Study of stress detection and proposal of stress-related features using commercial-off-the-shelf wrist wearables

Francisco de Arriba-Pérez*, Juan M. Santos-Gago and Manuel Caeiro-Rodríguez, Mateo Ramos-Merino
Department of Telematics Engineering, University of Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Galicia | Spain
farriba@uvigo.es, Juan.Santos@det.uvigo.es, Manuel.Caeiro@det.uvigo.es, mateo.ramos@gist.uvigo.es
*Correspondence: farriba@uvigo.es; Tel.: +34 986 814 073

Abstract

This paper discusses the possibility of detecting personal stress making use of popular wearable devices available in the market. Different instruments found in the literature to measure stress-related features are reviewed, distinguishing between subjective tests and mechanisms supported by the analysis of physiological signals from clinical devices. Taking them as a reference, a solution to estimate stress based on the use of commercial-off-the-shelf wrist wearables and machine learning techniques is described. A mobile app was developed to induce stress in a uniform and systematic way. The app implements well-known stress inducers, such as the Paced Auditory Serial Addition Test, the Stroop Color-Word Interference Test, and a hyperventilation activity. Wearables are used to collect physiological data used to train classifiers that provide estimations on personal stress levels. The solution has been validated in an experiment involving 19 subjects, offering an average accuracy and F-measures close to 0.99 in an individual model and an accuracy and F-measure close to 0.85 in a global 2-level classifier model. Stress can be a worrying problem in different scenarios, such as in educational settings. Thus, the last part of the paper describes the proposal of a set of stress related indicators aimed to support the management of stress over time in such settings.

Keywords: COTS wrist wearables; stress quantification; wearables analytics; wearable stress detection.

1. Introduction

Selye defined stress as the adaptive response of our body to stressful events (Selye 1973). Usually, stress is seen as a negative thing or even harmful, but from a pragmatic point of view, it can be positive under certain circumstances. In the medical literature, stress featured as “distress” is considered harmful (García-Ros et al. 2012), affecting to the thoughts, demotivating, contributing to give up hope, etc. If stress is produced on a regular basis, the risk to develop a physiological or psychological condition increases. Some problems related to stress in working environments are (Colligan and Higgins 2006): absenteeism, truancy, organizational dysfunction, a decrease in productivity, etc. Meanwhile, also in the medical literature, “eustress” is considered as good stress that can be positive for the person (García-Ros et al. 2012). Experienced in small chunks at the beginning of activities, this kind of stress can support motivation and favor positive thoughts, that will contribute to achieve goals. Therefore, stress is an important issue and it deserves to be measured and managed.

Several proposals can be found in the scientific literature to detect, measure and prevent stress, even solutions to manage it (Cooper and Cartwright 1997; Kompier and Cooper 1999; Sandhu et al. 2015). In the educational domain, which is the focus of this paper, existing pieces of research can be classified into two broad categories: focused on students or on academic staff. There exist problems that affect both types of users in a similar way, such as the Burnout

syndrome (Maslach et al. 1986b; Travers and Cooper 1997). This syndrome appears as a result of attaining too demanding challenges with scarce resources, which produces a constant over-stress in work (Gonzáles–Romá et al. 2002). The Burnout syndrome can be identified by the following symptoms: emotional exhaustion, depersonalization-effrontery of the tasks and work, and low personal development and professional effectiveness. In case of the students, usual stressful activities are (García-Ros et al. 2012): "classroom's presentations" (46,7%), "lack of time to perform the academic assignments" (41,7%), "academic overload" (39,2%) and "performing exams" (35,7%). The effects of stress can be found in students, particularly in the first-year college students (Lu 1994; Deberard et al. 2004), where the general causes are combined with other components that increase the stress: changes at social and family level, adoption of new learning strategies, differences in assessment methods, etc. This may explain why between a 20% and 30% of the students drop out during the first year of college studies (Deberard et al. 2004; Kitsantas et al. 2008). To reduce the Burnout syndrome, an emotional intelligence helps students to tackle it in a more effective way their goals (Extremera et al. 2007). Proper management of stress allows to learn faster and solve problems to achieve success.

Our final goal is to propose a solution to monitor and manage stress in educational contexts providing stress-related indicators. This is intended in a hassle-free and non-intrusive way to avoid unnecessary burdens on learners and on the learning process. There already exist some works that have tried to estimate stress in the educational domain based on different methods: focusing on keystrokes and linguistic features (Vizer et al. 2009), or using specialized devices (Kikhia et al. 2015; Sano and Eng 2016; Costa et al. 2019). By contrast, our approach is based on the use of nowadays popular commercial-off-the-shelf (COTS) wrist wearables. These devices are widely available in the market and their sales grow significantly each year (IDC 2017; Statista 2017, 2018). They are particularly popular among students as a result of their good features and prices. These devices allow us to collect physiological information from the student through sensors, such as the heart rate sensor or the accelerometer, in a non-intrusive and fully automatic way. These sensors are mostly present in the smartbands and smartwatches of the best-selling brands, such as Fitbit, Xiaomi or Apple (IDC 2016a, b, c). Other important factor to choose these devices is their availability, as long as smartbands and smartwatches are the most common COTS wearables (IDC 2016d) (more than 80% of market share by 2020). Different wearables and smartphones have already been used in the educational domain in various initiatives and projects to study certain behaviors of students and identify patterns and indicators. In the Studentlife project (Wang et al. 2014, 2015; Ben-Zeev et al. 2015; Mohr et al. 2016; Harari et al. 2016, 2017), data collected from mobile devices are used to provide enhanced student models with performance related indicators, such as the Grade Point Average (GPA), or behavioral trends, such as sleep patterns. There are other examples, such as (Espinosa et al. 2015), where the impact of inertial sensors for building engagement in educational activities is analyzed; or (Xu and Zhong 2018), where a portable EEG is studied as a wearable option to analyze different states of students; or (Mastrandrea et al. 2015), where wearable collected data is used to estimate relationship patterns among students. There exist also projects involving teachers to provide information related to their activities (Prieto et al. 2016).

