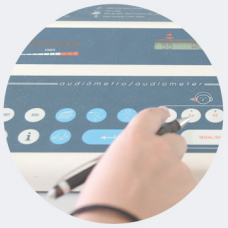
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA BIOMÉDICA






iMind: Uma ferramenta inteligente para suporte de compreensão de conteúdo

MIGUEL DUARTE SOUSA GOMES
Outubro de 2022

P.PORTO

iMind:

An intelligent tool to augment content comprehension

Miguel Duarte Sousa Gomes

With Bachelor's degree of Biomedical Engineer by the Superior Institute of Engineering of Porto

"Thesis to be presented at the Superior Institute of Engineering of Porto to obtain the Master's degree in Biomedical Engineering"

Supervisors:

Eng. Haytham Hijazi, (DEI-FCTUC & CISUC)

Prof. Dr. Isabel Praça (DEI-ISEP & GECAD)

[10/2022]

“The essence of knowledge consists in applying it, once possessed.”

Confucius

Acknowledgment

I would like to thank Engineer Haytham Hijazi and Professor Henrique Madeira for their sympathy, patience, and excellent guidance throughout this process. Thank you for all the care and availability that you have guaranteed throughout this school year. Thanks for the help, learning and motivation.

I would also like to thank the BASE project, Biofeedback Augmented Software Engineering (POCI - 01-0145 - FEDER - 031581) and FCT, for the financial support provided in the development of this work.

Abstract

Usually while reading, content comprehension difficulty affects individual performance. Comprehension difficulties, e. g., could lead to a slow learning process, lower work quality, and inefficient decision-making. This thesis introduces an intelligent tool called “iMind” which uses wearable devices (e.g., smartwatches) to evaluate user comprehension difficulties and engagement levels while reading digital content. Comprehension difficulty can occur when there are not enough mental resources available for mental processing. The mental resource for mental processing is the cognitive load (CL). Fluctuations of CL lead to physiological manifestation of the autonomic nervous system (ANS), which can be measured by wearables, like smartwatches. ANS manifestations are, e. g., an increase in heart rate. With low-cost eye trackers, it is possible to correlate content regions to the measurements of ANS manifestation. In this sense, iMind uses a smartwatch and an eye tracker to identify comprehension difficulty at content regions level (where the user is looking). The tool uses machine learning techniques to classify content regions as difficult or non-difficult based on biometric and non-biometric features. The tool classified regions with a 75% accuracy and 80% f-score with Linear regression (LR). With the classified regions, it will be possible, in the future, to create contextual support for the reader in real-time by, e.g., translating the sentences that induced comprehension difficulty.

Keywords: Biometrics measurement, cognitive load, content comprehension, eye-tracking, Heart rate variability, machine learning.

Resumo

Normalmente durante a leitura, a dificuldade de compreensão pode afetar o desempenho da leitura. A dificuldade de compreensão pode levar a um processo de aprendizagem mais lento, menor qualidade de trabalho ou uma ineficiente tomada de decisão. Esta tese apresenta uma ferramenta inteligente chamada "iMind" que usa dispositivos vestíveis (por exemplo, smartwatches) para avaliar a dificuldade de compreensão do utilizador durante a leitura de conteúdo digital. A dificuldade de compreensão pode ocorrer quando não há recursos mentais disponíveis suficientes para o processamento mental. O recurso usado para o processamento mental é a carga cognitiva (CL). As flutuações de CL levam a manifestações fisiológicas do sistema nervoso autônomo (ANS), manifestações essas, que pode ser medido por dispositivos vestíveis, como smartwatches. As manifestações do ANS são, por exemplo, um aumento da frequência cardíaca. Com *eye trackers* de baixo custo, é possível correlacionar manifestação do ANS com regiões do texto, por exemplo. Neste sentido, a ferramenta iMind utiliza um smartwatch e um *eye tracker* para identificar dificuldades de compreensão em regiões de conteúdo (para onde o utilizador está a olhar). Adicionalmente a ferramenta usa técnicas de *machine learning* para classificar regiões de conteúdo como difíceis ou não difíceis com base em *features* biométricos e não biométricos. A ferramenta classificou regiões com uma precisão de 75% e *f-score* de 80% usando regressão linear (LR). Com a classificação das regiões em tempo real, será possível, no futuro, criar suporte contextual para o leitor em tempo real onde, por exemplo, as frases que induzem dificuldade de compreensão são traduzidas.

Keywords: Biometrics measurement, cognitive load, content comprehension, eye-tracking, HR variability, machine learning.

Contents

List of Figures

List of Tables

List of abbreviations

ANOVA – Analysis of Variance

ANS – Autonomic Nervous System

BVP – Blood Volume Pulse

CL – Cognitive Loads

CLT – Cognitive Load Theory

CNS – Central Nervous System

ECG – Electrocardiogram

EEG – Electroencephalogram

EDA – Electrodermal Activity

HR – Heart Rate

HRV – Heart Rate Variability

PNS – Peripheral Nervous System

pNN50 – Percentage of successive RR intervals bigger than 50 ms

PPG – Photoplethysmography

PRV – Pulse Rate Variability

RF – Random Forest

ROC – Receiver Operating Characteristic

SNS – Somatic Nervous System

*Additionally, the abbreviation used for HRV and EDA features are in Tables 3 and 4 and the abbreviations for the Classifiers used are in Table 5.

Chapter 1 – Introduction

1. Introduction

1.1. Context

Commonly, when reading an English passage, we may not fully understand an expression, a sentence, or region of content at first glance. Difficulties in comprehending content can hinder professional work or learning as it increases the time necessary to read and analyze information. Those difficulties can occur because our mind is not focused on that task or because the concepts behind that region are too complex for us at that point in the day. With measures of mental capacity, it would be possible to improve tasks that are dependent on reading content.

Cognitive load (CL) is the mental resource used for mental processing. CL has a limited capacity as such the mental ability to comprehend content is limited as well. Changes in CL can lead to physiological manifestations. This physiological manifestation comes mainly from the ANS. ANS is composed of the sympathetic and parasympathetic branches, which are antagonistic to each other. Under mentally challenging situations, the sympathetic nervous system increases the Heart rate (HR), while the parasympathetic branch decreases the HR in calm situations.

Considering that psychological states [1], [2] can affect the physiological manifestations of ANS, studies tend to use a multimodal approach (i.e., integrating more than one biometric signal) to further distinguish CL [3]. HR biometric features can be obtained with electrocardiogram (ECG), but the use of a smartwatch has a more practical use for daily life scenarios and applications (although they are less precise). The incorporation of wearables to measure the manifestation of ANS (like a smartwatch) in assessing CL, represent a small but emerging topic in the literature as the technology of wearables is growing enormously.

To identify which content caused the comprehension difficulties, the synchronization of the eye-movements on the content and biometrics (from wearables) is necessary. With eye trackers, is possible to extract the location of the screen which the user is looking at, also known as eye gaze. If the displayed content changes its position, on the screen, that means the user is looking at another region of content. To synchronize the content position with the position in which the user is looking is necessary to correlate the eye gaze with, e. g., the page scroll.

In addition to multimodal approach studies also incorporate the use of machine learning (ML) to help identify the physiological patterns associate with CL. With the incorporation of ML is possible to identify which regions are difficult by the prediction of high CL, mainly influenced by comprehension difficulties. In future work, the identified regions of comprehension difficulties can be used for content feedback.

1.2.Motivation

According to O'Rourke [4], the struggles in learning a foreign language are related to difficulties in lexicon structure, grammatical and syntactic rules, and also sociologic-communicative factors.

Comprehension difficulty is not exclusive to learning spoken languages, but also programming languages (e.g., JAVA). The skills involved in learning a programming language are similar to the ones engaged in learning a spoken language as both skills involve lexical and syntactic rules ad as both present a set of constraints and structural rules that need to be understood.

Code learning can differ from spoken language learning as arithmetic logic can be a key point of its comprehension [5]. Jens et al. [6] reported according to UNESCO data, the failure rate of introductory programming courses progress from 2007 to 2019 as 33% to 29% failing rate. Jens et al. justified that these percentages are not as high compared to other failure rates like college algebra in the US in those years (42 to 50%), but still are significant.

Peter et al. [7] described an 18% to 21% failure rate in the first semester of an English degree in Japanese Universities in 2007. According to Peter et al. [7], this degree teaches English mainly as a first foreign language. Even though these percentages are smaller than in programming languages still show the struggle exhibited by a student in learning a foreign language. Another example that also reveals these struggles, even with more cultural and language proximity, is Europe. In Europe, the majority of countries mandate, through legislation, that students must learn at least two foreign languages (English being the most learned with 5 to 7 mandatories years). Despite this, some European countries with moderate English proficiency, like Italy and Spain, present results of 535 and 540 points, corresponding to only B2 classification in the CEFR score

(Common European Framework of Reference for Languages) [8]. Since 5 to 7 years of English in the education system result in medium proficiency, maybe incorporating a tool to help language learning is a relevant path to a more efficient education.

Nowadays, most content presents a digital version or is solely presented as digital content. Using digital content in learning and work environments opens the possibility for multiple uses for digital content tools. During the spread of the SARS-CoV-2 between 2020 and 2021, multiple confinements were imposed, which pushed the educational system to remote classes. Del Arco et al. [9] described that, in universities that were less adapted to online classes, there was a decrease in communication with students, which could be suppressed with a remote assessment of student comprehension.

The use of support tools in the language learning process can be incorporated not only for a study supported by teachers but also for self-taught. Self-taught students are more prone to recurring errors when learning [10] and could potentially benefit more from a learning support tool.

One of the crucial motivations behind this work is the availability of affordable and precise wearables and biometric sensors. E. g., smartwatches and desktops low-cost are good technologies to use in assessing individual comprehension difficulties in learning and work environments. Nonetheless, very few studies have used them in CL classification at region-level (evaluating different regions of content) [11].

1.3. Objective

This thesis aims to develop an intelligent tool, iMind, capable of assessing individuals' comprehension and engagement levels by classifying CL using smartwatches and low-cost desktop eye trackers.

Content comprehension assessment at the region content levels occurs by integrating multimodal measures (HR and Electrodermal Activity (EDA) measurements to assess CL) and using an eye-tracker to identify content regions associated with high CL.

1.4. Contributions

The main contributions from this thesis are as follows:

- Development of an intelligent tool that measures CL to predict content comprehension difficulties.
- Development of a new computational method to synchronize a multimodal set of physiological data with the eye-tracker.
- Comparison of different classifiers models to assess content comprehension using wearables.
- Synchronization of content location (page scroll) to achieve fine-grained analysis of content regions over time.

1.5. General outline

This document will present the details of the study of measuring CL through the analyses of multiple biometric features, their acquisition and their physiological meaning. Chapter 2 will present the background of biometric features, their relation to the CL, and related work. Chapter 3 presents the methodology with the tool architecture and implementation. Chapter 4 presents the protocol, dataset and data analysis. And chapter 5 presents the main results taken from the features selection and ML pipeline. Lastly, chapter 6 describes the main conclusion, contribution and future work regarding this thesis.

CHAPTER 2 – BACKGROUND AND RELATABLE WORK

2. Background and relatable work

The goal of this thesis is to develop a tool to classify content comprehension using biometric data. Biometrics are related to the physiological manifestation of changes in ANS induced by different factors such as mental effort. That data will be synchronized with eye-tracking to identify the content regions that are associated with high CL (potentially high difficulty in comprehension). This chapter presents the background related to CL and is related to the ANS and how its manifestation can be measured with biometric signals. After that describes the state of the art of measurements of CL changes by biometric features.

2.1. Cognitive Load

CL represents the mental resources available in mental processing [3]. The mental resource of CL changes can be decomposed into long-term and working memory. Long-term memory is responsible to store and organize information and working memory is related to the processing of it. Considering that the working memory is limited in capacity, the same can be extrapolated to CL, which means that CL has a limited capacity, which is used for managing all mental tasks [12].

It was proposed by Newell et al. [13] in 1980, that comprehension is the mental processing of data by CL, and it can be decomposed into a model called Cognitive load theory (CLT). This model describes that initially our mind interpretation raw input independently. After that, the understanding of the context of what the subject is doing/reading is created. Only by understanding the content is the individual capable of executing a course of action related to the context, e. g., selecting the right answer after reading a question [14].

The mental representation created while interpreting content uses the individual mental capacity, which is directly influences by the ability to understand the content. Consequently, if the mental capacity is overloaded the individual cannot grasp the notion begin that content [15].

According to CLT, learning a spoken language evolves the ability to comprehend and interpret text according to its lexical and grammar rules. For new programming and new spoken language learning, both processes can evolve around learning different

syntactical rules leading to comprehension difficulties dependent, e. g., on the learner's effort and proficiency to learn new skills [16].

Grammatical and syntactic rules can influence comprehension as the type and complexity of content will influence the CL usage. In spoken languages, some regional factors like cultural barriers and learning methodology can also influence learning. There are also individual factors as, e. g., children tend to learn spoken languages faster [17]. On the other hand, code comprehension is affected by technical and structural rules [16] which, as stated in chapter 1, can lead to additional difficulty as showed by Nikula et al. [18], which reported that globally, in 1999, that more than 30% of students in computer science have failed introductory programming courses.

CL does not only change according to the type of content presented by also the tasks involved with that content. CL can be divided into three components: extraneous, intrinsic, and extraneous CL. Extraneous CL is related to how the information is presented to the learner and so it is related to the representation of the task. Intrinsic CL is related to the difficulty level of the task and its content. Germane CL refers to the pattern of thoughts and behaviour, relating to the production of new patterns in the information (e.g., flowchart representation) [19].

Most studies do not differentiate between the different CL's and consider CL as an owl. Larmuseau et al. [2] tried to differentiate between the different types of CLs using different biometric features that correlated with CL, by using different types of tasks in their protocol. This was done by trying to induce different CL, intrinsic load induced by interactivity element and extraneous load by the provision of hints. They concluded that it was impossible to detect this measure without self-reports of associated mental states, considering the multiple factors that can influence the measured physiological features, and that these distinctions are more relevant in the field of psychology [2].

Considering the joint components of CL, its effects translate into manifestations of the central nervous system (CNS) and ANS as the first is constituent by the brain, which receives and coordinates all body responses, and the second for the systems responsible for regulating involuntary responses [20].

2.2.Nervous System

The nervous system is responsible for transmitting signals between different parts of the human body to produce voluntary and involuntary reflexes and responses. It is responsible for receiving and processing internal and external stimuli, maintaining homeostasis and functions associated with consciousness, memory and thought. The nervous system receiving of stimuli and response occurs through a vast web of neurons.

Within neurons, conduction can be fast or slow (1-120 msec) depending on whether myelin is present or not. Signal conduction within by electrical conduction and between neurons occurs by electrical or chemical conduction. Electrical conduction occurs through the electrical polarization of the membrane by the trafficking of potassium and sodium ions. Chemical conduction occurs by neurotransmitters [21].

As shown in Figure 1, the nervous system can be divided into the CNS and peripheral nervous system (PNS). The first is composed of the brain and spinal cord and is responsible for receiving and coordinating corporal responses, this system is integrated with the PNS as it is composed of the elements of the body that allows communication (sensory receptors, nerves, ganglia, and the plexuses).

