed with statistical transformations of standard HRV features, computed at different time-resolutions, i.e., using time frames from 3min to 10sec. The goal of these classifiers is to discriminate between low and high complexity code sections. By testing different time frame sizes, we expect to observe its influence in the features extracted and consequently in the classifier results, even for ultra-short HRV required for real-time programmers' support.

It is important to highlight the use of features' statistical transformations, which are associated to different complexity

[1]A. Bernardes, R. Couceiro, J. Medeiros, P. Carvalho, J. Henriques, C. Teixeira, H. Madeira, J. Duraes, are with with the Centre for Informatics and Systems of the University of Coimbra, Polo II, University of Coimbra, 3030-290, Coimbra, Portugal, (e-mails: ambernardes11@gmail.com, {rcouceir, juliomedeiros, carvalho, jh, cteixei, henrique}@dei.uc.pt and jduraes@isec.pt).

[2]J. Duraes is with the Polytechnic Institute of Coimbra, 3045-093, Coimbra, Portugal, (e-mail: jduraes@isec.pt).

Fundação para a Ciência e Tecnologia (FCT) under the national project BASE (POCI - 01-0145 - FEDER- 031581)

code sections. The statistical transformations are performed after the concatenation of the standard HRV features, extracted during the periods when the programmer gazed at a certain code section. Following this procedure makes our study conclusions specific to code inspection or similar contexts and, consequently, not as generalizable as the prevailing literature, which methods do not involve statistical transformations.

The rest of the paper is organized as follows: Section II provides the description of the data collection procedure, section III details the proposed approach for the analysis, and section IV presents the description of the classification process. Finally, in section V, the results are outlined and section VI summarizes the main conclusions and future work.

## II. Data Collection

For the development of the current analysis a dataset collected during the BASE project has been used. This dataset contains biometrical signals including Electrocardiogram (ECG), the Photoplethysmogram (PPG), the Electrodermal activity (EDA) and eye movements (Eye Tracking). In this paper we will focus on the analysis of the HRV obtained from the ECG signals during the time periods associated to the code sections analysed by the subjects.

The dataset contains data from 21 programmers with different levels of expertise in C language. All the subjects selected for the experiment were male with ages between 19 and 40 years, and an average age of 22 years old.

The protocol to which the selected subjects were submitted consisted of 4 different runs that always started with a fixation cross positioned in the center of the screen, followed by the next three tasks:

- 1) Task consisting in a natural language reading. Four small literary excerpts were used here and randomly selected each run. Estimated duration of the task: 60 seconds
- 2) Task consisting in a neutral code reading. The four code snippets employed here were selected at random and were bug free and straightforward. Estimated duration of the task: 300 seconds
- 3) Task consisting in code inspection and bug detection. In this task 4 code snippets written in C language were used; these were selected randomly at each run as well. Estimated duration of the task: 600 seconds

Between tasks, a fixation cross was presented to the subjects. In order to reduce any bias in the obtained results, the three tasks were performed in a random order at the different runs.

A written consent for the data collection was signed by the subjects that took part in the study. The study, in accordance with the Declaration of Helsinki, received clearance by the Ethics Committee of the Faculty of Medicine of the University of Coimbra.

## III. Methods

### A. ECG - Pre-Processing

Given that the experiment was conducted with simultaneous acquisition of MRI, an initial pre-processing was necessary to remove the gradient artifact (GA) induced by the MRI-scanner on the ECG signals. To this end, an average artifact subtraction (AAS) technique based on the algorithm from Niazy et al. [8] wwas performed to reduce this artifact on the ECG data. Beside the GA correction, the ECG by being recorded inside a MRI-scanner, the recorded signals presents some changes in its morphology due to the magnetic field. Therefore, traditional QRS detection algorithms tend to fail and, consequently, lead to a bad R-R intervals calculation. Nevertheless, the R-peak detection algorithm proposed by Christov et al. [9] is commonly used in these scenarios, given its robustness and high performance in the R peak detection on ECG signals recorded inside MRI-scanners. After having the R-peaks detected, we proceeded to the computation of the R-R intervals to obtain the HRV time-series.