Currently, available COTS wrist wearables do not offer measures of stress. Just there exist some devices that provide stress-related warnings and suggest to perform tasks to reduce stress when changes in pulsations or breathing are detected (Caddy 2018). Nevertheless, these devices do not provide stress estimations, they are conceived as intelligent assistants to support healthy habits and self-emotional control. This paper is mainly focused on studying if popular COTS wrist wearables (affordable commercial wearables) can be used as tools to estimate the stress experienced by a subject at a specific time, namely instant stress. To this end, a lab experiment involving different types of stressing activities and wearables has been designed to collect some key physiological data: heart rate (HR), skin temperature

(ST), galvanic skin response (GSR) and accelerometer (ACC) data. Then, these data are analyzed using machine learning classifiers to estimate stress and measure the accuracy and F-measure. Once the feasibility of COTS wrist wearables as stress detectors is verified, it is proposed new indicators based on the variation along time of the instant stress to support the monitoring and management of stress in educational contexts. The goal of these indicators is to enable learners to enhance their self-awareness, to improve their self-management and learning. These indicators are also intended to support teachers during the development of educational activities.

The rest of this paper is divided into four parts. Section 2 includes a review of the methods described in the scientific literature to measure stress. This review has a double purpose, to identify existing instruments, methods, and the factors for stress detection and to explore the possible use of COTS wrist wearables. Section 3 introduces the lab experiment and the results that validate the use of wearables for stress detection. Then, a set of indicators with potential use in educational scenarios is proposed in section 4. Finally, the last part of the paper explains the conclusions of this study.

2. Methodologies to detect and analyze stress

This section reviews existing methods to estimate stress and stress-related features. These methods can be found mainly in the medical literature, as part of psychological studies. Two different types of methods can be distinguished: subjective tests and physiological signals analysis.

2.1. Clinical subjective tests

Several different clinical subjective tests have been proposed to provide measures of stress-related features. These tests are based on the use of questionnaires to collect subjective data from subjects. Generally, questions are about the perception of stress or about the frequency of stressful events, usually considering different scenarios and situations. All the instruments are based on the use of Likert scales to collect the subjective user answers. The most well-known subjective stress tests found in the literature are the following ones:

- **State-Trait Anxiety Inventory (STAI)** (Spielberger et al. 1970). This is the most used test worldwide (Cano et al. 2007) and it has been used as a support of multitude of studies (Grös et al. 2007). This test is made up by a total of 40 items that measure two different anxiety concepts: state anxiety and trait anxiety. Each one of the two parts is made up by 20 questions. Questions related to state anxiety are about the subject feelings at this moment (such as “I feel at ease”, “I feel upset”), while questions related to trait anxiety are about the subject feelings in his/her daily life situations (such as “I am a steady person”).
- **Maslach Burnout Inventory (MBI)** (Maslach et al. 1986b). It is made up by 22 items referred to different stress situations, including features such as professional and emotional exhaustion or depersonalization. Three versions for different application areas have been proposed: the MBI-General Survey (MBI-GS) with a generic approach (Schaufeli and Leiter 1996); the MBI-Human Services Survey (MBI-HSS) (Maslach et al. 1986b)) focused on the professionals in the human services domain; and the MBI-Educators (MBI-ES) (Maslach et al. 1986a), focused on teachers, administrators and other staff members, working in any educational setting. Each one of these focused tests involves an adaptation of the generic questions to the specificities of the domain. For example, depending on the test, question 4 is provided as follows: “I feel I can understand patients easily” and “I can easily understand the feelings of my students.”