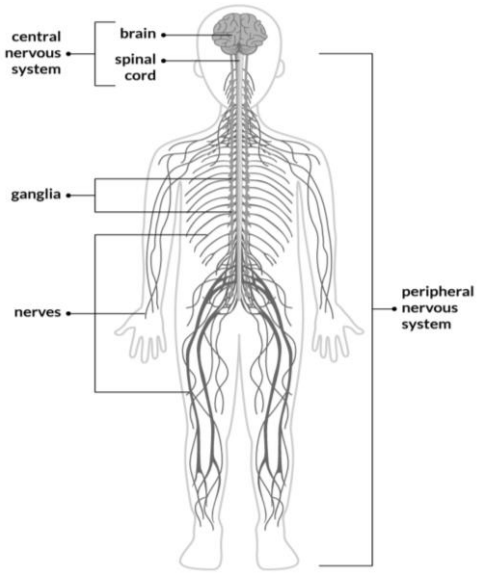


Figure 1-Nervous system division [22]

The PNS can be divided into three groups: autonomic, somatic, and enteric systems. The enteric system is the one responsible for the sensing and regulation of the visceral system and can also be considered an independent system from the peripheral nervous system [23]. The somatic nervous system (SNS) is associated with the motor

functions derived from voluntary actions and the (ANS) from involuntary ones. Within the ANS two subsystems work antagonistically to each other, there are called the parasympathetic and sympathetic systems [24].

Regarding the structures of the encephalon (brain, brainstem, spinal cord, and cerebellum), the brain is the largest and most important of all and can be divided into 4 lobes, frontal, parietal, occipital and temporal lobe. The frontal lobe is associated with motor functions, the parietal lobe is associated with the reception and evaluation of sensory information, the occipital lobe, the occipital lobe is associated with the reception and integration of visual stimuli, and the temporal lobe is associated with the reception of olfactory stimuli and hearing aids and memory management [25].

The motor function of the brain is correlated to the frontal lobe, as previously mentioned, motor function is controlled by both ANS and SNS. ANS is responsible for the regulation of involuntary response by balancing the response from parasympathetic and sympathetic systems, which are antagonistic to each other. These two elements of ANS create the responses for events of panic or relaxation by inducing g different responses via the release of hormones to, e. g., accelerate the HR in stressful situations or in calm situations to, e. g., reduce the breathing rate. This particular example is called the "fight or flight" response of the ANS [20]. It is described that psycho-physiological loads lead to an activation of the ANS which translates into changes in blood pressure, temperature [26] and Heart Rate Variability (HRV) and other factors. These factors can be compared and correlated to measure CL to infer comprehension [27] and also engagement levels [28]. Some articles also considered personality trails to better understand CL response [29], [30] because it has been demonstrated that anxiety and arousal can affect working memory capacity [31] and consequently ANS response.

2.3.Biometric Features

From the physiological effects of CL changes, manifestation on the frontoparietal areas of the brain can be measured with an electroencephalogram (EEG) which reflexes the motor function which allows the identification of different levels of CL as specific power waves as gamma band power (20-25 Hz), e. g., correlates to hyper brain activity [32]. Even though EEG allows for reliable results [33], but it may not be fitted for daily use scenarios for its practicability. Brain imaging can also be used to measure the same

physiological manifestation but add impracticability and high cost. For this reason, papers resort to the use of measurements more practical measurements such as HR, EDA and pupillography data [34], in which papers have explored the extraction of ultra-short features [35].

Electrocardiogram (ECG) is also used to measure CL as it was proven that HRV is controlled by both branches of ANS, and its measurement is sensitive to mental task level and duration. ECG is acquired by electrodes in the skin and typically uses 12 points (6 on the limps and 6 on the torso). The reduction of the number of points is used to simplify the ECG setup as not all channels are needed [36]. With the evolution of processing technology is possible to acquire a 3-point ECG or use photoplethysmography (PPG) to acquire the necessary data to distinguish heart pathologies and acquire HR features.

HRV is based on variations between HR components taken from the HR, which can come from ECG or PPG. In the time-domain measurements of HRV, the most common features are the standard deviation of RR intervals (SDRR), root mean square of successive differences (RMSSD), and the percentage of successive normal sinus RR intervals more than fifty milliseconds (pNN50) as representatives of vagal tone. From these features, it was proved the relation between SDRR, RMSSD, and intrinsic CL [37]. Complementary, it has been described by the literature that the values of mean RR, SDRR, RMSSD, and pNN50 decrease from low stress to high cognitive stress and an increase in mean HR [32]. The variation of RMSSD values is described as more noticeable in comparison to SDRR [38]. Some articles also explore non-linear features like the ratio of Pointcaré plot components (named SD1 and SD2), which are features obtained by the transformation of consecutive data using the area of a time plot (Pointcaré plot), which corresponds to the ratio of minor and major axis of the Pointcaré beat to beat time and it shows the balance between sympathetic and parasympathetic with non-periodic oscillations [39].

Other papers recurred to features related to frequency domain features of HRV as the comparison between very low (0-0.04 Hz), low (0.04-0.15 Hz), and high (0.15-0.4 Hz), these frequencies represent the mix of sympathetic and vagal influences. Features like the ratio of low and high frequencies show the influence of both sympathetic and parasympathetic branches, resulting in an increase of low frequencies in distinguish to high frequencies leading to the increase of this feature with the increase of CL [40].

HRV is generally extracted from ECG signals, but this approach requires multiple electrodes with expensive monitoring devices and may not be practical to daily use of learners or education environment, so this measure can be done from PPG [41] that is available nowadays in wearable devices, like, smartwatches.

PPG is a technique that implies the measurement of light absorption related to blood flow and correlates this flow with the rhythmic changes of the heart cycle to identify the HR. PPG can be done in blood vessels that are distant from the heart and uses two different light sources for better results or to extract blood oxygenation [42].

PPG sensor is a low-cost sensor commonly used in smartwatches and smart bands, and it has been the center of multiple studies to determine its precision in approximation to HRV [43]. HRV features acquired from PPG are also called Pulse rate variability (PRV). The PRV can avoid some ECG artifacts but is more sensitive to movement artifacts [43] and breathing patterns [44]. The literature proved the correlation of PRV to HRV [43], [45] but some described estimation errors in a patient with cardiac diseases [44].

EDA is the measurement of the electrical activity of the skin produced by changes in the activity of sweat glands. The increase in activity of sweat glands increases electrical conductivity. EDA is related to CL because the peaks in skin conductivity are associated with sympathetic activation [46]. EDA signals are divided into two components, tonic and phasic. Phasic is related to rapid response resulting in quick peaks with high amplitude, denominated skin conductance response (SCR). The tonic component of the signal is related to the background signal characteristic by slower response resulting in lower variation of amplitude, being described as the skin conductance level (SCL) [47]. The SCR can be correlated to an event (event-related SCR), and compared to the values outside that event, to distinguish from different mental states but given the fact that the sympathetic nervous system can respond with arousal or stress in situations cognitively challenging, that distinguishment is unnecessary [46].

Some articles describe the correlation of CL changes to EDA features from both SCL and SCR, like the mean, maximum and minimum value of SCL [48] and the mean peak height, peak rate (peaks/min) and quantiles for SCR, where all of this features increase with higher CL [46].

Our developed tool uses HRV and EDA as an index for CL. However, to identify which content region was difficult to comprehend, we used a desktop eye tracker. Eye trackers enable us to record eye activity. Some articles describe the use of eye activity measurements as an index of CL changes, e. g., pupillography and bink rate [49]. Task-invoked pupillary response is a feature associated with pupil dilatation [27]. Blink rate is another eye-related feature that is expressed as the frequency of spontaneous blink being, as pupil dilation, correlated to the activation of the central nervous system [50]. It is also described that behaviour features can be extracted from the eye movement regarding the fixation time and the transition between fixation referred to as saccades [51], unfortunately, the acquired device in this thesis does not allow to acquire this class of measurements.

The acquisition of eye measurements occurs with an eye tracker, which typically measures the position where the eyes are looking at the screen, denominated eye gaze. This measurement allows the identification of content regions that can later be synchronized temporally with the biometric feature to determine comprehension difficulty. Eye gaze is determined by the vector produced between the cornea and the pupil center. The cornea position is obtained with the contrasts created by infrared light that allow the identification of the cornea created by the reflection of its outer surface, which is called “First Purkinje image” [52]. The pupil center is determined by an algorithm concerning the limit between the iris and the pupil. The human eye is described as a structure that presents a crystalline lens with muscles and its surrounded by liquid, which separate the cornea and the retina. The fact that light passes through various materials in the human eye leads to a difference between the measured axis and the real axis of the eye, which requires calibration [53].

2.4.Related work

Regarding the analyses and measurement of the CL, there has been a growth in this topic in the last 20 years with the evolution in data acquisition and models related to CL. According to the results of the search from the website “ScienceDirect”, a database for the publisher “Elsevier”, there has been an increase of about nine times the number of papers per year in the last 20 years (from 649 in 2001 to 6613 papers in 2021) regarding measurements of CL changes.

One of the first papers to describe the measure of CL changes was a study by Winsum et al. [54], which proved in 1984 the relation between CL and EEG by showing the decrease of a specific EEG frequency interval to the increase of CL. It also described that the level of the CL could be inferred from the duration and amplitude of that decrease. Considering the technology available at the time, this was the best indicator of CL changes that could be achieved. In recent years, some articles have explored the same feature of the EEG as Winsum et al. [54], but with the advance in science and technology, it is possible to extract more features and use classifiers to achieve the best possible model. Articles like Candela-leal et al. [55] prove the previously stated point but using an EEG helmet from OpenBCI. With these physiological data, multiple features were extracted, and the best features were selected by a hybrid features selection method. The best features were then put in different models to test their accuracy. The models used were random forest (RF), support vector machine (SVM), gradient boosted machine and classification and regression trees (CART). The best accuracy was achieved at 92,69% with RF.

Even though EEG is a good indicator of CL changes, measurements of ANS manifestation like PPG can be light-weight non-intrusive alternatives and PPG devices may be more accepted by the general population as a component of the tool when compared to EEG helmets, e. g. [56]. PPG also presents itself as a good alternative for ECG equipments as, e. g., a smartwatch can be more practical and comfortable for daily use than an ECG chest band.

Some paper use eye trackers to give information about the user's eye gaze. Eye gaze is the position of the screen the user is looking at. The article of Kang et al. [50] in 2015, was one of the first papers to incorporate eye gaze synchronization in CL assessment and the first to apply that methodology to support learning a foreign language by identifying unknown words. The goal of this paper was to identify comprehension difficulty and correlate it to the corresponding content location to identify a specific predefined content area. This concept was designed to help the users learn Korean as a foreign language. It used an eye tracker from Tobii to extract the eye gaze and used an EEG band with two points from Brainno to extract EEG features synchronized with the eye gaze. Even though the 2-point EEG band can be more comfortable, than the EEG helmet referred to in Candela-leal et al. [55] the previous mention, results in less accuracy than the EEG helmet and it still presents high costs. The accuracy of 74.76% of an SVM

model of the EEG features of this EEG band supports the multimodal approach previous mention. Also, is important to notice that the experimental protocol did not integrate the eye tracker and the EEG modules as they were validated separately; the eye tracker data were tested to estimate the unknown word and the EEG was tested in distinguishing texts with known and unknown words, in that sense this study does not synchronize the eye gaze with measurements of CL changes.

Some studies also explored more proficient assessments of CL measurement, such as Ayres et al. [49] in 2020, where the CL changes were evaluated using EEG, functional magnetic resonance imaging and Functional near-infrared spectroscopy, these acquisition methods present high costs and reduced practicability with long acquisition times and high weight, which reduce is possible use in real life scenarios.

Bianco et al. [57] in 2019 proposes the use of HR, EDA and perinasal perspiration (acquired with a smartwatch from Zephyr, a GSR sensor from Shimmer, and a camera from FLIR Systems) to identify a state of cognitive, emotional and sensorimotor stress that can jeopardize driving safety. This article stands out from the others that wore already referenced by the fact that uses a thermal camera to determine perinasal perspiration, which is related to the breathing rate by checking the flow of air in the nostrils. The protocol is set to create a general sense of distraction state using stimulus by smartphone but can be interpreted as a lack of distinction between arousal, stress, and cognitive overload. Despite the interesting approach of applicability of this paper to other multimodal studies, the identification of poorly defined features for the specific psychological state that has been assessed makes the results of the study less significant.

Very few studies evaluate the use of light-weight wearables to assess CL or the assessment of CL at the region-level. Hijazi et al. [11] describe the concept of using lightweight wearables to assess CL at the region-level. Hijazi et al., an article from 2021, extract HRV from ECG and pupillography from a low-cost eye-tracker from Tobii, being the same eye-tracker described to be capable of being used to indicate the region of code that the user is looking at, consequently, identifying the content source of comprehension difficulty. It also makes use of ML algorithms to classify the cognitive state of the user in real-time recurring to different methods as other previous mention like SVM, k-nearest neighbor (KNN), and also, RF, decision tree, and gaussian Naive Bayes (NB) classifiers, it is the most relevant article in terms of results as it has a protocol which tested the combination of features with different ML algorithm which, according to the article, let

it to outperformed the stated of the art in precision & recall by 23% and 17%, resulting in an accuracy of 83.00%± 0.75. This paper describes the extraction of features to classify comprehension with the synchronised content region which differentiates it from other articles.

The use of ML algorithms allows the distinction from different states of CL and levels, most articles, even those differentiating CL from other types of mental states like stress in Setz et al. [46] and arousal in Markova et al. [58], identify general states, identify two general states but some article make the distinction from three levels of CL. Romine et al. [59], are one of those and present the use of a wearable watch called Empatica E4, to measure the CL related to problem-solving. This paper uses PRV features, EDA, body and temperature and compares these data with reported learning experience and performance, describing the possibility of use for personal environment and educational environments. Regarding the equipment used, Empatica, has been reported as effective to extract PRV with results very close to HRV according to Lascio et al. [56]. The distinction between the 3 states, resulted in the identification of a medium cognitive state between overload and a low cognitive state. Such distinction is interesting in a school environment, as it can allow identifying a state before overloaded/high mental load according to the author.

Other papers, like Abbad-Andaloussi et al. [60], explored code comprehension by using only eye tracking. The extracting of both behaviour features (pupil, fixations and saccade features) and eye gaze allowed to correlate content to comprehension, only lacking on use and comparison of multiple classifier methods to improve prediction as only a decision tree was used.

Romine et al. [59] use the same set of classifier algorithms as Hijazi et al. [11] but also include others, like the logistic regression model, AdaBoost, and random forest algorithm. Concerning the first set of algorithms, the baseline models, it presented the best classification as KNN with 81% of accuracy and in the black box model, random forest performed even better with 85% accuracy. Even though these results have shown higher accuracy when compared to Hijazi et al., these cannot be compared as the minimum window lengths of both papers are not described. Comparing the window length used on the baseline on both papers, Romine et al. used 2 minutes while Hijazi et al. used 30 seconds. The bigger the window length, the more precision and accuracy the

result will be, but the smaller the window length, the more proximal it is for real-time [11].