### B. Features Extraction

For the HRV analysis, a thorough survey of the most suitable features, i.e., the most reliable extracted using small time frames (Ultra-short HRV measurements) has been conducted, leading to the selection of a set of 31 features across Time, Geometric, Non-Linear and Frequency Domains [10]–[16]. The set of features used in this paper are summarised in the table I.

Starting from the HRV signals corresponding to the data collected during the code inspection and bug detection task, we applied a sliding window with a step of 1 sec. for the feature extracting process, giving origin to a vector of individual measurements, associated to an instant in time, for each feature. Every single measurement value is calculated based on a RR signal portion with the size of the sliding window used. The time instant associated with that individual measurement will be the instant corresponding to the centre of the RR signal portion used on the computation.

The feature extraction step was repeated using 18 different sliding windows, creating 18 different datasets. The first features were extracted using a sliding window of 180 seconds, which was iteratively reduced until 10 seconds, the smallest window used in the current analysis.

### C. Features Normalization

All the features extracted during the "code inspection and bug detection" task were normalized regarding the "natural language reading" periods. With this intent, the features mentioned in the table I were also extracted from the data collected during the "natural language reading" periods, using a similar procedure as the explained in the previous section (to facilitate the reference let's call these ones the 'rest features', and the others the 'code features'). In the "natural language reading" period the subjects are supposed to be in a low cognitive stress state, which corresponds to an optimal state for the normalization process. The normalized features were obtained by calculating the ratio between each "code feature" and the median of the corresponding "rest feature". Here, the median has been selected since the data does not follow a normal distribution (assessed using the Kolmogorov–Smirnov test).

TABLE I

SET OF FEATURES USED ON THE STUDY.

| Initials | HRV Measurments |
|---|---|
| time domain | |
| mNN | mean of RR intervals |
| SDNN | standard deviation of RR intervals |
| SDSD | standard deviation of the differences between heart beats |
| RMSSD | root mean square of the differences between heart beats |
| NN50 | number of RR intervals that fall within 50 milliseconds |
| pNN50 | percentage of RR intervals that fall within 50 milliseconds |
| | References: [6] [10] |
| geometric domain | |
| TI | HRV Triangular Index |
| TINN | Triangular Interpolation of RR interval Histogram |
| SI | The Baevsky's Stress Index |
| | References: [6] [10] [11] [12] [13] |
| non-linear domain | |
| ApEn | Approximate Entropy |
| SD1 | standard deviation of the Poincare' plot perpendicular to the line-of-identity |
| SD2 | standard deviation of the Poincare' plot along the line-of-identity |
| PTM | Point Transition Measure |
| KFD | Katz Fractal Dimension |
| HFD | Higuchi Fractal Dimension |
| | References: [4] [6] [14] [15] [16] |
| frequency domain | |
| VLF | very-low frequency band power (<=0.04Hz) |
| LF | low frequency band power (0.04 - 0.15 Hz) |
| HF | very-low frequency band power (0.15 - 0.4 Hz) |
| VLFnu | VLF power normalised |
| Lfnu | LF power normalised |
| HFnu | HF power normalised |
| VLFpeak | VLF power frequency peak |
| LFpeak | LF power frequency peak |
| HFpeak | HF power frequency peak |
| VLFpeak-nu | VLF power frequency peak normalised |
| LFpeak-nu | LF power frequency peak normalised |
| HFpeak-nu | HF power frequency peak normalised |
| totPow | Total Power |
| Peak | Overall frequency power peak |
| LF/HF | Ratio of LF and HF band powers |
| LFpeak/HFpeak | Ratio of LF and HF band power frequency peak |
| | References: [6] [10] |

### D. Features Transformation

As mentioned before, the code snippets where the subjects search for bugs have sections with different complexities, labelled as low and high complexity, according to a classification performed by a panel of experts. The individual measurements from the feature vectors, produced in the feature extraction process, associated with all instants corresponding to the periods that the subject was looking to a certain section during a run, were grouped. From each group, a set of features was computed using simple statistic transformations, i.e., mean; standard deviation; maximum; minimum; median; quantile 0,50; quantile 0,75; quantile 0,85; quantile 0,95; peaks mean; peaks standard deviation; peaks maximum; peaks minimum; peaks median; peaks quantile 0,50; peaks quantile 0,75; peaks quantile 0,85; peaks quantile 0,95; and peaks rate. The scheme of the statistical transformations method used is presented in the figure 1.