- **Perceived Stress Scale (PSS-14)** (Cohen et al. 1983). This test was developed in 1983, but it continues in common use nowadays. In the original form, it includes 14 questions referred to potential stressful situations produced during the last month. For example: “In the last month, how often have you been upset because of something that happened unexpectedly?”. There exist two simplified versions with just 10 questions (PSS-10) and with just 4 questions (PSS-4) (Cohen and Williamson 1988).
- **Anxiety Sensitivity Index (ASI)** (Reiss et al. 1986). This test includes 16 items designed to assess the sensitivity to anxiety situations, even the fear to the anxiety feelings. Several variants have been published. One of the most well-known ones is the ASI-3 (Taylor et al. 2007), made up by 18 items with evidence of improved psychometric properties over the original ASI.
- **Hamilton scale for Anxiety (HAM-A) or Hamilton Anxiety Rating Scale (HARS)** (Hamilton 1959). This test with 14 items measures the importance of anxiety symptoms. It is used to measure both the mental anxiety (psychological distress) and the somatic anxiety (physical aches or complains related to the anxiety).
- **Generalized Anxiety Disorders** (**GAD**). This test is focused on the detection of anxiety disorders and its involvement levels. There are several proposals, such as Anxiety Screening Questionnaire (ASQ—15) made up by 15 items (Wittchen and Boyer 1998) or the Screening Scale for DSM–IV GAD (Carroll and Davidson 2000) made up by 12 items.
- **Beck Anxiety Inventory (BAI)** (Beck and Steer 1990). This test is made up by 21 items. Each item checks the presence of anxiety symptoms. For example, “Fear of losing control” or “Fear of worst happening.” Its main goal is to differentiate anxiety from depression.
- **Anxiety Situations and Responses Inventory (ISRA)** (Cano-Vindel and Miguel-Tobal 1999). This test made up by 224 items assesses the level of stress at cognitive, psychological and physical functions. The test shows the tendency to stress in four contexts: situations where the person can be assessed, social scenarios, phobic situations for the person and daily/ordinary situations.
- **Depression, Anxiety and Stress Scales (DASS)** (Lovibond and Lovibond 1995). This test is designed to measure the seriousness of a range of common symptoms in three scales: depression, anxiety and stress. It is made up by 42 questions arranged in 14 items for each scale. There also exists a simplified version to reduce the number of questions and to improve the psychometric properties (Norton 2007). Both versions have shown good psychometric properties in several studies (Brown et al. 1997).

From this review, we would like to notice the variety of contexts and ways in which stress is considered:

- Related to the contexts, tests usually involve a specific focus on a particular context or variations for different contexts. For example, the MBI focuses on professional settings, PSS is about stressful situations produced during the last month and ISRA distinguishes four areas: situations where the person can be assessed, social scenarios, phobic situations for the person and daily/ordinary situations.
- Related to the variety of ways, different stress-related features are recognized. The MBI is focused on emotional exhaustion, ASI analyses the anxiety sensitiveness in physical, cognitive and social dimensions, HAM-A analyses mental and somatic anxiety, BAI tries to differentiate anxiety from depression, DASS distinguishes between depression, anxiety and stress and finally, STAI distinguishes between state and trait anxiety.

Other idea that can be obtained from this review is that these tests provide an estimation of accumulated stress in a period of time. These tests are used when the specialist detects a health problem or condition related to stress. In this

situation, the tests are not measuring the "snapshot" stress, but the stress accumulated during a period of time and the events related to such accumulation.

2.2. Physiological signal analysis

In contrast to the previous section, addressing subjective tests based on questionnaires and focused on accumulated stress, the scientific literature also includes many studies to detect instant stress. These studies are based on the analysis of physiological signals. Stress produces some physiological changes: variations in the cortisone levels, heartbeats, sweat, skin temperature, etc. These variations can be measured through clinical tests or using body sensors. Next list summarizes commonly used physiological signals to detect stress (Table 1 indicates studies in which these signals have been analyzed):

- **Heart rate (HR) and heart rate variability (HRV)**. Heartbeats are very related to stress. Several studies have demonstrated a strong relationship between stress and heart rate. When a subject is under a stressful situation the HR frequency increases. These changes in frequency are accurately measured by the HRV, also known as the instant HR. HRV can be calculated from an electrocardiograph by detecting R-wave peaks, but this method is very costly and uncomfortable for the subject. Nevertheless, the HR can also be measured using a photoplethysmograph from light signals over the skin (Healey 2000). Variations of this measurement are less precise than the HRV ones obtained from an electrocardiograph, but it is much more comfortable and affordable.
- **Galvanic skin response (GSR), electrodermal activity (EDA) or skin conductance (SC)**. When a person is in a stressful situation, an increase in the level of sweating is automatically produced. This increase provokes a variation in the electrical resistance of the skin. In this way, the EDA can be used to estimate the state of our nervous system. Several studies have shown very good results using the GSR as a stress detection physiological signal, particularly some recent works offer success results close to 100% using GSR in combination with HR (Cano et al. 2007; Santos 2012).
- **Muscle activity (MA).** Based on the assumption that muscular activity increases with stress, some studies have been focused on the analysis of MA (Healey 2000). The registration of muscle activity is performed by an electromyograph (EMG) detecting surface voltages that occur when a muscle is contracted.
- **Blood pressure (BP) or blood volume pressure (BVP)**. Blood pressure is a physiological feature that varies for multiple reasons, such as the physical exercise, ingestion of food or stress. Using a sphygmomanometer, it is possible to measure the BP. The main drawback of this device is that it is not possible to take continuous measurements. Nevertheless, the photoplethysmograph used to measure the BP can also measure differences in blood volume through a reflected signal (infrared or red) over the skin in a controlled way.
- **Skin temperature (ST).** The temperature of the skin has been proposed as a useful estimator of the stress level used in multimodal systems (Karthikeyan et al. 2012). Its use improves the accuracy of classifiers, providing figures similar to the BP ones, around 88.75%.
- **Respiration (RESP)**. Variations in the breath speed, deep breathes and irregular breathes are indicators of stressful situations. As it is described in (Healey 2000), a hall effect sensor can be used for measuring respiration through chest cavity expansion.