Ahmad et al. [61], which also distinguish three levels of CL used features from pupillography, blink rate and HRV with the objective of underlying student attention. The Cor Sense extract PRV from the finger rather than the wrist, and the eye tracker records the eye position by recording the pupil motion with a sensor in front of the eye. The data from the 41 subjects resulted in an F-score of 0.85 with the use of an RF model.

Some papers like Mills et al. [62] used eye tracking to extract behavioural features, as it was used to detect attention lost while reading and intervene by accessing the user comprehension of the last paragraph with questions. This paper used Tobii TX300 and Tobii T60 eye trackers to extract fixation features, saccades features, blink rate, pupil diameter and eye movement features. Even though comprehension assessment is validated with feedback from the user, comprehension difficulty can occur without mind wandering. Features related to behaviour like eye movement are less significative than physiological manifestations of mental stress like pupil diameter.

To correlate the measurements of biomarkers with cognitive state most article develop experiments to trigger such cognitive state and validate that information with self-reports but other studies such as Keller et al. [63] obtains their results with performance tests which can avoid fake self-reports according to the author. Keller et al. [63] used a pair of eye-tracking glasses from Tobii to measure eye gaze and pupillography and a PPG device from Cor Sense, similar to Ahmad et al. [61]. This article intended to determine comprehension through CL measurements with an interface where the subject needs to find invented words, words that do not exist in the English dictionary or that cannot be used in that sentence’s context. The use of the eye-tracing (Tobii pro glasses) results in the measurement of parameters closer to the target eye, being less susceptible to errors in the identification of the gaze direction; however, the chosen equipment presents smaller sampling than other "conventional" eye-tracking equipment as those placed on the monitor. The Tobii 5L can extract data up 120Hz with pupillography and eye movement data in comparison to the Tobii pro glasses with a sampling of 100Hz. The use of eye-tracking glasses may not be compatible with users who wear glasses, or as comfortable, due to the weight that the battery can have on the ears.

Table 1 represents a table view of the objective, features results and limitations of the most relevant article described in this thesis.

Table 1-Most relevant related work

Reference	Biomarkers	Methodology	Results	Limitations
Mils et al. [62]	Eye tracking, behaviour	Eye features correlation to identity low attention moments in real-time	The model presented with a weighted precision of 72.2% and a weighted recall of 67.4%	The experiment protocol was not designed for the specific goal of the paper, the experiment only used one text as reading material and the use of high cost, non-commercial eye tracker.
Romine et al. [59]	Temperature, PRV, EDA	Distinguish between 3 levels of CL by physiological data and machine learning techniques.	The best classifier is random forest with an F-score of 0.79-0.80 (random forest)	The lack of behaviour features or eye tracker simplifies this model as only evaluating the time domain of comprehension and not localizing the content
Ahmand et al. [61]	HR, HRV, Blinking rate, pupil diameter	Distinguish between CL in 3 levels of attention in real-time	The best classifier is random forest with an F-score of 0.85-0.95	The use of not so practical devices to measure the features on learning environments
Peng et al. [1]	Facial and eye tracking, HRV, audio	Identification of student mental state from facial, HR and acoustic modalities	Best performance in concentration and boredom with an accuracy of 0.842 and 0.810 respectively	Some states are hard to identify and subjective and the data wore obtained from only lab settings
Hijazi et al. [11]	HRV, eye tracking, pupillography	Identify regions of digital content that cause learner Comprehension difficulties	Accuracy of 83% with precision and recall of 0.89, 0.79	The experiment protocol was not designed for the specific goal of the paper

In Table 1, Mills et al. [62] describe the use of only one equipment, the eye tracker. to extract both eye gaze and features. Even though the setup, proposed by Mills et al, does not incorporate features like HRV, that same decision led to low results. Features like HRV which have minimum window length, will increase the processing time but also are more sensitive manifestations of ANS compared to pupillography. Also, the price of the equipment used was high, existing for that reason papers like Sandhu et al. [64] use low-cost equipments, in this case, an eye tracker from Eye Tribe, which shows the importance of the selection of low-cost equipment as it was done in this thesis. Romine el al. [59]

used light-weight wearables equipment to measure CL with a relatively adequate time resolution, which is also limited for assessing the CL without giving attention to the content regions that cause the comprehension difficulty as opposed to the region-level identification in this thesis tool. Peng et al. [1] explore the distinction of CL from other mental states, like boredom and frustration but very few articles have explored that differentiation by using wearable [58], unfortunately, it lacks the evaluation of different classifier techniques, to compare precision and recall. Complementary, even though boredom states can lead to additional information about the user state, a low mental effort state can still allow evidence of CL change [12]. Additionally discriminating frustration from CL is also not relevant as frustration can be an effect of overloaded CL [46]. As previously stated, Hijazi et al. [11] propose the concept of real-time evaluation of binary classification of CL with adequate time resolution during code review, which presents itself as the most relevant article in Table 1. Considering that this paper was the first to describe the concept CL assessment using light-weight devices, this paper was the foundation of the thesis methodology as it introduces most of the ideology behind the development of the tool of this thesis.

This thesis tool differentiates itself from the described literature by its use of low-cost wearables at region-level with fine-grained analysis and even though articles like Hijazi et al. [11] give the fundament of this thesis methodology, their implementation was limited to the use of more heavy or more expensive equipments like clinical ECG as Mills et al. [62] and the classification of CL changes on only task level like Romine et al. [59].

CHAPTER 3 – METHODOLOGY

3. Methodology

The developed tool uses a web extension to open the reading content and start the data acquisition. In this chapter, we will analyse the architecture decisions, components, methodology and implementation of the tool. After the data acquisition, the data is filtered and processed to extract features. These features can then be selected to be trained as an ML classifier. This chapter includes the feature analysis and the ML pipeline description.

The first subchapter presents the tool architecture and the progressive adaptation made to the tool, the second subchapter presents the details of the data acquisition, the third subchapter describes the device setup and the fourth and fifth chapters present the data synchronization, feature extraction and the ML pipeline.

3.1. Tool architecture

In iMind, we hypothesize that analysing the reader's biometrics and eye movement will allow us to determine the temporal and spatial location of the content that was difficult for the reader to comprehend.

The tool architecture comprises the tool components and pipelines involved in transforming data acquisition into CL prediction, as shown in Figure 2.

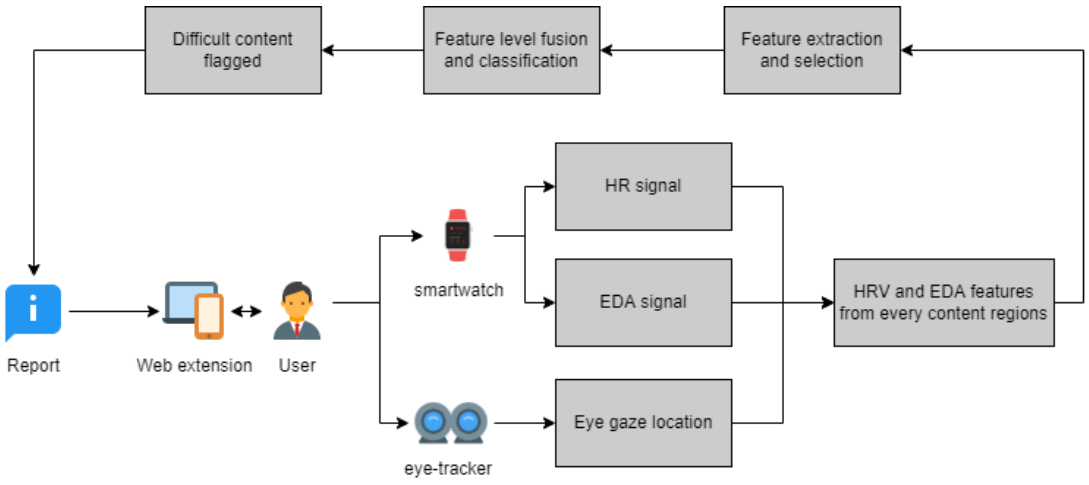


Figure 2-Diagram of the tool architecture

The web extension is the first component of the tool as it is with the interface that the user reads its content and starts the data acquisition. While the data acquisition starts, the data from the user devices are recorded. With the smartwatch is possible to acquire HR and EDA data and with the eye-tracker we can acquire the eye gaze (the position in which the user is looking at the screen). Since the content displayed is correlated with the

eye gaze to define content regions, if the location on the screen in which the content has been displayed changes, then the displayed content location needs to be updated. The content location synchronization represents, in this case, the synchronization of the eye gaze with the content scroll. The content scroll is acquired by the web extension from the web page where the content has been displayed.

As shown in Figure 2, the HR and EDA data, after being extracted, must be synchronized with the eye gaze position and page scroll to extract the eye position on the content. HR and EDA data are divided per region so that features (HRV and EDA) can be extracted to then classify that content region. The use of classifiers allows us to, according to the best features, predict if a region is "difficult" or "non-difficult".

The resulting information from the classification module can then be used to flag the difficult content, which may be used on the interface as feedback in future work.

We will now elaborate on the main components of the tool. While reading content, the user data will be acquired by the smartwatch and the eye tracker, leading to HR, EDA, and eye gaze data.

3.1.1. Smartwatch

The goal of the smartwatch is to acquire HRV and EDA features that can be correlated to CL with sufficient resolution to be synchronized with eye tracking to analyse the content, as shown in Figure 2. The use of a smartwatch is centred on its practicability for daily use.

Considering that to measure physiological data, more precisely CL, the best setup is to acquire EEG data, as mentioned in subchapter 2.3, which can be measured through an EEG cap, but sine such types of equipment can be impractical for certain environments. The growth of wearable use and their improvement in sensitivity over the years make smartwatches and other types of wearables adequate candidates to acquire biometrics.

Initially, the acquired smartwatch was Fitbit sense brand, which was described to measure EDA and HR. However, there were limitations to its use. The following are some examples of these limitations:

1. Lack of time fluctuation on HR. That means the HR data was smoothed with various filters, which can be a challenge to detect peaks or spikes.

2. EDA acquisition implies putting the hand-palm on the watch frame for some seconds, which is impractical.

Regarding the first limitation, the HR data did not present significant HR fluctuation in time, most likely justified by the strong data normalization and filtering. The second limitation was the fact that the acquisition of EDA data implied that the user placed his palm on the watch's frame, which meant that the user could not write while the data acquisition was happening, which contradicted the practicality of the tool.

The commercial app of the Fitbit sense allowed for a more precise HR measure by using electrodes present on the watch frame and the back to be equivalent to a single-channel ECG, unfortunately, this data was not accessible to developers. This function also needed the contact of two hands, restricting movement, and the data acquisition was restricted to a proprietary application.

Considering the lack of sensitivity from the Fitbit Sense, a new smartwatch was acquired, Empatica E4, displayed in Figure 3**.**

Figure 3-Empatica E4 Specifications [65]

The new smartwatch, Empatica E4, uses a communication protocol that requires the use of a Bluetooth dongle to stream data to the computer directly or an android application. The Empatica E4, compared to the Fitbit sense, allows access to more precise raw data like BVP signal, which is used to extract the HR.

Empatica E4 is a smartwatch with no display showing only a multiple-colour LED light. It works by measuring BVP with a 64 Hz sampling rate by using a PPG sensor in the back panel. It also includes an accelerometer to measure the movement of the user

with a 32 Hz sampling rate and an infrared temperature sensor that extracts the temperature of the wrist with a 4 Hz sampling rate, two electrodes in the wristwatch to measure EDA with also 4 Hz. The device communicates with a low-energy Bluetooth connection, and it uses its internal memory to store recordings in case of a lost connection or to simply work without a server. The device can communicate to the computer using a streaming server and a dongle or to a proprietary server using an Android application.

After acquiring the new device, a side experiment was done to evaluate its results with CL changes. The side experiment was done by reading two texts of different English levels using Empatica E4 and an ECG. The side experiment showed similar HR features fluctuation between the new device and the ECG, showing the potential of using the new device. Additionally, the data analysis shows that the new device did not suffer from the lack of time fluctuation as Fitbit Sense.

3.1.2. Eye tracker

The Eye tracker's goal is to measure the position the user is looking at on the screen over time, to then identify the content region with high CL, which probably caused comprehension difficulties to the user.

Eye trackers can operate in different modalities, mainly by either a set of cameras on the monitor or by glasses equipped with cameras. Eye tracker typically includes multiple cameras to obtain the axis created by the eyes and the screen resulting in the gaze point, the location in which the user was the location at the screen. This data can be used in multiple ways, from video games to attention factors in reviews/tests. Eye trackers have a variety of prices, so the recent rise of low-cost eye trackers makes their integration more appealing.

As described in subchapter 3.2, Keller et al. [63] described that eye-tracking glasses might not be compatible with users who wear glasses or as comfortable, due to the weight that the battery can have on the ears; this is the main reason why we chose a desktop eye tracker, for that reason, the Tobii 5L eye tracker was acquired. A desktop eye tracker is put in the monitor, and from there, it tracks the user's eyes. The Tobii 5L works for screens up to 27 inches and presents two modes of operation, 33 Hz and 120 Hz, with data regarding pupil dilation and eye gaze.

The Tobii 5 L works using a set of cameras that observe the eye with visible and infrared light (sensors are visible in Figure 4) allowing it to observe the centre of the pupil and the reflex of the cornea and create the axis that gives the direction in which the user is looking, this axis is used to calculate where the user is looking at.

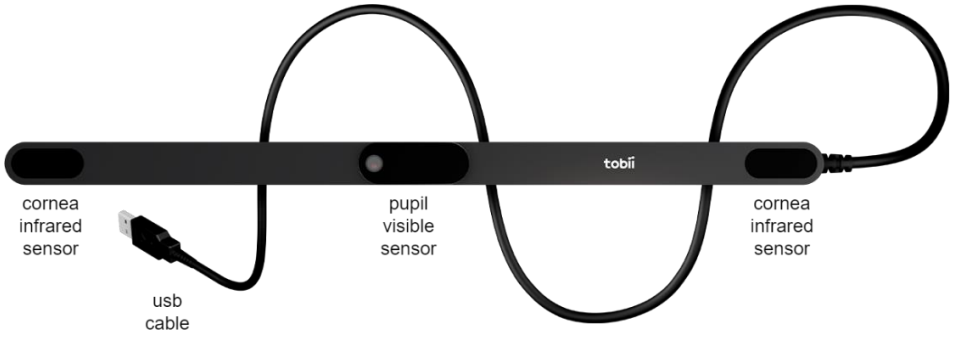


Figure 4-Tobii 5L from Tobii [66]

The API sold by Tobii presented some computational challenges in terms of integration as it is a platform development kit (PDK) connected to a streaming server that presents a complex system with multiple dependencies making it difficult to integrate in the tool. Instead of using the PDK of Tobii, an executable was developed to extract the instant gaze point measured by the device resulting in a 1000 Hz sampling rate without access to the user features like pupil dilation or saccades.

3.1.3. Web Extension

A web browser extension was chosen to implement the tool because it would be easy to integrate with various platforms (e.g., English learning platforms).