In summary, 589 features (31 features x 19 statistical transformations) were extracted from the 18 datasets (one dataset for each different sliding window size used in the initial extraction) labelled according to its difficulty.

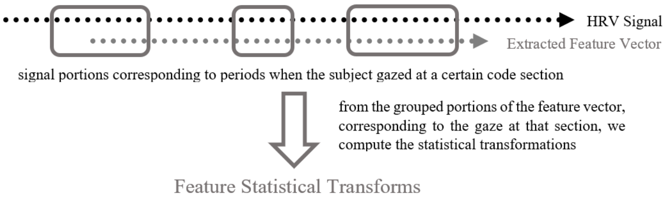

HRV Signal
Extracted Feature Vector
signal portions corresponding to periods when the subject gazed at a certain code section
from the grouped portions of the feature vector, corresponding to the gaze at that section, we compute the statistical transformations
Feature Statistical Transforms

Fig. 1. Feature statistical transformations scheme.

## IV. CLASSIFICATION

### A. Feature selection and approach

The goal of the classification is to discriminate between low and high complexity code sections. To reduce the number of features by dataset, we performed the Kruskal-Wallis test for each transformed feature divided in two groups by the complexity label. This algorithm returns the p-value for the null hypothesis that both groups come from the same distribution. The rejection of the null hypothesis was considered when the p-value was below 0.05.

From this procedure it was observed that the datasets resulting from the use of the smaller sliding windows, had a larger amount of discriminative transformed features when compared to the datasets corresponding to the larger windows. Thus, in order to produce a fair comparison between the classification models, three different approaches were used:

- **Approach 1)** Selection of the 15 most discriminative transformed features with the lowest p-values (in the Kruskal-Wallis analyse) from the 180 seconds sliding window dataset;
- **Approach 2)** Selection of the 15 most discriminative transformed features with the lowest p-values from the 10 seconds sliding window dataset;
- **Approach 3)** Selection of the 15 most discriminative transformed features, i.e., with the lowest p-values, for each window size dataset.

The dataset produced by each of these approaches was used to train and test a SVM classifier, with linear kernel and a regularization parameter of 2, allowing us to evaluate the influence of the window size in the different classifier results.

### B. Classifier Train and Test

The classification process was performed using a Leave-One-subject-Out cross validation scheme using each 18 datasets from our 3 approaches.

To access the performance of each classifier, the F1 Score was computed for training and test in each loop of the Leave One Out technique, resulting a vector of F1 Scores for each window size dataset (with the size equal to the number of subjects). The F1 Score assigned to each dataset was the mean of all repetitions using that window size dataset, the standard deviation was also computed. The data distribution analysis of each resulting F1 Score vector was done using the Kolmogorov-Smirnov test, and it was possible to conclude that non exhibited a normal distribution. Therefore, the Wilcoxon rank sum test was used to verify the differences in the F1 Score distributions from each window size dataset against the 180 second dataset, assumed as reference. The set were considered to be from different distributions for a p-value below 0,05 (null hypothesis rejected). These steps were repeated for the 3 different approaches.

## V. Results and discussion

First, it should be underlined in the discussion the reason why, in this context, we opted to use statistical transformation of the HRV features, instead of using directly the HRV features. When a programmer is inspecting code, with multiple sections of different complexities, it is expected for him to go back and forward through the sections, searching for code flaws and bugs. This fact makes that a single section could be gazed several times, also in each gaze the cognitive load may be different. To illustrate this logic, we may think on the following example: a programmer inspects a certain section for the first time and does not feel any difficult, although through the code inspection continuation he has a doubt and suspects that the answer to his doubt may be in a previous section, so he goes back to that initial section, he may now feel a higher difficulty there, where previously that level of difficulty was not felt.

With the use of statistical transformations, each sample on the dataset built corresponds to a certain section gazed during a run. The importance of this step is to capture and enhance the state of the subject on each specific code section over the experiment. The use of different size sliding windows in the HRV features extraction (before the transformations) is proposed since different time frames may access different ANS dynamics. Furthermore, the use of smaller time frames may provide greater granularity in the analysis, making it possible to capture the ANS dynamics during smaller code sections gazing periods. This way, the analysis done in this paper reinforces the use of Ultra-Short-term features, more specifically which time frames are more suited to employ in the features extraction process in this specific software development or similar contexts.