- **Pupil diameter (PD).** The size of pupil varies under stressful situations. Using methods such as video-pupillography, studies have been performed to analyze how stress affects to the PD (Pedrotti et al. 2014). Mistaken and missed measurements due to automatic flickering are discarded or interpolated.
- **Salivary level (SL).** Variations in salivary alpha-amylase and salivary corrosion are related to the body response under stressful situations. In (Rashkova et al. 2012), SL is identified as a good objective biomarker to detect stress when used in combination with a psychological test like STAI.

Table 1. Scientific publications relating stress and physiological signals.

Bibliographic Ref.	**BP/BVP**	**RESP**	**PD**	**SL**	**MA**	**ST**	**HR**	**HRV**	**GSR**
(Lundberg et al. 1994)					X				
(Healey 2000)	X	X			X		X	X	X
(Vrijkotte et al. 2000)	X						X	X	
(Dishman et al. 2000)								X	
(Healey and Picard 2005)		X			X		X	X	X
(Zhai et al. 2005)	X		X						X
(Lin et al. 2005)	X						X		X
(Zhai and Barreto 2006)	X		X			X			X
(Setz et al. 2010)									X
(Mokhayeri et al. 2011)			X				X	X	
(Hernandez et al. 2011)									X
(Rashkova et al. 2012)				X					
(Karthikeyan et al. 2012)						X			
(Santos 2012)							X		X
(Pedrotti et al. 2014)			X						X
(Sano and Eng 2016)									X
(Kothgassner et al. 2016)				X			X	X	

As a summary, there exists a variety of physiological signals that can be used to estimate the stress. As it can be observed in Table 1, GSR and HR/HRV are the most common ones as they are considered in 8 of the 17 reviewed papers. These studies are focused on validating stress detection using the physiological signals. In 8 of these studies (Zhai et al. 2005; Zhai and Barreto 2006; Setz et al. 2010; Hernandez et al. 2011; Mokhayeri et al. 2011; Santos 2012; Pedrotti et al. 2014; Sano and Eng 2016) a machine learning approach is used to validate the stress detection and the results obtained are used to show the performance of the classifiers as predictors. These machine learning techniques facilitate the creation of adaptive algorithms to detect stress automatically.

These papers also include other interesting issues: methods to induce stress in subjects during experiments and method to analyze the data collected. Next, a summary of the devices and method used to stress detection is shown (cf. Table 2):

- In (Zhai et al. 2005) a clinical NI DAQPad-6020E device was used to collect physiological signals. This is a multi-channel data acquisition system produced by National Instrumentation Corp. It was used in combination with the “Stroop Color-Word Test Interference Test” (Stroop 1935), a common stressor activity, in which

words referred to colors (e.g. “blue”) are shown colored in the same or in a different color to the word (e.g. the word “blue” can be painted in yellow color).

- In (Zhai and Barreto 2006) there is no information about the equipment used to collect stress-related data. As in the previous case, the “Stroop Color-Word Interference Test” was used as stressor activity. The whole experiment comprises four consecutive sections: an Introductory segment (IS); a Congruent segment (C); an Incongruent segment (IC); and a Resting Section (RS). During the IC segment, the idea is to confuse the user, indicating wrong colors.
- In (Setz et al. 2010) the Emotion Board device was used. This is a prototype for measuring EDA in a non-intrusive way. To provoke stress, subjects were asked to solve some mathematical equations. A time bar showed the remaining time for solving the task, and a color bar showed a comparison between the individual performance and the performance of a simulated, representative population. The test also involved a kind of evaluator and a supervisor that tried to induce stress in the subject providing low marks or entering into the experimentation room to ask damning questions.
- In (Mokhayeri et al. 2011) a video camera was used to collect the pupil diameter. This device records the subject and detects variations in the stress level induced using the “Stroop Color-Word Interference Test”.
- In (Hernandez et al. 2011) data were recorded from dry Ag-AgCl 1cm diameter electrodes on the wrist, using an early beta version of the discontinued Affectiva QSensor device. The stressor activity was developed in the context of a call center, where participants were asked to respond calls and rate each call-in terms of stress. This study provides one of the lower values of accuracy for stress detection: 78.03% using SVM.
- In (Santos 2012) a multichannel research device I-330-C2 PHYSIOLAB (J &J Engineering) was used to collect the measurements. The stressor activity was the Hyperventilation and Talk Preparation activity. In this activity, first subjects (in this case students) are required to perform a hyperventilation activity that produces variations in physiological signals like a stressor activity. Then, subjects are asked to prepare a presentation that will be recorded. The results from both activities are compared. This study provides one of the best results in stress detection, with an accuracy of 99.5% using a Fuzzy Logic classifier.
- In (Pedrotti et al. 2014) it was used an eye tracker device (RED 4) for PD measurement; an Analog /Digital(A/D) converter (MP36R acquisition system for science researches) for measuring illumination and EDA. To generate a stressor situation, it was used a driving simulator, which consists of driving on a traffic-free straight three-lane road, changing lanes according to the information displayed.
- In (Sano and Eng 2016) an ad-hoc device was used, but there is no information about it. Stress analysis was performed through the quantification of activities along several days, not provoking stress with a specific stressor activity.