Its goal was to open any select text and pdf files and be able to start and pause the data acquisition when wanted. As the first interface, a web browser extension has a popup that is always accessible by clicking on the extension icon. This main popup has four buttons defined to start, pause, resume and stop an input to select the document to open.

When the start button is pressed, both the HR, EDA, and gaze data should start recording. At the same time, the interface opens a web page that includes a grey screen used for calibration, and after 30 seconds, it opens the content to be read and understood, as shown in Figure 5.

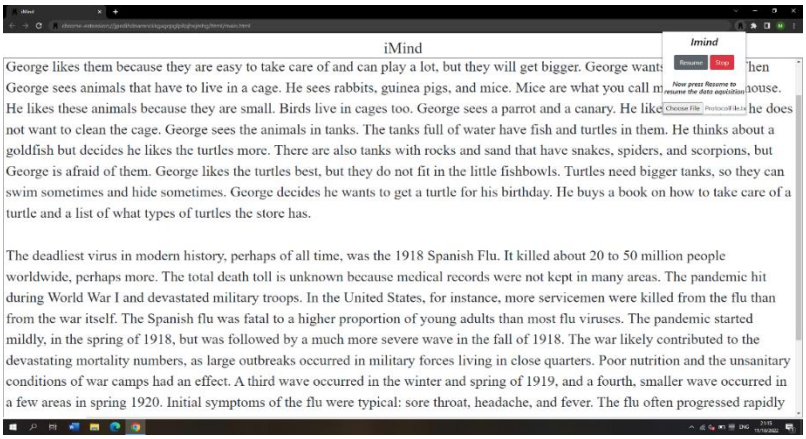


Figure 5-Content Interface and popup

Selecting the web browser to develop the tool is essential. Therefore, we compared the main web browsers (Safari, Edge, Chrome, Firefox). From the main browsers, in terms of compatibility, security and performance they are almost indistinguishable [61], with the majority, presenting tools for developers [62]. Apart from that, most global users use Chrome when accessing search engines [67] and websites [68], which corresponds to a relatively large difference when compared to the statistics of other browsers, like Safari and Firefox, which are the second and third most used browser in 2021 with 17.7% and 5.8% of user's use correspondingly [69].

It was initially believed that the web extension would not need a server to write and read local files. Extensions are executed in the browser, which for security reasons, cannot access, alter, or delete content from the user's computer. For this reason, it was required the addition of a server to manage the beginning of data acquisition and reception, for which a server is developed in Node.js, which is a backend development environment [70].

This server receives requests from the extension through an HTTP connection to start and stop the collection of data referring to HR, EDA, and eye gaze. An HTTP connection is a communication protocol commonly used to start a connection to the internet [71]. The server is also responsible for adding a counter id number for each run and storing in the computer all records of the tool, which include the identification number for each run, timestamps, scroll and selected options, as the tool allowed to select a file to open and exceptionally other information for the experimental protocol version, described in subchapter 4.1.

3.2.Data acquisition

The Empatica E4 is a smartwatch that uses PPG. It works by measuring the light absorption allowing detention related to blood volume changes to estimate HR.

As explained in subchapter 2.4, most PPG technologies use two light sources, red and infrared as these light sources allow for the extraction of both HR and oxygenation levels. Empatica E4 uses infrared and green light-emitting diodes as light sources, as green light-emitting diode contributes to higher battery life compared to infrared.

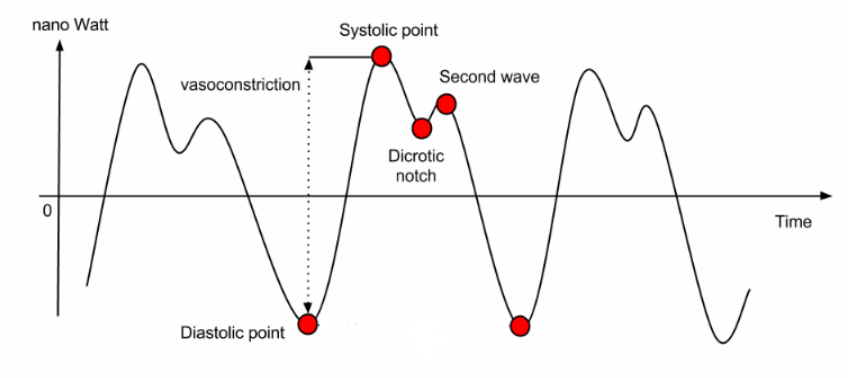


Figure 6-BVP processing [72]

This type of measurement is cheap but is sensitive to sudden movement because it is carried out on the wrist, which must be compensated in the processing. By measuring the variations of absorption of the light beam, the watch can measure values referring to the volume of blood.

Figure 6, taken from the Empatica website, shows how the rhythmic changes in the BVP signal (which represents the blood volume) have distinct peaks and fluctuations related to the heart cycle. The cardiac cycle is created by the blood flow which goes from the cells to the heart and then to the lungs and back to the heart, to pump blood the cells. This cycle is possible by the heart atriums and ventricles. The first peak on a BVP signal is created by the ventricle bombing blood to the body and the second peak is created by the ventricle bombing blood to the lungs. The intervals between two consecutive events translate into the heartbeat [73].

The EDA measurement of the Empatica E4, uses two electrodes at the opposite end of the sensors to measure changes between a defined range of electrical current 2-20 µS, this is referred to as skin conductivity [73].

Now, regarding the eye tracker, as described in subchapter 2.3, it works by using two cameras and infrared light-emitting diodes to give a depth perspective that allows tracking of both eye's orientation axis, and distance to the sensor. The eye tracker can also be used as a cognitive indicator related to behaviour features such as saccades and fixation, considering the eye gaze presents 1000 Hz of sampling [74]. Behaviour features and mean blink rate and mean eye movement speeds are incapable of obtaining results in real time since those features tend to present temporal variations within the same individual [75]. In this sense, the use of the eye tracker is mainly centred on the location of the individual's gaze point.

After extracting the HR and EDA features of each content region is also possible to extract features. After labelling those features is possible to use classifiers to classify that content region. In the future the classification of the content region could be integrated into the tool to create feedback, giving utility to a data report, as shown in Figure 2.

Before describing the feature extraction and ML pipeline is important to describe the tool implementation regarding setup and pre-processing.

3.3.Device setup and Pre-processing

Regarding the pre-processing, the eye tracker configures the setup and the calibration of the eye tracker for the user, which is done by a proprietary software of Tobii to save the setup and user information on the device.

The configuration of the setup is done by aligning the eye tracking with the screen and recording the distance between the eye tracker and the screen. This configuration only needs to be done once with new computers where the tool will be installed.

The calibration of the eye tracker takes place every time there is a new user or when the user's condition changes (like putting on glasses). It is done by giving feedback about the user's distance to the screen and then looking at pointed dots on the screen. The feedback about the head position helps if the user is too far away as the tracker can have difficulties recognizing the eyes.

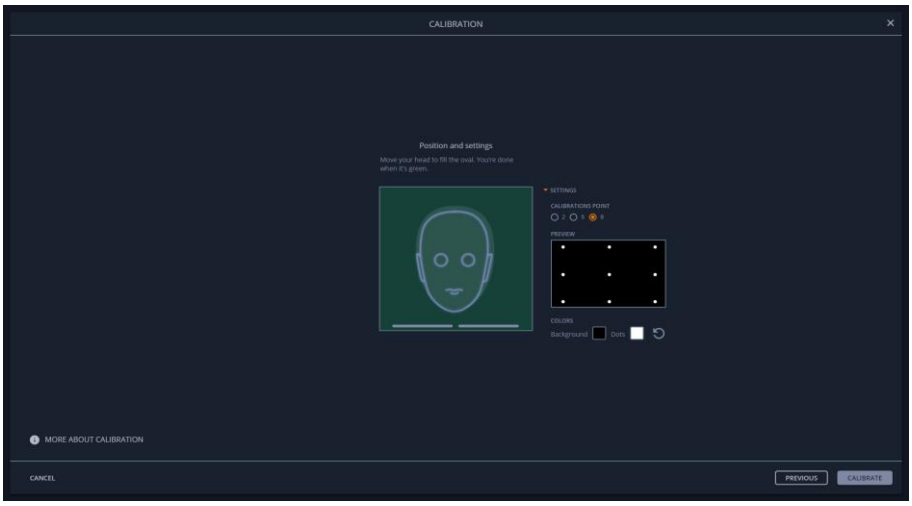


Figure 7-Tobii eye-tracker calibration

The calibration takes place using 2 to 9 dots on the screen. To achieve better calibration, 9 dots are selected as shown in Figure 7. After pressing next, the calibration forces the volunteer to look at each of the dots in random order, this procedure allows the correction of the eye's real axis as the eye tracker through its perspective only sees the optical axis and not the visual axis, this deviation occurs by the fact that humans eyes are made of different materials with different refractive indexes leading to the distortion of the eye axis [76], this is exemplified on Figure 8.

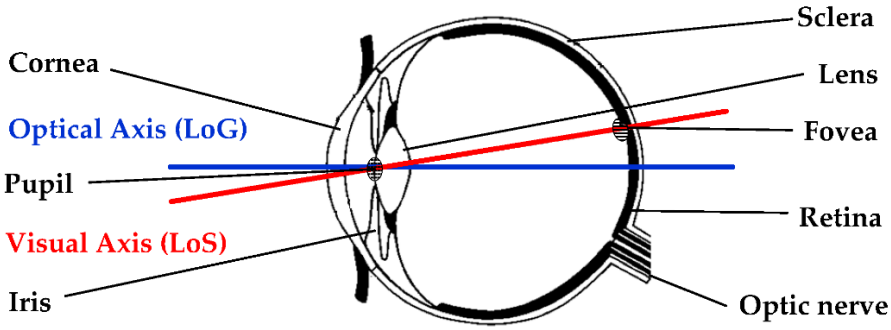


Figure 8-Schematic of eye optical and visual axis [76]

Now regarding the smartwatch setup, as it was previously explained, the smartwatch uses PPG sensors to extract BVP. BVP data can be converted to HR to extract HRV features. With the smartwatch, Empatica E4, it is possible to use two setups, as the data can be sent directly to the computer or a server. If the data is sent to the server, the results are only available after the end of the data acquisition.

Considering the two possible setups, there are two stages of processing:

1. Processing performed in the smartwatch (transformation of PPG measurements into BVP).
2. Processing performed outside the smartwatch (transformation of BVP into HR).

The pre-processing done to acquire BVP is a common procedure done in PPG devices. Its description has been taken from PPG features descriptions provided by Schuurmans et al. [73] regarding this device. The first stage applies a lowpass filter to remove the aperiodic variation of the BVP signal and a median filter to avoid baseline drift.

As the data is stored in the computer/server, the second stage takes place as heartbeat intervals obtained from the intervals of the consecutive waves of the BVP are non-uniform, forcing the application of a cubic spline interpolation to resample the data [73]. The resample is used to correct unevenly sampled into one single sample and to compensate for missing values. Cubic spline resample is also used to avoid errors at the beginning and end of data done by polynomial function in resampling, also known as Runge's phenomenon [77]. The resample results in a sample of 1.25 Hz, applied a moving average to round the sampling frequency to simplify synchronization, as 1 Hz.

The EDA measured by the watch is obtained with two electrodes in the watch which are in contact with the wrist skin. According to Benedek et al. [78], the Empatica E4 background software measures the current by its electrodes and then applies a filter cut band related to the skin conductivity intervals of amplitude. After that, a resample process results in a sampling frequency of 4 Hz.

3.4.Data Synchronization and Features Extraction

Data synchronization aims to map each content region provides to its associated biometric (HR and EDA) data. Through this synchronization, we can identify the content regions associated with high CL. To synchronize, process and extract features per region of content, a MATLAB script was developed. This script extracts the timestamps created by the tool. With the intention of in future work improving this tool to work in near real-time, the script was also adapted to work as an executable. The executable can run the

data from the last run as the tool adds an identification number to every file created related to the identification number given to each run.

The data synchronization occurs by using the timestamp recorded by the tool. These timestamps are stored on the computer by the server, as described in subchapter 3.1.3. These timestamps represent the start, end, and pause timestamps and are an essential component of data synchronization.

After synchronizing the data (EDA and HR), that data needs to be divided according to each content region.

Each content region can be pre-defined by specifying the number of lines of each region or by pre-defining the number of regions. With the information taken from the eye tracing and the defined regions is possible to identify which region of content the user is looking at over time, allowing the synchronization of HR and EDA accordingly.

The eye tracker data allows us to determine the eye gaze, but the scroll of the window is also necessary to define content region's position over time. In that sense, the scroll is recorded with a sampling rate of 10 Hz and stored by the tool. Using the scroll and the eye gaze (sampling rate of 1000 Hz), is possible to define visits and revisits to content regions. The synchronization of the scroll data and eye gaze is done by updating the Y value of the eye gaze with the scroll value. The resulting array is then “downsampled” to 5 Hz for improved computational speed.

As previously stated, with the eye gaze synchronized with the scroll is possible to identify the regions the user is looking at. To analyse the number of times the user looks at a particular region its needed visits and revisits definition.

Visits are identified as the first time the user looks at a particular content region and revisits as the times the user looks at a particular content region that the user has already looked at previously. In the developed script, thresholds are used to define minimal time periods to differentiate visits/revisit from small fixation and guarantee minimum time windows in visits, as some papers do [79], [80]. The data regarding the eye tracker gives us information about when the user looks out the screen and when the sensors cannot determine the position of the eyes, which can happen if there are obstacles or if the user leaves the device's field of vision. Revisits are defined, in the script, as regions that the user have been already looked at previously, which can happen by looking back to a previous region. Looking outside the screen or if the eye tracker losses the eye

gaze temporally does not count as revisits by itself. Each region's data includes the visits and the revisits to that region.

The visits and revisits timestamps are used to divide the HR and EDA data according to the content regions (visits and revisits) to then extract the features for each region.

Before processing HR and EDA data, is possible to extract two features from the eye tracker, one regarding the time spent looking at a region and another regarding the number of revisits, as shown in Table 2. According to some papers, the number of revisits reflects premature shifts of attention and is more evident in tasks that involve searching words [81]. Regarding time spent, according to CLT, a reduction in reading time suggests a reduction in CL [3].

Table 2-Eye tracking features

Feature name	Feature meaning
TotalTimeSeconds	The number of seconds spent looking at a given region
RevisitsNumber	The number of revisits represents the number of times the user returns to a given region

Reading time (time spent in each region) and the number of revisits features have proved to be good indicators of the CL associated with comprehension [82], [83].

3.4.1. HRV features analysis

Following HRV guidelines [77], each HR Data is transformed into RR intervals. RR intervals correspond to the inverse of HR, multiplied by 60 (in case HR is recorded in beats per minute). The RR intervals are filtered with a low pass filter to remove noise from the acquisition. With a sliding window with a window length of 30 seconds is extracted four-time domain features and one non-linear feature as shown in Table 3. Since it is not possible to acquire data with only one or two values of RR with significance, that why a sliding window is applied. A 30-second window length is used because studies proved that even signals of 60, 30 and 10 seconds could be a potential marker of mental stress and other ANS manifestations [35], [42]. Time domain features are obtained by

mathematical calculations of the difference between RR intervals as explained in subchapter 2.3.