The results achieved from the current analysis are presented in figure 2. In the table II we present the results of the Wilcoxon rank sum test analysis of the F1 Score distributions.

Through the analysis of Figure 2, it is possible to observe that the three different approaches achieved similar results. In approach 1 the F1-Scores remains particularly stable until the 60 seconds time frame and for time frames below 60 seconds it is observed that the F1-Scores are not from the same distribution as the reference F1-scores, i.e., the 180 seconds dataset F1-Scores. It is also possible to observe that the performance of the classifiers did not change substantially, which is confirmed by the range of the F1 scores between 0.64 and 0.75. The largest difference found across the results are in the standard deviation values, which are significantly larger in the results related to the use of smaller time frames. The best result in this approach is the result correspondent to the use of the 180 seconds sliding window. This is in accordance to the expected outcome, since all the datasets were based in the most discriminative transformed features of the 180 seconds time frame.

Analysing the second approach, it is possible to notice an increase in performance when using the datasets obtained with the smaller window sizes. One possible reason for this observation is that in this approach the selected features were the most discriminative assessed by the Kruskal Wallis test for the 10 seconds dataset. Relatively to the first approach, there was an increase in the performance of the classifiers fed with features extracted from the smaller time frames, although the classifier based on the 10 seconds decreased its score value. The score correspondent to the 160 seconds dataset with this approach was the best of the overall study, fact that was not initially expected, but can be explained by some compensatory behaviour between transformed features. Finally, regarding the third approach, the conclusions taken are not much different from the other two. This approach had the most discriminative features for each dataset and therefore it is expected that the different time frames classifiers presented the higher F1-Score attributed from the 3 approaches, and with the lower variability (standard deviation). Nonetheless, one can observe some exceptions that are believed to be the result of the before mentioned compensatory behaviour of the different set of transformed features.

## VI. Conclusion

In this paper we presented the analysis of several classifiers following three distinct approaches. The classifiers were fed by statistical transformations of standard HRV features extracted with multiple time frames from 180 secs. to 10 secs. and the F1-Scores regarding each classifier were computed.

Considering the higher F1-Score values of each time frame across these approaches, it is possible to conclude that the reduction of the time frame used for the extraction of the HRV features, before the statistical transformations method application, does not affect substantially the results obtained in the classification process. This is a relevant result since most published results regarding the window size impact on HRV as well as our own results suggest that the lower the window size the higher is the uncertainty of the extracted HRV feature and the lower its correlation with respect to the feature extracted using a large window (e.g. the standard 5 minute duration window). However, in the study context, when applied in a classifier, paired with the proposed statistical transformations method in the dataset construction, the obtained results show that the performance degradation is much smaller. This might be due to the complementary nature of the feature which is able to compensate the uncertainty in each feature, i.e., these interesting results are thought to be the consequence of compensatory effects between features, since although individually some sets of transformed features were selected in the feature selection process in a certain window size dataset (approach 3), in some cases other different set of features ended up having the best F1-Scores. This compensatory behavior seems also