Table 2. Summary of machine learning scientific publications.

Bibliographic Ref.	**Device**	**Physiological signal**	**Machine Learning technique**	**Stress activity**	**Accuracy**
(Zhai et al. 2005)	NI DAQPad-6020E	BP/BVP, PD, GSR	SVM	Stroop Color-Word Interference Test	80%
(Zhai and Barreto 2006)	-	BP/BVP, PD, ST, GSR	SVM, Naïve Bayes, Decision Tree	Stroop Color-Word Interference Test	90%

(Setz et al. 2010)	Emotion Board	GSR	SVM, LDA, NCC	Mathematical equations	81.30%
(Mokhayeri et al. 2011)	Video camera	PD, HR, HRV	GA, FSVM	Stroop Color-Word Interference Test	78.5%
(Hernandez et al. 2011)	Discontinued Affectiva QSensor device	GSR	SVM	Call center	78.03%
(Santos 2012)	I-330-C2 PHYSIOLAB	HR, GSR	SVM, LDA, MM, K-nn, Fuzzy Logic	Hyperventilation and Talk Preparation	99.5%
(Pedrotti et al. 2014)	MP36R acquisition system	PD, GSR	Neuronal Network	Driving simulator	79.20%
(Sano and Eng 2016)	-	GSR	SVM	Several activities during sleep and daily life	82.4%

As a summary, the SVM (Support Vector Machine) classifier has been the most used technique. Only 2 out of the 8 studies did not use this technique. Another interesting feature is the accuracy of the models. In general, the accuracy is between 78% and 99.5%, with 2 studies obtaining a value over 90%. The main differences among results are related to the classifier selected and to the number of the physiological signals used. The use of a combination of signals has provided good results (Santos 2012). In another way, all the papers except one (Pedrotti et al. 2014), just estimate two levels of stress: relax, low stress or non-stress versus stress. Another important point is the variety of measurement devices and the activities to provoke stress. The "Stroop Color-Word Interference Test" has been used in several studies and can be taken as a reference. This activity, or other similar ones, have been used in combination with an audio stimulus (noise), social events (people entering or leaving the room) with the purpose of creating an atmosphere that can cause stress to the subjects.

3. Feasibility analysis of stress detection

From the results described in section 2, it is clear that the measurement of physiological signals and the use of machine learning methods can be used to provide a reasonable estimation of stress. Nevertheless, such signals have been collected using specialized devices with precise sensors. Our purpose is to validate the use of COTS wrist wearables and their sensors as collectors of such physiological signals to estimate stress.

In the current market, there are COTS wrist wearables that have sensors that measure physiological signals with a precision similar to specialized devices. Some examples are:E4 wristband (empatica 2016) and Shimmer3 (Shimmer). Nevertheless, these devices have been designed to be used for research purposes (Burns et al. 2010; Koskimäki et al. 2017). The main problems with these devices are their high cost (over € 800) or the hassle caused to the users (small probes and wiring to measure the GSR in the fingers).

For these reasons, our research has focused on the popular COTS wrist wearables. This type of device may involve concerns about their performance. In this regard, there are publications in the literature that try to quantify the accuracy

of COTS wrist wearables (Guo et al. 2013; Stahl et al. 2016; Wallen et al. 2016; Wang et al. 2017). The results reveal that the accuracy depends on the level of movement of the subject (Stahl et al. 2016). However, in conditions of low movement, sensors such as the HR offer values of absolute error in a range of 2,8% - 5,41% (Stahl et al. 2016). We can say that, although COTS wrist wearables have a lower accuracy than clinical devices, the use of these devices allows studies to avoid the annoying wiring of clinical tests and thus eliminating the psychological biases that would result from their use.

3.1. Methods

From the available devices in the market, we selected the Microsoft Band 2 wrist wearable. The selection was mainly based on the included sensors: optical HR monitor, accelerometer/gyrometer, thermometer, barometer, GSR monitor, light and UV meters, microphone and GPS. Another selection criterial were the options available to collect and transfer sensor data to our analytics system. This is not a simple task, because several issues can be involved (de Arriba-Pérez et al. 2016): interoperability among systems, energy consumption, storage capacity, etc. In the Microsoft Band 2 case, data can be collected by an Android/Windows/iOS app using available libraries. Therefore, from the most relevant physiological signals described in section 2.2, we could collect the next ones: HR, ST, GSR and motion/acceleration. Table 3 shows the technical spec of this device. This information was obtained from the official web page of Microsoft (Microsoft 2015).

Table 3. Microsoft Band 2 sensors information.