While the frequency domain captures the linear variation of the ANS manifestations, non-periodic oscillations on the RR interval are only be captured by non-linear features [84]. Non-linear features evaluate non-linear patterns like the area of a plot which is the case of the Poincaré plot. Poincaré plot is a non-linear geometric analysis that allows extracting features from the axis of an ellipse fitting done to the time plot of RR. Each axis of the ellipse width and length are called SD1 and SD2 and the ratio of these two is SD12. Since SD1 represents short-term variability of HRV and SD2 long-term change, SD12 reflects non-linear non-periodic variability of RR intervals, which may include additional non-periodic oscillations that are not expressed in frequency domain feature [85].

To extract the frequency features of HR and follow the same HRV guidelines, the Hanning window is applied, which reduces the ripple effect [86], leading to less leakage and distortion of the signal frequency domain, which is the method used for analysing the frequency domain of a signal. The signal frequency domain of a signal can be obtained, e. g., through a method called Fourier Transform. To obtain the power spectral density of a signal, the burg method was selected as parametric methods have the better results with HRV to obtain the low (0.04-0.15 Hz) and high frequency (0.15-0.4 Hz) [87]. Burg method is an autoregressive method which uses the Fast Fourier Transform algorithm to convert a signal to its frequency domain and its order was calculated using a partial autocorrelation sequence [88]. All the described HRV features are represented in Table 3.

Table 3-HRV features

Feature abbreviation	Feature type	Feature meaning	Feature fluctuation with increased cognitive load	Reference
Mean HR	Time-Domain	Average HR.	Increase	[89]
SDSD	Time-Domain	The standard deviation of the difference between successive RR intervals in ms.	Decrease	[89]

RMSSD	Time-Domain	The root mean square of the difference between successive RR intervals.	Decrease	[89]
SDRR	Time-Domain	The standard deviation of RR intervals in ms.	Decrease	[42]
SD12	Non-linear	The ratio between the minor and major axis of the Pointcaré beat to beat time.	Decrease	[30]
LHFratio	Frequency-Domain	The ratio between low and High frequencies.	Increase	[37]

Concerning Table 3, as described in subchapter 2.3, most temporal features have a relation to a parasympathetic activity like RMSSD, SDRR which represents the size of the rapidly changing components of the sympathetic and is characterized as a decrease with the increase in CL. Other temporal features, like SDSD and average HR represent an overall estimate of variation of the motor division of the PNS (both ANS and SNS, which are related to the involuntary and voluntary response) describing an increase of average HR and decrease of SDSD with increased mental stress [89].

The ratio of Low/High frequencies, LHFratio partly represents the ratio of sympathetic and parasympathetic branches as it is composed of a mix of sympathetic and vagal influences showing an increase in magnitude, being described to show better results than features like SDSD and HR as expected [40]. Similar results are obtained with the ratio of the Poincaré plot (SD1/SD2) which reflect a relation from the branches of ANS as well [85].

3.4.2. EDA features analysis

Now concerning EDA features, the EDA signal is divided into two components tonic and phasic components SCL and SCR. The EDA components can be defined by a filter cut using various value thresholds or using a polynomial model to extract the tonic component, SCL. The pipeline uses a function called cvxEDA [90] which uses an optimized convex model to extract the tonic component concluded by some papers as cable to differentiate phasic from tonic [91], [56].

The phasic component is taken from the difference between the raw data and the tonic component. With the phasic component, is possible to identify rapid chances of EDA characterized by the phasic peaks. Phasic peaks can be identified by their prominence, which is the distance between the peak and the base of the wave. The

literature described the prominence peak threshold as typically 0.05 µS, with some articles also describing 0.04, 0.03 and 0.01 µS [78]. For this reason, a function was design to select the best threshold between the four values using the peak rate as decision input. The peak rate is calculated by the number of peaks in a time window. The peak rate is typically from 0.016 to 0.050 peaks per second and up to 0.33 in some situations [47]. All the described EDA features are represented in Table 4.

Table 4-EDA Features

Feature abbreviation	Feature meaning	Feature fluctuation with increased cognitive load	Reference
SCL	Tonic component of EDA signal	Increase	[92]
SCR	Phasic component of the EDA signal	Increase	[92]
EDAPeakrate	The ratio of the number of SCR values divided by the data time	Increase	[78]

Concerning Table 4, as described in subchapter 2.3, EDA is only influenced by sympathetic activity, which differentiates from other features like average HR being possible to use it as an indicator of stress over CL within known data events. EDA processing guides are recent, but decomposition is the most common process as EDA can be decomposed of rapid changing (phasic) known as skin conductivity response (SCR) and slow changing data (tonic) known as skin conductivity level (SCL) increasing both with the increase in CL. SCR can be associated with events, justifying the classification between different data events. SCL is described as the most discriminated from overall cognitive states and SCR for event-related activations [46].

3.5.Machine Learning pipeline

The overall goal of using ML in this tool is to unveil hidden patterns of the physiological data to get insights into comprehension difficulties. Statistical and traditional methods of analysis might be inappropriate to capture the complex and non-linear relationships between the physiological patterns and user comprehension process. Moreover, the inter-variability among subjects in exhibiting physiological response is high, thus, ML could be a solution.

The data used in the machine learning pipeline incorporated features from HRV and EDA and eye tracking features (TotalTimeSeconds and RevisitsNumber).

After extracting the features, we labelled the dataset with binary labels (difficult to comprehend/not difficult to comprehend) to train the biometric and behavioural (e.g., reading time) features on these labels. Labelling is a challenging component of an ML pipeline. Therefore, we tried to come out with different formulas to see the best that can express the "comprehension status". The labelling occurs using thresholds (obtain by experimentation) of the number of highlighted words and the number of wrong answers, which reflect the volunteer's understanding of the content.

After extracting features, we ended up with 69 features (EDA, HRV, and Eye-tracker). Therefore, it's possible to use various data-driven and hand-crafted feature selection methods. The data-driven approaches are statistical approaches that examine the correlation between the features and the labels, and the relation between the features, selecting the most correlated ones with the label and the features with the least correlations to other features (to avoid redundant features). On the other hand, we used hand-crafted features selection from our knowledge of the domain and established literature, such as revisits to a given region.

Feature selection methods, like the ANOVA feature selection, allow the selection of features according to the f-value to correlate statistical measures regarding a variance to select the best features. The higher the variance in the feature space the more impactful this feature will be on the labelling.

As described in Figure 2, after the selection of the features, the settings of the models, known as hyperparameters, are changed manually with a grid search method, which is a method that from a randomised selection or a grid of parameters, selects the features with the best performance.

While partitioning the data sample to train and to test the model with, e. g., 70% for training and 30% for test, some data will only be used for either training or test, considering that if the model encounter only similar samples it can lead to overfitting, as the model will be trained only to a specific portion of the sample, for that reason cross-validation is required. There are multiple cross-validation techniques, like k-fold and leave-one-out which do multiple iterations according to the number of divisions selected

for the dataset. K-fold has greater results when compared to leave-one-out cross-validation when dealing with a big dataset.

Now regarding the classifier method that are tested, considering the classifier used in the literature, described in subchapter 2.4, KNN, CART, NB and SVM were tested to check which will have the best results; additionally, logistic regression (LR) and linear discriminant analysis (LDA) were also tested, resulting in the 6 supervised classifiers tested as shown in Table 5.

Table 5-Classifiers methods

Classifier abbreviation	Classifier name	Classifier explanation
CART	Classification and regression trees	Splits data according to best prediction into one decision tree
KNN	K-nearest neighbour	Predict values by correlation to a certain number of proximal values
LR	Logistic regression	Predict the probability of a binary event by a linear combination
LDA	Linear discriminant analysis	Predict the probability of a binary event by Fisher's linear discriminant
NB	Naive Bayes	Prediction of probability by Bayes' theorem with kernel density estimation
SVM	Support vector machine	Linear classification to maximize the margins between classes with kernel methods

CHAPTER 4 – PROTOCOL AND DATA ANALYSES

4. Protocol and Data analyses

In this chapter, we will present the developed experimental protocol, the dataset and the results obtained from data acquisition using the protocol. The data analysis includes text and region-level analysis and feature selection.

The first subchapter described the experimental protocol developed and the second subchapter includes de dataset and data analysis.

4.1. Protocol

Two protocols were developed for this thesis, one for English text comprehension and other for code comprehension. The English comprehension protocol was developed to validate measurements of CL using questions and different difficult texts to support the results. Unfortunately, on the other hand, the code protocol was not possible to be used with volunteers as it was difficult to find volunteers with proficiency in the programming language selected, C. Since the code protocol was not implemented it will not be elaborated in the thesis.

The protocol is composed of three runs. As can be seen in Figure 9, each protocol run starts with a rest state using a grey screen, followed by the content which has the option of highlighting words using two colours. After the reading content, the user is presented with questions regarding the user experience, based on NASA-TLX questions [93]. The task ends with questions regarding the previously presented content.

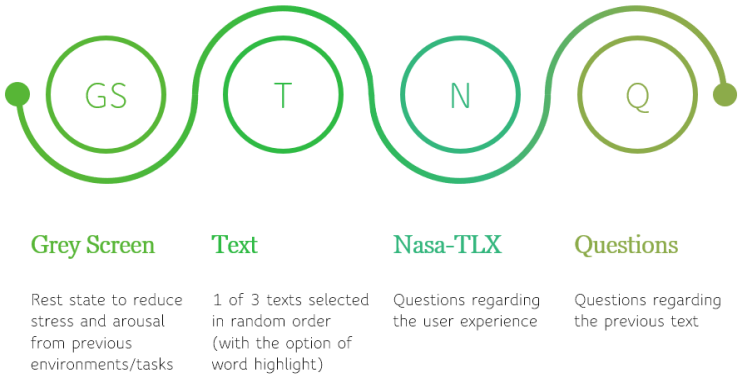


Figure 9-Schematic of each of run

For each run, every volunteer goes through each text as the tool randomly selects the order in which the three texts will appear.

With the intention of using the tool in the controlled experiments of the protocol, there was a need to add some elements to the tool. First, add a grey screen before the content, establish a baseline of signals, and also add a button that allows the user to highlight the words that the user had difficulty with (which will be used in the ML pipeline labelling). It was also added multiple HTML page to show the questions and additional tasks (translation task). To manage the moments of each page event, the timestamp in which each event takes place is sent to the server. The input of the popup of the tool, which allows to select the file to open, was replaced by a selection for either text or code protocol.

4.1.1. Text Comprehension test

Firstly, the experience starts with the tool web extension opening a new tab on the browser with a grey screen. The grey screen is used to relax the volunteer and reduce arousal and stress, it goes away automatically after 30 seconds. After the grey screen disappears the text, itself is shown with a next button and two highlight buttons.

The volunteers are instructed to read the text carefully and try to convey all of the context and information provided by the text as there will be questions afterwards. During the experience, it is possible to highlight words that the volunteer considered hard. The two highlight colours are red and yellow as red should be used for words the volunteer does not know the meaning of. Yellow highlight should be used for words that the user is not certain about the word meaning.

All text has a time limit of 5 minutes. After pressing the next button, the volunteer is prompted with questions regarding the volunteer experience during the task. This is adapted from NASA-TLX task load scales and questions. The adapted NASA-TLX questions are related to factors, such as volunteer mental effort, pressure with time, task fulfilment and discomfort and ask the volunteer to select a value on a scale from 1 to 6 for each factor. The NASA-TLX questions also include a preference question to understand which factor the volunteer considers more important.

Lastly, appears a set of multiple-choice questions regarding the text, this is done to evaluate if the volunteer understood the content.

The third task required the volunteers to translate a paragraph from the text to get more insights into the volunteer's English proficiency.

Each text come from various sources and with distinct levels of difficulty as shown in Table 6. Readability tests are used to evaluate the readability difficulty of a text. Readability test like Flesch-Kincaid readability tests [94] uses the length of words and sentences to give a score. The readability scores can be correlated with English grading systems like the Common European Framework of Reference for Languages (CEFR) and the International English Language Testing System (IELTS).

Table 6-English text characteristics

	Text Source	Flesch-Kincaid score	CEFR level	IELTS level	Score interpretation
Text 1	[95]	61.9	A2	(3-4)	“Easy”
Text 2	[96]	58.7	B2	(6-7)	“Medium”
Text 3	[97]	44.2	C2	(8-9)	“Difficult”

The readability scores in Table 6 show that from Text 1 to Text 3 there is a progressive increase in readability difficulty as the English level increase with the decrease of the Flesh-Kincaid score.

4.2.Data analysis

4.2.1. Dataset

Now regarding the data obtained from volunteers using the developed protocol described in chapter 3. The protocol was used to validate the tool. The protocol developed was created to measure comprehension difficulty in different contexts by selecting texts with different readability scores. All volunteer consent was taken previously to experiments and they were rewarded monetarily for participating. The protocol included is composed of English content. Each task involves reading 3 pieces of content with different comprehension difficulties. Texts 1, 2 and 3 are classified as A2, B2 and C2 in CEFR score, the readability score is in Table 7.

With the inclusion of self-reports for the volunteers is possible to categorize content zones to then train the ML classifiers. Regarding the data acquisition on the experimental protocol developed. The dataset is constituted of 5 volunteers. Table 7 shows information about them.

Table 7-Volunteer information

Volunteer Nº	Age	Sex	Education Level	Level of English
1	37	M	Master	C2
2	42	M	PhD	C1
3	22	M	Bachelor	C1
4	27	M	High School	C1
5	62	F	Bachelor	B2

From Figure 10 is noticeable that the pressure with time increased more significantly with Text 3 as the five-minute time limit imposed to read all content may induce more stress. Also, the fact that the readability score of Text 3 has higher than the average level of English proficiency of the volunteers also justified the description of the volunteer regarding text 3. Considering that the increase in pressure with time is also accompanied by an increase in discomfort, it is predicted that results should be more significant when comparing Text 3 to Text 1 and 2.

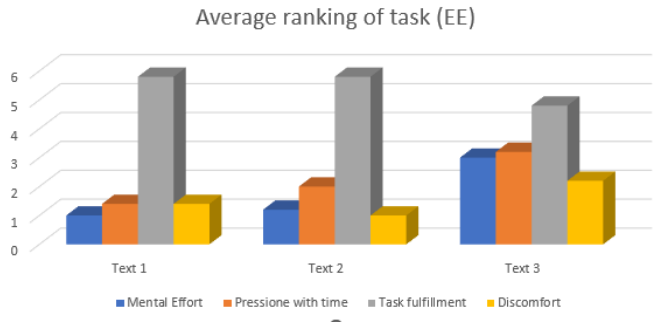


Figure 10-NASA-TLX volunteer's results

To analyse the data acquired from the volunteer is relevant to first analyse that data at text-level and then at region-level as if there is no significance in the data at text-level there cannot be at region.