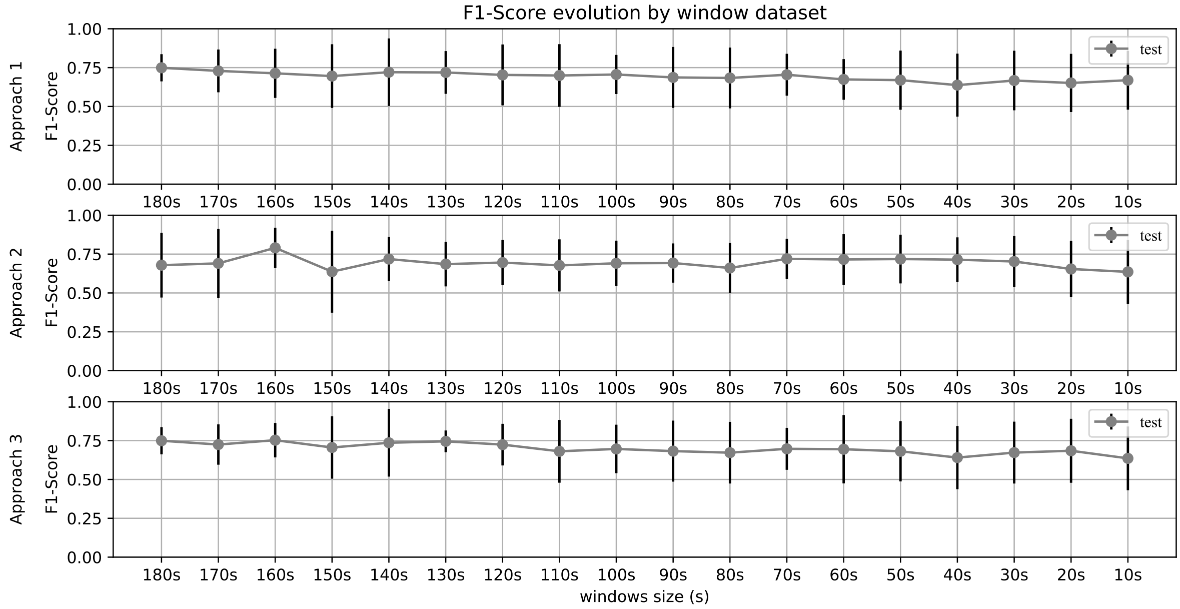


Fig. 2. Mean F1-Scores and Standard Deviations plot obtained using the datasets based on the diferent time frames in study

TABLE II
WILCOXON RANK SUM TEST ANALYSIS OF THE F1 SCORE DISTRIBUTIONS COMPARED TO THE 180 SECONDS DATASET

| | 180s | 170s | 160s | 150s | 140s | 130s | 120s | 110s | 100s | 90s | 80s | 70s | 60s | 50s | 40s | 30s | 20s | 10s |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Approach 1 | Accepted<br>p=1.00 | Accepted<br>p=0.95 | Accepted<br>p=0.77 | Accepted<br>p=0.79 | Accepted<br>p=0.98 | Accepted<br>p=0.68 | Accepted<br>p=0.38 | Accepted<br>p=0.51 | Accepted<br>p=0.18 | Accepted<br>p=0.23 | Accepted<br>p=0.22 | Accepted<br>p=0.26 | Rejected<br>p=0.02 | Rejected<br>p=0.04 | Rejected<br>p=0.01 | Rejected<br>p=0.03 | Rejected<br>p=0.01 | Rejected<br>p=0.02 |
| Approach 2 | Accepted<br>p=1.00 | Accepted<br>p=1.00 | Accepted<br>p=0.10 | Accepted<br>p=0.90 | Accepted<br>p=0.74 | Accepted<br>p=0.63 | Accepted<br>p=0.87 | Accepted<br>p=0.98 | Accepted<br>p=0.77 | Accepted<br>p=0.95 | Accepted<br>p=0.51 | Accepted<br>p=0.88 | Accepted<br>p=0.65 | Accepted<br>p=0.93 | Accepted<br>p=0.84 | Accepted<br>p=0.79 | Accepted<br>p=0.60 | Accepted<br>p=0.42 |
| Approach 3 | Accepted<br>p=1.00 | Accepted<br>p=0.61 | Accepted<br>p=0.71 | Accepted<br>p=0.60 | Accepted<br>p=0.57 | Accepted<br>p=0.69 | Accepted<br>p=0.63 | Accepted<br>p=0.34 | Accepted<br>p=0.35 | Accepted<br>p=0.18 | Accepted<br>p=0.20 | Accepted<br>p=0.15 | Accepted<br>p=0.69 | Accepted<br>p=0.13 | Rejected<br>p=0.01 | Accepted<br>p=0.07 | Accepted<br>p=0.18 | Rejected<br>p=0.04 |

to be present across the different time frames since the F1-Score values hold relatively stable with the decrease of the window sizes. To further explore these compensatory effects a possible approach may be applying a Principal Component Analysis to the features, under penalty of losing some of its interpretability. Another explanation to the obtained results stability, may be the statistical transformation method used in the dataset construction. This procedure reduces the probability of, in a sample, the programmer difficulty sensation being different from the actual labeled complexity of the code section gazed. With this method implementation, it is possible to account every time that a certain section is gazed, and the different difficulty levels felt at each gaze, in each sample, helping to further extend the compensatory behavior discussed. Nevertheless, the standard deviation of F1-Scores obtained with the cross-validation method used, reveal higher variability in the smaller time frame datasets.