Sensor	Details	Frequency
Accelerometer	Provides X, Y, and Z acceleration in g units. 1 g = 9.81 meters per second squared (m/s2).	62/31/8 Hz
Gyroscope	Provides X, Y, and Z angular velocity in degrees per second (°/sec) units.	62/31/8 Hz
Heart Rate	Provides the number of beats per minute.	1 Hz
Skin temperature	Provides the current skin temperature of the wearer in degrees Celsius.	1 Hz
UV	Provides the current ultraviolet radiation exposure intensity.	1 Hz
Galvanic Skin Response	Provides the current skin resistance of the wearer in kohms.	0.2/5 Hz
Ambient Light	Provides the current light intensity (illuminance) in lux (Lumes per sq. meter).	2 Hz
Barometer	Provides the current raw air pressure in hPa (hectopascals) and raw temperature in degrees Celsius.	1 Hz
Altimeter	Provides current elevation data like total gain/loss, steps ascended/descended, flights ascended/descended, and elevation rate.	1 Hz

The experiment involves the induction of different stress levels to the subjects. The protocol to perform this experiment has been described in a previous publication (de Arriba-Pérez et al. 2018) and here we focus on the description of the

stressing activities and the data analysis. To support the performance of stressing activities an Android app was developed for both smartphone and tablet. This app proposes 4 different activities to the subject under experimentation. Each one of these activities is explained to the subject previous to the performance of the experiment, including also the description of the measured variables and the purpose of the measurement. Besides, previous to the performance of each activity, the app provides a brief textual description. The 4 activities are as follows:

- The first activity (VIDEO) is a 4 minutes 20 seconds relaxing video. The aim is to take the subject to an entire relaxing situation (cf. Fig. 1).

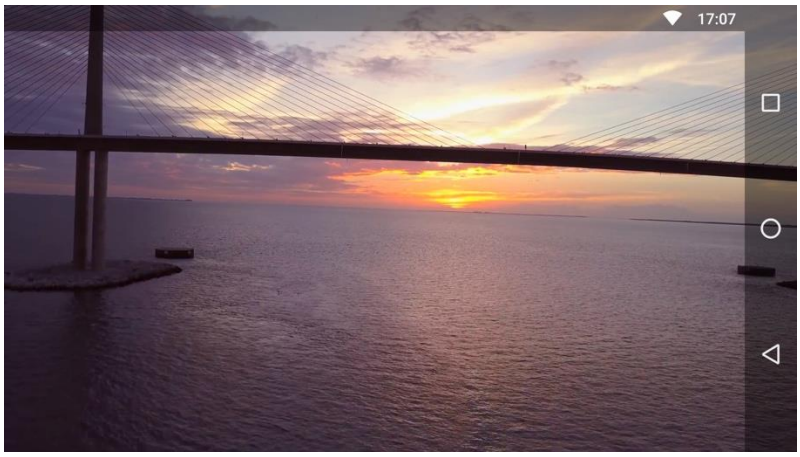

Fig. 1. Snapshot of the video activity.

- The second activity (SCT) involves an adapted version of the "Stroop Color-Word Interference Test". As it is indicated in section 2.2, this has been extensively used in the experiments about stress. In all of them, colored words appear in the screen and then disappear when the subject clicks in the button corresponding to the color shown with a time limit of 2 seconds (cf. Fig. 2), that is the maximum time available for answering. This activity is composed of 3 different levels of difficulty:
    - In the first level (SCT1), the colors and the words are in correspondence. For example, if the word is "Green", then it is in green color. Besides, the words/colors are shown following the same sequence. The aim of this level is to train the user and to take into account the accelerometer movements in the absence of stress.
    - In the second level (SCT2), the colors and the words are in correspondence also, but the sequence of appearance is randomized. Also, the user will have just 2 attempts to get over this level. Every time the subject makes a mistake, a buzz is emitted. The aim of this level is to check the existence of a variation in the physiological signals as an indication of a level of stress produced by the concentration needed to perform the task.
    - In the third level (SCT3), the colors and the words are not in correspondence. The rest of the conditions are equal to the previous level. The aim is to take the subject stress to the highest level.

Fig. 2. Snapshot of the Stroop Color-Word Interference Test (SCT) activities.

- The third activity (PASAT) involves an adaptation of the PASAT test (Tombaugh 2006). The aim of this test is to increase attention and concentration, as demonstrated by several studies (Tombaugh 2006). In this test, the subject hears numbers in a sequence, and he/she has to answer the result of adding the actual number with the previous one (cf. Fig. 3). For example, if the sequence is initiated with the number "1" and then the number "2", the subject should write "3"; if the next number is "5", he/she should answer "7", and so on. This test has been included to compare the results with the ones obtained in the "Stroop Color-Word Interference Test", trying to achieve a high concentration/attention state different from high stressed and relaxed ones.