4.2.2. Text-level data analysis

Comparing the statical features extracted from the volunteers, it is possible to infer the sensibility of the tool to evaluate comprehension difficulty at text and region-level and to conclude about the best features to discriminate CL. Some features did not represent a normal distribution as most of the normalized features did not pass the one-sample Kolmogorov-Smirnov test [98]. As an example, RMSSD from Text 1 had a p-value of 4.8e-49. For that reason, the data represented in statistical analysis is not standardized nor normalized.

Statistical functions like t-tests can be used with not normalised data and with extremely small sample sizes ($N \leq 5$) resulting, in that case, in a high rate of false positives. Rank-transformation test and Wilcoxon test are generally not recommended for extremely small samples [99]. The test applied in the statistical analysis was a two-sample t-test from MATLAB which uses Satterthwaite's approximation when the variance is selected as unequal, also known as the welch t-test [100].

Between Text 1 and Text 2 (easy to medium text) there are some features which visually show an apparent fluctuation of mean value from text 1 to text 3. Unfortunately, using statistical analyses none of these is statistical significance. Considering that Text 1 is easier than Text 2, we should see a reduction of the ratio of frequencies domain feature, LHFratio, which happened with 75%. Quantile. As stated in subchapter 4.2.1, LHFratio represents the ratio of sympathetic and parasympathetic branches. Using the welch t-test it was possible to obtain a p-value of 0.17, which does not prove, with significance, that these are two independent groups with unequal means. Other features like SCL presented a p-value of 0.29 with the minimum of SCL.

Between Text 2 and Text 3 (medium to hard text), there is a set of features that presented results that reject the hypothesis that the samples come from two independent groups with equal means. These results presented the minimum of SDSD and the standard deviation and maximum of LHFratio, as significant differences do exist in these samples. Figure 11 represents a Figure where each plot has Text 2 on the left and Text 3 on the right, showing as described the decrease of SDSD and increase of LHFratio with an increased CL with p-values of 0.035, 0.038 and 0.046 correspondingly.

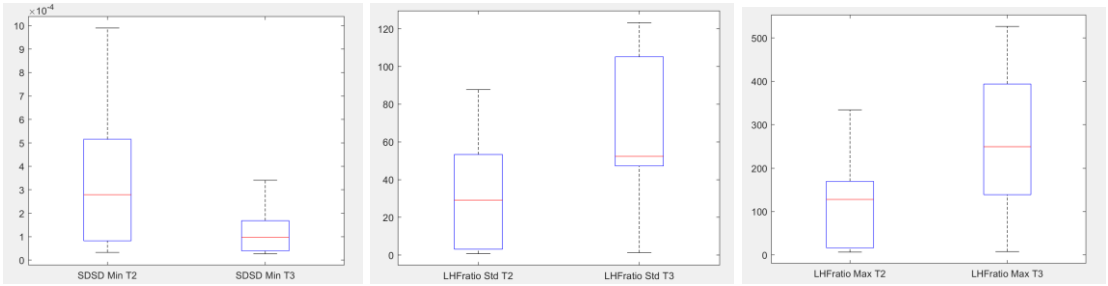


Figure 11-SDSD Min, LHFratio Std and LHFratio Max on Text 2 and 3

Between Text 1 and Text 3 (easy to hard text), only one feature showed significant results and that was, as represented in Figure 12, the 75% quantile of LHFratio with a p-value of 0.012, showing an increase in value with the increase of mental load as expected.

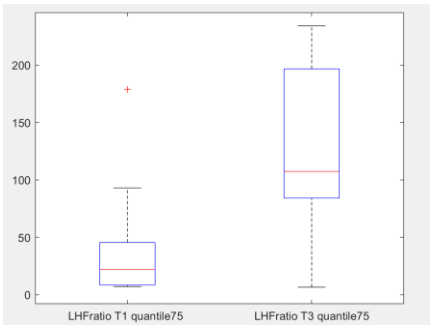


Figure 12-LHFratio 75% quantile on Text 1 and 3

At text-level, in an overall comparison of results, the best features are LHFratio and SDSD.

4.2.3. Region-level data analysis

The regions of the texts were defined by classifying paragraphs with a readability score. The readability score test used was SMOG readability score [101], which evaluates the number of words with more than three syllables per sentence. If two consecutive paragraphs had similar scores, the two paragraphs are turned into one and their score is averaged. The regions defined are represented in Table 8.

Since SMOG score does not take into consideration the total number of words, it is possible to compare the time spent without contemplating if the time spent is influenced by the size of the text. The result in Table 8 may suggest that the time spent is not influenced by the size of the text as fluctuations in the SMOG score are accompanied by similar fluctuations in the average time spent. Time spend reading may be an indicator of effort as a reduction in reading time may suggest a reduction in CL [3]. Table 8 results

would suggest that the average time spent showed an increase with an increase in the difficulty of the content on texts 2 and 3 as the increase in SMOG score indicates an increase in readability difficulty.

Table 8-Final regions, difficulty score and average time spent

Regions	Average SMOG readability score	Average time spent
Text 1 Region 1	9.63	52.2
Text 1 Region 2	10.945	46.8
Text 2 Region 1	12.115	61.6
Text 2 Region 2	11.54	56.4
Text 2 Region 3	13.265	69.8
Text 3 Region 1	13.04	73.6
Text 3 Region 2	18.145	130.8

Between the regions of Text 2, there were seven features with significant results comparing the regions 2 to 3 with the best features of those seven, but not one comparing region 1 to 2 and region 1 to 3 being shown in Table 9.

Table 9-Best features of region-level analysis

Feature	Welch t-test p-value	Regions
LHFratio 95% quantile	0.274	Region 1 to 2 of Text 1
SDSD 95% quantile	0.114	Region 1 to 2 of Text 2
SDSD min	0.0875	Region 2 to 3 of Text 2
RMSSD 95% quantile	0.0317	Region 1 to 3 of Text 2
SDSD 95% quantile	0.0265	Region 1 to 3 of Text 2
SDSD min	0.0350	Region 1 to 2 of Text 3
RMSSD min	0.0289	Region 1 to 2 of Text 3

Between regions of Text 3 p-value did show significance. Between the regions where the p-value did not show significance, which might be because of the limited data sample. From Table 9, Text 3 has the lower p-values, probably influenced by the fact that the two regions have the biggest disparity in difficulty scores between two regions.

EDA features presented higher p-values when compared to the average HR features. To understand the relationship between HR and EDA features, it is necessary to compare them. In the interest of comparing both HR and EDA, EDA sample frequency was reduced to the same as HR. Since HRV time domain features are extracted with a sliding window there is no time function associated with HRV features as each feature value corresponds to the evaluation of 30 seconds of HR values.

By comparing Figure 13 is possible to see that some event triggers a response in volunteer 1. From HRV on the left to EDA on the right the fluctuation changed, occurring in EDA about 10 seconds after HRV. Considering how a sliding window works a 10 seconds "delay" may not represent a smaller response in EDA features as it is smaller than half of the window length used in the sliding window.

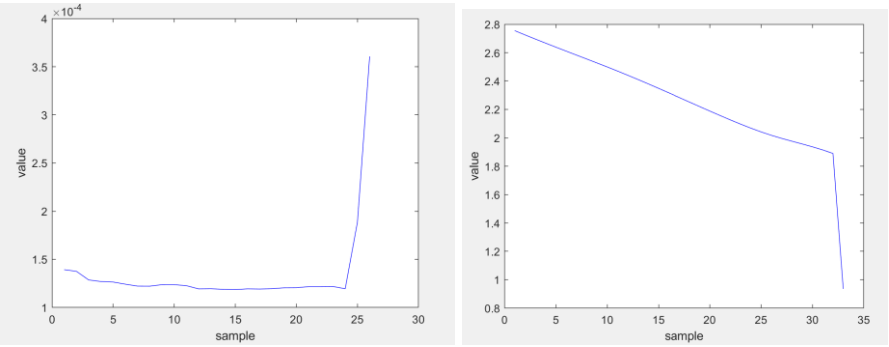


Figure 13- Region 1 Text 3 RMSSD (left) and SCL (right), Volunteer 1

As a general comparison of the result of the data analysis in region-level, it is noticeable that the same feature that obtained the best result in text-level also obtained the same results in region-level, LHFratio and SDSD. Additionally features like RMSSD also shows significant results as well as the average time spent reading.

4.2.4. Feature selection analysis

As stated, the data processing results in sixty-nine features for each of the seven content regions distinguished with this protocol. To train the classifiers, the twelve best features were selected using ANOVA features selection. Table 10 shows the twelve features with the highest f-value using ANOVA.

Table 10-ANOVA f-value feature selection

Specs	Score
TotalTimeSeconds	6.712087
SDSDMin	3.183291
LHFratioMin	2.816770
RMSSDMin	2.691713
SDSDMean	2.380576
SDSDMedian	2.369680
SDSDQuantile50	2.369680
SCLMin	2.151741
SDSDQuantile75	2.138913
SCRMin	2.112393
RMSSDMean	1.994668
SCLMedian	1.930422

Table 12 complies with the previous data analysis as features like the minimum of SDSD appear twice in the best region-level features table (Table 8) and appear here as the second best features. The best feature according to this test was the total time spent on that region, which can be interpreted as the volunteer's time "investment" in comprehending the content.

To validate the features selected a correlation matrix is used to compare the best features and evaluate dependencies. The following heatmap is a correlation matrix between the features and features/labels as shown in Figure 14.

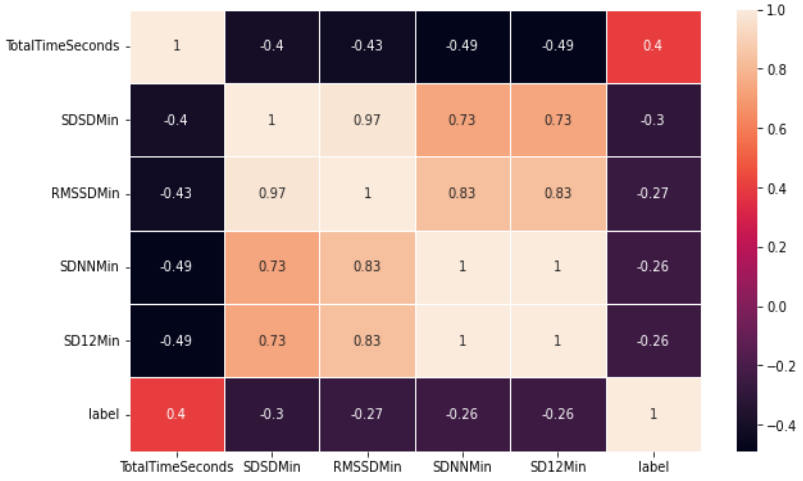


Figure 14-Correlation Matrix of ANOVA selected features

As we can see in Figure 14, there are many dependencies between the features, therefore, we used a different approach called "forward feature selection".

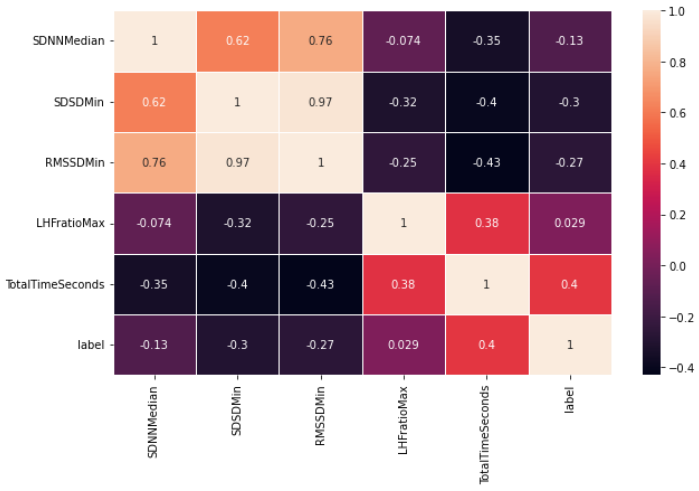


Figure 15-Correlation Matrix of Forward Feature Selection

In forward feature selection, n models are trained on n features individually, where n is the number of features. The best performing feature in terms of accuracy is ranked the 1st. After that, we add each feature to the best feature and analyze the results [102].

In Figure 15, we can see how the dependencies among features decreased and increased on the label in accordance using forward feature selection.

CHAPTER 5 – RESULTS AND DISCUSSIONS

5. Results and Discussion

On data analysis on region-level the minimum of SDSD and RMSSD and the 95% quantile of RMSSD and SDSD were the best features, between some regions of the same text. Features selection show that the time spent in each region and the HRV features the minimum of LHFratio and SDRR are, from the group of features extracted, among the best features in terms of variance and dependencies. EDA features like minimum of SCL and SCR are also included in the best features, as well as the RMSSD, from HRV.

Considering the extremely low dataset, hyperparameter tuning will not be explored, nonetheless, to evaluate the performance of the selected models cross-validation was used. Classifiers overall performance indicated that classifiers like LR and SVM lead to good results and as the best classifier in the f-score and accuracy was LR as shown in Figure 16. The results were rather low in CART describing the worst prediction.

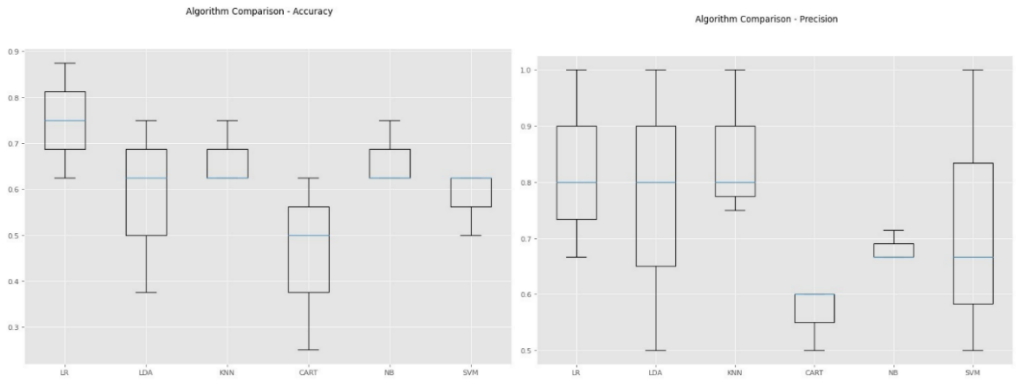


Figure 16-Accuracy and Precision in overall classifiers

Overall precision was close to recall showing that the balanced of true positives was similar to the predictive positive in all observations as shown in Figure 17.