Considering the use case reported in this paper where the extraction of HRV features under ultra-short time periods is vital to capture fine events such as the inspection of short, but complex code sections, a classifier fed with statistical transformations of features extracted using different time frames could be an optimal solution. The conclusions presented in this paper are context specific and should be carefully analysed and further studied. Furthermore, another limitation of this work is that the experiment design protocol does not account for daytime HRV variations neither for longer rhythms dynamics, which can not be captured by ultra-short HRV measurements.

## ACKNOWLEDGMENT

The authors acknowledge the financial support by FCT under the national project BASE (POCI - 01-0145 - FEDER-031581).

# References

[1] Ricardo Couceiro, Raul Barbosa, Joáo Duráes, Gonçalo Duarte, Joáo Castelhano, Catarina Duarte, Cesar Teixeira, Nuno Laranjeiro, Júlio Medeiros, Paulo Carvalho, et al. Spotting problematic code lines using nonintrusive programmers' biofeedback. In *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, pages 93–103. IEEE, 2019.

[2] Ricardo Couceiro, Gonçalo Duarte, João Durães, João Castelhano, Catarina Duarte, César Teixeira, Miguel Castelo Branco, Paulo Carvalho, and Henrique Madeira. Biofeedback augmented software engineering: monitoring of programmers' mental effort. In *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pages 37–40. IEEE, 2019.

[3] Roberto Minelli, Andrea Mocci, and Michele Lanza. I know what you did last summer-an investigation of how developers spend their time. In *2015 IEEE 23rd International Conference on Program Comprehension*, pages 25–35. IEEE, 2015.

[4] Muhammad Zubair and Changwoo Yoon. Multilevel mental stress detection using ultra-short pulse rate variability series. *Biomedical Signal Processing and Control*, 57:101736, 2020.

[5] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017.

[6] Leandro Pecchia, Rossana Castaldo, Luis Montesinos, and Paolo Melillo. Are ultra-short heart rate variability features good surrogates of short-term ones? state-of-the-art review and recommendations. *Healthcare technology letters*, 5(3):94–100, 2018.

[7] Giuseppe Forte, Francesca Favieri, and Maria Casagrande. Heart rate variability and cognitive function: a systematic review. *Frontiers in neuroscience*, 13:710, 2019.

[8] Rami K Niazy, Christian F Beckmann, Gian Domenico Iannetti, J Michael Brady, and Stephen M Smith. Removal of fmri environment artifacts from eeg data using optimal basis sets. *Neuroimage*, 28(3):720–737, 2005.

[9] Ivaylo I Christov. Real time electrocardiogram qrs detection using combined adaptive threshold. *Biomedical engineering online*, 3(1):1–9, 2004.

[10] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.

[11] Tanmay Kumar Sahoo, Ashutos Mahapatra, and Nersisson Ruban. Stress index calculation and analysis based on heart rate variability of ecg signal with arrhythmia. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, volume 1, pages 1–7. IEEE, 2019.

[12] Mücahid Yılmaz, Hidayet Kayançiçek, and Yusuf Çekici. Heart rate variability: Highlights from hidden signals. *J. Integr. Cardiol*, 4:1–8, 2018.

[13] Marcus Vollmer. A robust, simple and reliable measure of heart rate variability using relative rr intervals. In *2015 Computing in Cardiology Conference (CinC)*, pages 609–612. IEEE, 2015.

[14] Paolo Melillo, Marcello Bracale, and Leandro Pecchia. Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination. *Biomedical engineering online*, 10(1):1–13, 2011.

[15] Tomoyuki Higuchi. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):277–283, 1988.

[16] Michael J Katz. Fractals and the analysis of waveforms. *Computers in biology and medicine*, 18(3):145–156, 1988.