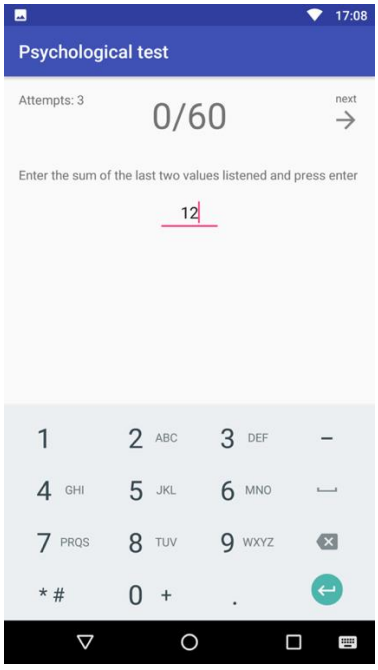

Fig. 3. Snapshot of the PASAT activity.

- Finally, the fourth and last activity (HYPERVENT) involves the development of a hyperventilation exercise (Santos 2012). In this exercise, the subject has to breathe deeply, inhaling and exhaling, following the rhythm

marked by the app (cf. Fig. 4). The duration of the breathing periods is of 3 seconds. If physiological signals do not change significantly, the subject is asked to increase inspiration and expiration rates. The aim of this test is to assess the variations of the physiological signals when the user is in a hyperventilation state.

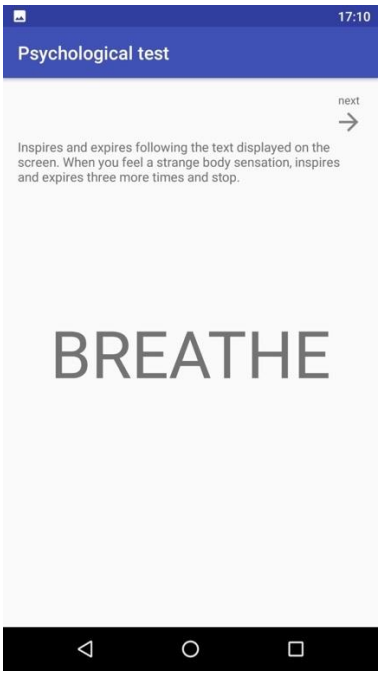

Fig. 4. Snapshot of the hyperventilation activity.

Each one of these activities is followed by a question about the stress perceived (cf. Fig. 5). The subject provides a value (using a 7-point Likert scale) of the perceived stress, indicating if he/she has felt relaxed, stressed, or has experimented any intermediate level of stress.

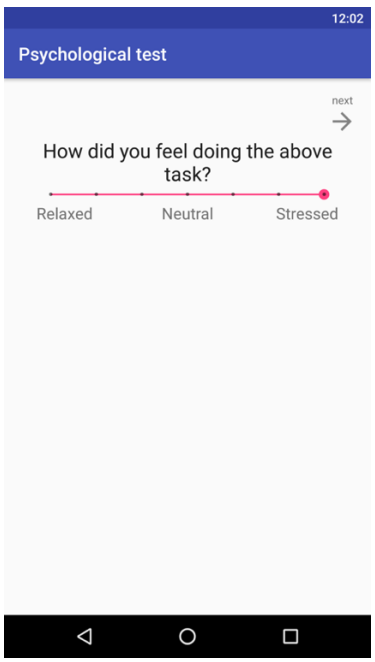

Fig. 5. Quiz to be completed at the end of each activity.

The physiological signals are collected by another app developed for this purpose, and the data is sent to an analytic server. The data can be shown on a web dashboard. Each second a sample of HR, Acc, GSR and ST is recorded. To eliminate the noise of HR samples, a FIR filter has been introduced. This filter is commonly used in the real-time signal analysis (Fan and Wang 2010). Several filtering levels were tested with different time window values. Finally, a 15 seconds time window was selected because the high-frequency component is reduced and the temporal shift and amplitude loss are almost insignificant.

For data processing, we have used a Java Server with a REST API with Jersey (Jersey 2016) including Weka as the machine learning library (Mark et al. 2011). Data is stored in a MongoDB server (MongoDB 2017), as objects in BSON (Binary JSON) format. The data structure includes an id for each user and timestamp and a record with the value of each sensor at sampling times. Finally, the data and results are rendered in the web using the Highcharts library (Highcharts 2017) that provides an excellent set of tools to show dynamic graphics (cf. Fig. 6).