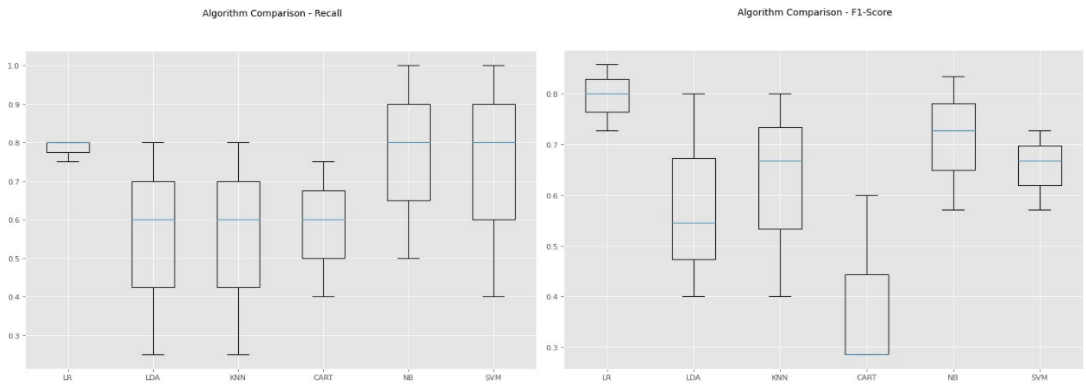


Figure 17- Recall and F-score in overall classifiers

Considering the low sampling, the prediction results can be produced by luck for that reason the permutation test score is used to create a distribution of classifier accuracy and a p-value is taken. The permutation score classification works by shuffling the labels and keeping the features as they are. After doing that, the performance of the shuffled data classifier is compared statistically with the original data classifier. After doing that, the results show a p-value of 0.0198 with SVM, as shown in Figure 18, which does show that the classifier can utilize the dependent variable to obtain good results.

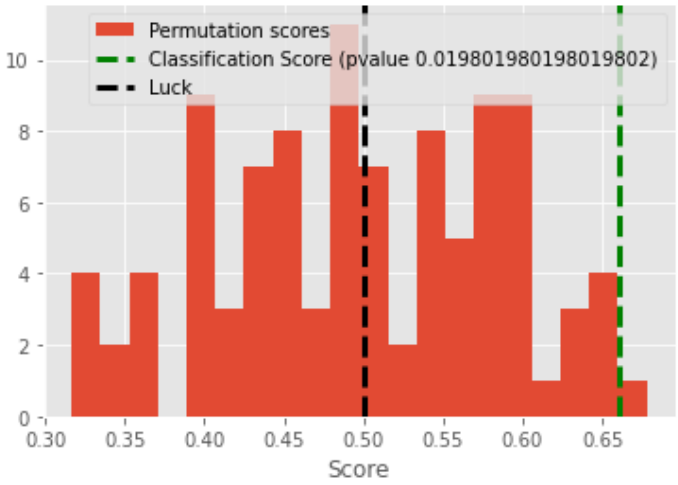


Figure 18-Permutation Scores Classification

The classification score is not much better than the score from permutations, but it does show an adequate p-value. To get better insights into the true positive rate and the false positive rate (the ability of the tool to correctly detect the comprehension difficulties), the analysis of the Receiver operating characteristic curve (ROC curve) of one of the classifiers (LR) was performed as showed in Figure 19.

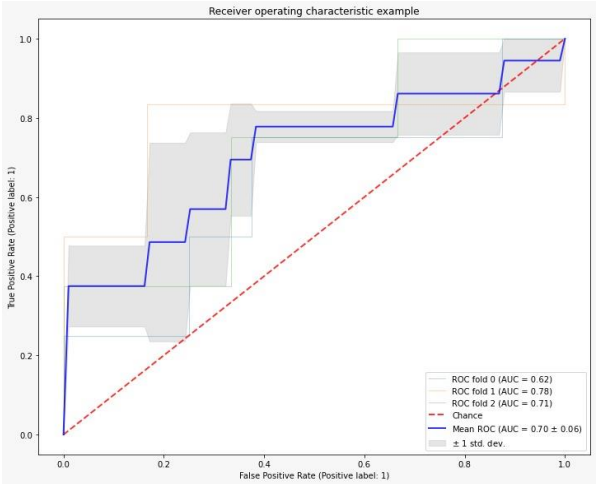


Figure 19-ROC Curve

Figure 19 shows 3-fold cross-validation as leave-one-out would not generalize with such a small dataset.

We notice that the mean ROC (AUC) is relatively good and converges to acceptable specificity and sensitivity.

Although the ML model is now established, we are still gathering more data, and we expect to find skews between the trained model and the deployed model when we install the tool in a realistic environment. Complementary the acquisition of more data will allow for a more realistic assessment of performance and comparison the classifier's results. However, with careful training and feedback, we will, most probably, achieve a stable model in terms of performance.

CHAPTER 6 – CONCLUSION AND FUTURE WORK

6. Conclusion

This thesis work was focused on developing an intelligent tool using a web extension that would allow the acquisition of biometric features synchronized with eye gaze and content location (scroll). The tool (called iMind) goal is to predict comprehension difficulties based on biometric and non-biometric data at region-level. The data acquired is fed to an ML module to classify content regions according to the associated features.

The tool used a smartwatch and a low-cost desktop eye tracker to acquire HRV and EDA features and eye tracking features. The smartwatch uses PPG sensors to measure BVP which is transformed into RR intervals. The same smartwatch measured EDA by two electrodes. The use of BVP can lead to more artifacts but according to the data analysis, there was significance in the results of these features.

Most features acquired correlate to one or both branches of the ANS. ANS being responsible to regulate involuntary systems in the body, create physiological fluctuation that can be a consequence of near overload CL or high mental stress.

Literature analysis showed that, to our extent, at this moment, no paper developed a tool based on the assessment of CL at region-level with the used of low-cost wearable devices as this thesis have done.

To validate the tool a protocol was developed in which volunteers would read multiple texts and respond to questions. The protocol was only used in 5 volunteers but still, there is significance between the data considering the results.

Feature selection showed that, from the features extracted, the best features are reading time and HRV features, like SDSD, LHFratio and RMSSD. The features selected also included both SCL and SCR features from EDA features, which, in a general comparison, had inferior scores to the HR features.

From the different classifiers tested, LR has the best overall results with the ROC curve analysis showing a relatively good balance of sensitivity and specificity. Even though the acceptable results, the extremely small sample does not allow for adequate evaluation of performance and more data is needed to test if the selected models can generalize to larger populations.

This work limitation is centered on the fact that the tool was only tested in a limited number of volunteers forcing the enlargement of the dataset for more adequate conclusions regarding its performance. The complexity of using the smartwatch software, is a limiting for the implementation of the tool. Considering these two points, the future directions of this tool could involve around developing a simplified “live mode” that allows for real-time prediction with the inclusion of more sophisticated ML and Deep Learning methods, needing for that a bigger dataset. The test of this tool in realistic environments, like training and learning centers or the classification of an intermediate CL state, could also be important steps in future work.

REFERENCES

[1] S. Peng and K. Nagao, “Recognition of Students’ Mental States in Discussion Based on Multimodal Data and its Application to Educational Support,” *IEEE Access*, vol. 9, pp. 18235–18250, 2021, doi: 10.1109/ACCESS.2021.3054176.

[2] C. Larmuseau, J. Cornelis, L. Lancieri, P. Desmet, and F. Depaepe, “Multimodal learning analytics to investigate cognitive load during online problem solving,” *Br. J. Educ. Technol.*, vol. 51, no. 5, pp. 1548–1562, 2020, doi: 10.1111/bjet.12958.

[3] J. Sweller, *Cognitive Load Theory*, vol. 55. Elsevier Inc., 2011.

[4] B. O’Rourke, “Whose Language Is It? Struggles for Language Ownership in an Irish Language Classroom,” *J. Lang. Identity Educ.*, vol. 10, no. 5, pp. 327–345, 2011, doi: 10.1080/15348458.2011.614545.

[5] H. A. Valdecantos, K. Tarrit, M. Mirakhorli, and J. O. Coplien, “An Empirical Study on Code Comprehension: Data Context Interaction Compared to Classical Object Oriented,” *IEEE Int. Conf. Progr. Compr.*, no. May, pp. 275–285, 2017, doi: 10.1109/ICPC.2017.23.

[6] J. Bennedsen and M. E. Caspersen, “Failure rates in introductory programming - 12 years later,” *ACM Inroads*, vol. 10, no. 2, pp. 30–35, 2019, doi: 10.1145/3324888.

[7] P. Gobel and S. Mori, “Success and failure in the EFL classroom,” *EUROSLA Yearb.*, vol. 7, pp. 149–169, 2007, doi: 10.1075/eurosla.7.09gob.

[8] EF, “Ranking of 112 Countries and Regions by English Skills,” *EF-Education First*, pp. 1–38, 2021, [Online]. Available: https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2021/ef-epi-2021-english.pdf.

[9] I. del Arco, P. Silva, and O. Flores, “University teaching in times of confinement: The light and shadows of compulsory online learning,” *Sustain.*, vol. 13, no. 1, pp. 1–16, 2021, doi: 10.3390/su13010375.

[10] J. Barbosa, A. Silva, M. A. Ferreira, and M. Severo, “The impact of students and curriculum on self-study during clinical training in medical school: a multilevel

approach," *BMC Med. Educ.*, vol. 17, no. 1, pp. 1–7, 2017, doi: 10.1186/s12909-016-0846-3.

[11] H. Hijazi, R. Couceiro, J. Castelhano, P. De Carvalho, M. Castelo-Branco, and H. Madeira, "Intelligent biofeedback augmented content comprehension (tellback)," *IEEE Access*, vol. 9, pp. 28393–28406, 2021, doi: 10.1109/ACCESS.2021.3058664.

[12] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educ. Psychol.*, vol. 38, no. 1, pp. 63–71, 2003, doi: 10.1207/S15326985EP3801_8.

[13] J. F. Lehman, R. L. Lewis, and A. Newell, "Integrating Knowledge Sources in Language Comprehension," *Science (80-.).*, 1990.

[14] C. Smallwood, "Reading comprehension," *Paris Rev.*, no. 206, pp. 133–141, 2013, doi: 10.1177/2372732215624707.

[15] W. Kintsch, *Comprehension: A paradigm for cognition*, Third edit. Cambridge university press, 1998.

[16] R. Cates, N. Yunik, and D. G. Feitelson, "Does Code Structure Affect Comprehension? On Using and Naming Intermediate Variables," *IEEE Int. Conf. Progr. Compr.*, vol. 2021-May, pp. 118–126, 2021, doi: 10.1109/ICPC52881.2021.00020.

[17] S. Long, "Tuning in to teacher–talk: a second language learner struggles to comprehend," *Reading*, vol. 36, no. 3, pp. 113–118, 2002, doi: 10.1111/1467-9345.00197.

[18] U. Nikula, O. Gotel, and J. Kasurinen, "A motivation guided holistic rehabilitation of the first programming course," *ACM Trans. Comput. Educ.*, vol. 11, no. 4, 2011, doi: 10.1145/2048931.2048935.

[19] B. F. Nguyen and R. C. Clark, "Practical Applications of Technology for Learning," *Learning*, 2005.

[20] A. S. P. Jansen, X. Van Nguyen, V. Karpitskiy, T. C. Mettenleiter, and A. D. Loewy, "Central command neurons of the sympathetic nervous system: Basis of the fight-or-flight response," *Science (80-.).*, vol. 270, no. 5236, pp. 644–646, 1995, doi: 10.1126/science.270.5236.644.

[21] F. G. Paas, J. J. Van Merriënboer, and J. J. Adam, “Measurement of cognitive load in instructional research.,” *Percept. Mot. Skills*, vol. 79, no. 1 Pt 2, pp. 419–430, 1994, doi: 10.2466/pms.1994.79.1.419.

[22] M. Irwin and D. Morgenstern, “Anatomy and physiology of the nervous system,” *Canadian cancer society*, 2020. https://cancer.ca/en/cancer-information/cancer-types/neuroblastoma/what-is-neuroblastoma/the-nervous-system.

[23] C. B. Saper, “The central autonomic nervous system: Conscious visceral perception and autonomic pattern generation,” *Annu. Rev. Neurosci.*, vol. 25, pp. 433–469, 2002, doi: 10.1146/annurev.neuro.25.032502.111311.

[24] D. H. Evans, *The physiology of FISHES 2nd edition*. 1998.

[25] Y. Gao *et al.*, “Cattle encephalon glycoside and ignotin injection improves cognitive impairment in APPswe/PS1dE9 mice used as multitarget anti-Alzheimer’s drug candidates,” *Neuropsychiatr. Dis. Treat.*, vol. 11, pp. 537–548, 2015, doi: 10.2147/NDT.S78025.

[26] V. Sandhu, “Evaluation of Learning Performance by Quantifying User’s Engagement,” pp. 180–183, 2017.

[27] A. Baddeley and G. Hitch, “The social design of virtual worlds: constructing the user and community through code,” *Med. Res. Counc.*, pp. 47–88, 1974.

[28] V. Sandhu, A. A. P. Wai, and C. Y. Ho, “Evaluation of learning performance by quantifying user’s engagement investigation through low-cost multi-modal sensors,” *Proc. 2017 Int. Conf. Orange Technol. ICOT 2017*, vol. 2018-Janua, no. December, pp. 180–183, 2018, doi: 10.1109/ICOT.2017.8336117.

[29] L. Yan, Y. Wang, C. Ding, M. Liu, F. Yan, and K. Guo, “Correlation among behavior, personality, and electroencephalography revealed by a simulated driving experiment,” *Front. Psychol.*, vol. 10, no. July, pp. 1–16, 2019, doi: 10.3389/fpsyg.2019.01524.

[30] M. Gjoreski *et al.*, “Datasets for cognitive load inference using wearable sensors and psychological traits,” *Appl. Sci.*, vol. 10, no. 11, 2020, doi: 10.3390/app10113843.

[31] S. L. Beilock and T. H. Carr, “When high-powered people fail: Working memory and ‘Choking under pressure’ in math,” *Psychol. Sci.*, vol. 16, no. 2, pp. 101–105,

2005, doi: 10.1111/j.0956-7976.2005.00789.x.

[32] R. Xiong, F. Kong, X. Yang, G. Liu, and W. Wen, “Pattern Recognition of Cognitive Load Using EEG and ECG Signals,” *MDPI*, pp. 1–13, 2020.

[33] A. Dan and M. Reiner, “Real Time EEG Based Analysis of Cognitive Load Enhance Instructional Analysis,” *J. Educ. Data Min.*, vol. 9, no. 2, pp. 31–44, 2017, [Online]. Available: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/160.

[34] C. H. Tan *et al.*, “Optical measures of changes in cerebral vascular tone during voluntary breath holding and a Sternberg memory task,” *Biol. Psychol.*, vol. 118, pp. 184–194, 2016, doi: 10.1016/j.biopsycho.2016.05.008.

[35] 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 (Switzerland)*, vol. 19, no. 17, pp. 1–16, 2019, doi: 10.3390/s19173729.

[36] A. M. Hughes, G. M. Hancock, S. L. Marlow, K. Stowers, and E. Salas, “Cardiac Measures of Cognitive Workload: A Meta-Analysis,” *Hum. Factors*, vol. 61, no. 3, pp. 393–414, 2019, doi: 10.1177/0018720819830553.

[37] S. Solhjoo *et al.*, “Heart Rate and Heart Rate Variability Correlate with Clinical Reasoning Performance and Self-Reported Measures of Cognitive Load,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019, doi: 10.1038/s41598-019-50280-3.

[38] N. Herbig, S. Pal, M. Vela, A. Krüger, and J. van Genabith, “Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation,” *Mach. Transl.*, vol. 33, no. 1, pp. 91–115, 2019, doi: 10.1007/s10590-019-09227-8.

[39] C. Wang and J. Guo, “A data-driven framework for learners’ cognitive load detection using ECG-PPG physiological feature fusion and XGBoost classification,” *Procedia Comput. Sci.*, vol. 147, pp. 338–348, 2019, doi: 10.1016/j.procs.2019.01.234.

[40] A. Singh Vijoriya and R. Maheshwari, “ECG Signal Acquisition, Feature Extraction and HRV Analysis Using Biomedical Workbench,” *Int. J. Adv. Res. Eng. Technol. (IJARET*, vol. 9, no. 3, pp. 84–90, 2018, [Online]. Available:

http://www.iaeme.com/IJARET/index.asp84http://www.iaeme.com/ijaret/issues.asp?JType=IJARET&VType=9&IType=3http://www.iaeme.com/ijaret/issues.asp?JType=IJARET&VType=9&IType=3.

[41] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 4289–4294, 2006, doi: 10.1109/IEMBS.2006.260607.

[42] M. Zubair and C. Yoon, "Multilevel mental stress detection using ultra-short pulse rate variability series," *Biomed. Signal Process. Control*, vol. 57, p. 101736, 2020, doi: 10.1016/j.bspc.2019.101736.

[43] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *Int. J. Cardiol.*, vol. 166, no. 1, pp. 15–29, 2013, doi: 10.1016/j.ijcard.2012.03.119.

[44] N. D. Giardino, P. M. Lehrer, and R. Edelberg, "Comparison of finger plethysmograph to ECG in the measurement of heart rate variability," *Psychophysiology*, vol. 39, no. 2, pp. 246–253, 2002.

[45] E. Gil, M. Orini, R. Bailón, J. M. Vergara, L. Mainardi, and P. Laguna, "Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions," *Physiol. Meas.*, vol. 31, no. 9, pp. 1271–1290, 2010, doi: 10.1088/0967-3334/31/9/015.

[46] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tr, and U. Ehlert, "Discriminating Stress From Cognitive Load Using a Wearable EDA Device," *Technology*, vol. 14, no. 2, pp. 410–417, 2010.

[47] J. J. J. Braithwaite *et al.*, "A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments," …, pp. 1–42, 2013, [Online]. Available: http://www.bhamlive.bham.ac.uk/Documents/college-les/psych/saal/guide-electrodermal-activity.pdf%5Cnhttp://www.birmingham.ac.uk/documents/college-les/psych/saal/guide-electrodermal-activity.pdf%0Ahttps://www.birmingham.ac.uk/Documents/college-les/psych/sa.

[48] V. Farrell, Y. Wang, F. Chen, and R. Calvo, “Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks,” *OZCHI*, no. 2, p. 692, 2015.

[49] P. Ayres, J. Y. Lee, F. Paas, and J. J. G. van Merriënboer, “The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load,” *Front. Psychol.*, vol. 12, no. September, 2021, doi: 10.3389/fpsyg.2021.702538.

[50] J. S. Kang, A. Ojha, and M. Lee, “Development of intelligent learning tool for improving foreign language skills based on EEG and eye tracker,” *HAI 2015 - Proc. 3rd Int. Conf. Human-Agent Interact.*, pp. 121–126, 2015, doi: 10.1145/2814940.2814951.

[51] E. Stuyven, K. Der Van Goten, A. Vandierendonck, K. Claeys, and L. Crevits, “The effect of cognitive load on saccadic eye movements,” *Acta Psychol. (Amst).*, vol. 104, no. 1, pp. 69–85, 2000, doi: 10.1016/S0001-6918(99)00054-2.

[52] M. Ntodie, S. R. Bharadwaj, S. Balaji, K. J. Saunders, and J. A. Little, “Comparison of Three Gaze-position Calibration Techniques in First Purkinje Image-based Eye Trackers,” *Optom. Vis. Sci.*, vol. 96, no. 8, pp. 587–598, 2019, doi: 10.1097/OPX.0000000000001405.

[53] K. Holmqvist and R. Andersson, “Eyetracking: A comprehensive guide to methods, paradigms and measures,” no. March, p. 537, 2017.

[54] W. van Winsum, J. Sergeant, and R. Geuze, “The functional significance of event-related desynchronization of alpha rhythm in attentional and activating tasks,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 58, no. 6, pp. 519–524, 1984, doi: 10.1016/0013-4694(84)90042-7.

[55] M. O. Candela-leal *et al.*, “Real-time Biofeedback System for Interactive Learning using Wearables and IoT Real-time Biofeedback System for Interactive Learning using Wearables and IoT,” no. November, 2021.

[56] E. Di Lascio, S. Gashi, and S. Santini, “Unobtrusive Assessment of Students’ Emotional Engagement during Lectures Using Electrodermal Activity Sensors,” *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–21, 2018, doi: 10.1145/3264913.

[57] S. Bianco and P. Napoletano, “Biometric Recognition Using Multimodal

Physiological Signals," *IEEE Access*, vol. 7, pp. 83581–83588, 2019, doi: 10.1109/ACCESS.2019.2923856.

[58] V. Markova, T. Ganchev, and K. Kalinkov, "CLAS: A Database for Cognitive Load, Affect and Stress Recognition," *Proc. Int. Conf. Biomed. Innov. Appl. BIA 2019*, 2019, doi: 10.1109/BIA48344.2019.8967457.

[59] W. L. Romine *et al.*, "Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: Development of a cognitive load tracker for both personal and cla," *Sensors (Switzerland)*, vol. 20, no. 17, pp. 1–14, 2020, doi: 10.3390/s20174833.

[60] A. Abbad-Andaloussi, T. Sorg, and B. Weber, "Estimating Developers' Cognitive Load at a Fine-grained Level Using Eye-tracking Measures," *IEEE Int. Conf. Progr. Compr.*, vol. 2022-March, pp. 111–121, 2022, doi: 10.1145/3524610.3527890.

[61] M. I. Ahmad, I. Keller, D. A. Robb, and K. S. Lohan, "A framework to estimate cognitive load using physiological data," *Pers. Ubiquitous Comput.*, 2020, doi: 10.1007/s00779-020-01455-7.

[62] C. Mills, J. Gregg, R. Bixler, and S. K. D'Mello, "Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering," *Human-Computer Interact.*, vol. 36, no. 4, pp. 306–332, 2021, doi: 10.1080/07370024.2020.1716762.

[63] I. Keller, M. I. Ahmad, and K. Lohan, "Multi-Modal Measurements of Mental Load," *Conference-W12*, no. April 2019, 2019, [Online]. Available: http://arxiv.org/abs/1906.10557.

[64] T. Čegovnik, K. Stojmenova, G. Jakus, and J. Sodnik, "An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers," *Appl. Ergon.*, vol. 68, no. March 2017, pp. 1–11, 2018, doi: 10.1016/j.apergo.2017.10.011.

[65] Empatica, "E4 wristband technical specifications," 2020. https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications.

[66] Tobii, “Tobii Eye Tracker datasheet.”

[67] K. Schoolov, “Google is winning in education, but Apple and Microsoft are battling for market share,” *cnbc*, 2019. https://www.cnbc.com/2019/03/20/apple-google-microsoft-are-battling-for-dominance-in-education.html#:~:text=In 2018%2C Chromebooks made up,to data from Futuresource Consulting.

[68] Imascientist, “Which devices and browsers are UK schools using?,” *IMS*, 2019. https://about.imascientist.org.uk/2019/which-devices-and-browsers-are-uk-schools-using/.

[69] W3counter, “Browser & Platform Market Share,” *w3counter*, 2021. https://www.w3counter.com/globalstats.php?year=2021&month=9.

[70] H. Shah and T. R. Soomro, “Node. Js Challenges in Implementation,” *Glob. Journals Comput. Sci. Technol. E Network, Web Secur.*, vol. 17, no. 2, pp. 73–83, 2017.

[71] R. Fielding *et al.*, “Hypertext transfer protocol – HTTP/1.1,” *Internet Eng. Task Force*, vol. 1, no. 11, pp. 1829–1841, 1999, [Online]. Available: https://www.ietf.org/rfc/rfc2616.txt.

[72] Empatica, “E4 data - BVP expected signal,” 2020. https://support.empatica.com/hc/en-us/articles/360029719792-E4-data-BVP-expected-signal.

[73] A. A. T. Schuurmans *et al.*, “Validity of the Empatica E4 Wristband to Measure Heart Rate Variability (HRV) Parameters: a Comparison to Electrocardiography (ECG),” *J. Med. Syst.*, vol. 44, no. 11, 2020, doi: 10.1007/s10916-020-01648-w.

[74] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, “An eye-tracking study of website complexity from cognitive load perspective,” *Decis. Support Syst.*, vol. 62, pp. 1–10, 2014, doi: 10.1016/j.dss.2014.02.007.

[75] T. Čegovnik, K. Stojmenova, G. Jakus, and J. Sodnik, “An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers,” *Appl. Ergon.*, vol. 68, no. October 2017, pp. 1–11, 2018, doi: 10.1016/j.apergo.2017.10.011.

[76] M. Q. Khan and S. Lee, “Gaze and eye tracking: Techniques and applications in ADAS,” *Sensors (Switzerland)*, vol. 19, no. 24, 2019, doi: 10.3390/s19245540.

[77] Task Force of The European Society of Cardiology, “Guidelines Heart rate variability,” *Eur. Hear. J.*, vol. 17, pp. 354–381, 1996, doi: 10.1161/01.CIR.93.5.1043.

[78] M. Benedek and C. Kaernbach, “A continuous measure of phasic electrodermal activity,” *J. Neurosci. Methods*, vol. 190, no. 1, pp. 80–91, 2010, doi: 10.1016/j.jneumeth.2010.04.028.

[79] J. Gwizdka and Y. Zhang, “Differences in Eye-Tracking Measures Between Visits and Revisits to Relevant and Irrelevant Web Pages,” pp. 811–814, 2015, doi: 10.1145/2766462.2767795.

[80] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky, “Eye tracking in web search tasks: Design implications,” *Eye Track. Res. Appl. Symp.*, no. 650, pp. 51–58, 2002.

[81] D. S. Niederhauser, R. E. Reynolds, D. J. Salmen, and P. Skolmoski, “The influence of cognitive load on learning from hypertext,” *J. Educ. Comput. Res.*, vol. 23, no. 3, pp. 237–255, 2000, doi: 10.2190/81BG-RPDJ-9FA0-Q7PA.

[82] P. D. Antonenko and D. S. Niederhauser, “The influence of leads on cognitive load and learning in a hypertext environment,” *Comput. Human Behav.*, vol. 26, no. 2, pp. 140–150, 2010, doi: 10.1016/j.chb.2009.10.014.

[83] K. A. Longstaffe, B. M. Hood, and I. D. Gilchrist, “The influence of cognitive load on spatial search performance,” *Attention, Perception, Psychophys.*, vol. 76, no. 1, pp. 49–63, 2014, doi: 10.3758/s13414-013-0575-1.

[84] T. M. Harrison and R. Brown, “Autonomic Nervous System Function after a Skin-to-Skin Contact Intervention in Infants with Congenital Heart Disease,” *J. Cardiovasc. Nurs.*, vol. 32, no. 5, pp. E1–E13, 2017, doi: 10.1097/JCN.0000000000000397.

[85] K. Tajane, R. Pitale, and J. Umale, “Non-Linear Feature Extraction for Heart Rate Variability: An Overview,” *Int. J. Comput. Appl.*, vol. 89, no. 10, pp. 17–19, 2014, doi: 10.5120/15667-4068.

[86] G. Gautam, S. Shrestha, and S. Cho, “Spectral Analysis of Rectangular, Hanning, Hamming and Kaiser Window for Digital Fir Filter,” *Int. J. Adv. smart Converg.*, vol. 4, no. 2, pp. 138–144, 2015, doi: 10.7236/ijasc.2015.4.2.138.

[87] J. M. Medeiros, “Development of a Heart Rate Variability analysis tool,” 2010, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.460.2891&rep=rep1&type=pdf.

[88] and L. C. G. Landoni, A. Zangrillo, “Annual Update in Intensive Care and Emergency Medicine 2012 Non- invasive Ventilation Outside the ICU,” *Updat. Intensive Care Emerg. Med.*, no. January 2011, pp. 29–37, 2012, doi: 10.1007/978-3-642-25716-2.

[89] T. Thong, K. Li, J. McNames, M. Aboy, and B. Goldstein, “Accuracy of Ultra-Short Heart Rate Variability Measures,” *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, vol. 3, pp. 2424–2427, 2003, doi: 10.1109/iembs.2003.1280405.

[90] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, “CvxEDA: A convex optimization approach to electrodermal activity processing,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, 2016, doi: 10.1109/TBME.2015.2474131.

[91] S. Subramanian, R. Barbieri, and E. N. Brown, “A Systematic Method for Preprocessing and Analyzing Electrodermal Activity,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 6902–6905, 2019, doi: 10.1109/EMBC.2019.8857757.

[92] M. Buchwald, S. Kupinski, A. Bykowski, J. Marcinkowska, D. Ratajczyk, and M. Jukiewicz, “Electrodermal activity as a measure of cognitive load: A methodological approach,” *Signal Process. - Algorithms, Archit. Arrange. Appl. Conf. Proceedings, SPA*, vol. 2019-Septe, pp. 175–179, 2019, doi: 10.23919/SPA.2019.8936745.

[93] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” *Adv. Psychol.*, vol. 52, no. C, pp. 139–183, 1988, doi: 10.1016/S0166-4115(08)62386-9.

[94] S. Studies and S. Bilgiler, “Evaluating Text Complexity and Flesch-Kincaid Grade Level,” *J. Soc. Stud. Educ. Res.*, vol. 8, no. 3, pp. 238–248, 2017.

[95] “A2 Key Reading Text and Questions,” *Exam English Ltd.*, 2020. https://www.examenglish.com/KET/KET_reading_part3_2020.html.

[96] “B2 Key Reading Text and Questions,” *Exam English Ltd.*, 2021.

https://www.examenglish.com/B2/b2_reading_health.htm.

[97] S. Whitelam, “Defensive scientific writing,” *Linacre Lines*, p. 23, 2003, [Online]. Available: https://www.linacre.ox.ac.uk/sites/default/files/michaelmas_2003.pdf.

[98] R. H. C. Lopes, I. Reid, and P. R. Hobson, “The two-dimensional Kolmogorov-Smirnov test,” *Proc. Sci.*, vol. 50, 2007.

[99] J. C. F. de Winter, “Using the student’s t-test with extremely small sample sizes,” *Pract. Assessment, Res. Eval.*, vol. 18, no. 10, pp. 1–12, 2013.

[100] N. A. Ahad and S. S. S. Yahaya, “Sensitivity analysis of Welch’s t -test,” *AIP Conf. Proc.*, vol. 1605, no. February 2015, pp. 888–893, 2014, doi: 10.1063/1.4887707.

[101] R. Formula, G. Harry, and M. Laughlin, “SMOG Grading-a New,” *Source J. Read.*, vol. 12, no. 8, pp. 639–646, 1969.

[102] V. Kumar, “Feature Selection: A literature Review,” *Smart Comput. Rev.*, vol. 4, no. 3, 2014, doi: 10.6029/smartcr.2014.03.007.