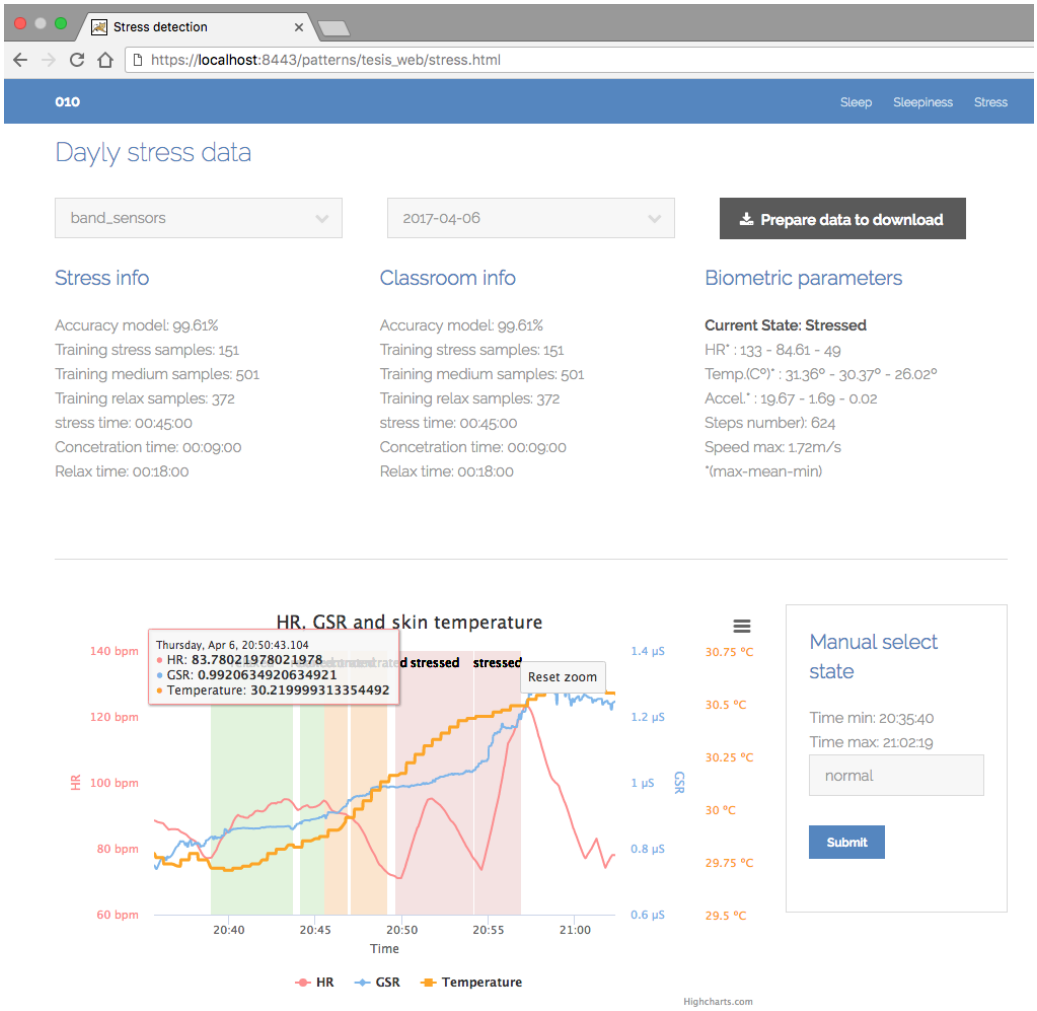

Fig. 6. Dashboard to visualize the physiological signals collected from the wearable.

3.2. Data Analysis

The experiment involved 19 subjects with an average age of 23.5 ± 6.04. It was divided into two different stages, but the same stress induction procedures were carried out in both of them. In the first stage, a total of 7 researchers and

postgraduate students participated. In the second stage, 12 undergraduate students from the University of Vigo were involved. The experiment was performance inside the range of 18ºC to 24ºC recommended for the performance of intellectual activities. The people participating in the experiment were informed and provided their consent. All of them were aware of the voluntary character of the test and had a good knowledge about the aim of the study and its development. Nevertheless, they do not have previous experiences in the performance of stress-related tests like this one. In addition, no one of the subjects had any cardiopathy or known heart disorders. In the following sections are summarized the visual and numerical analysis performed on the generated datasets.

3.2.1. Visual Analysis

Firstly, we analyzed the physiological signals variations in a visual way. Fig. 7 shows results of the 7 subjects involved in the first stage. The most remarkable issues in the evolution of the physiological signals are the following ones:

- The HR experiences an evident increase in the hyperventilation activity. At the beginning of this activity, the HR value is in a medium or low level, compared to the previous activities. Only the subject number 3 shows a very high HR value at the beginning of the activity. This behavior is anomalous in comparison to the other subjects. We asked the subject about this issue, and he explained that he was particularly nervous and worried about the next activity.
- The GSR increases in a staggered way with the complexity of the activities.
- The temperature shows several variations among subjects. Subjects 1, 3, 4 and 5 increase the temperature in the first SCT activities and this increase continues until the most demanding activities are experienced, such as the SCT3 and PASAT. From here, and mainly during the hyperventilation activity, the temperature decreases. Subject number 7 does not show this trend in the decrease of temperature during the hyperventilation, but the increase of temperature slows down. Finally, subject 2 shows a strange behavior, because the temperature decreases from the beginning, with a huge decrease at the beginning of the hyperventilation activity.

The answer provided by the subjects to the question about the perceived stress is shown in the top part of each graph. When we analyzed the physiological signals of each subject, we observed that the perceived stress is highly subjective. For example, in the fifth graph, it is shown a subject that qualifies the video activity as medium stress. In the other activities, the perceived stress provided by the subject changes in correspondence with the physiological signals. Nevertheless, in the case of the hyperventilation activity, despite the subject does not consider it as a stressful situation it produces a high physiological activation, similar to a situation of high stress (Cano et al. 2007). In another way, the sixth graphic shows a subject that has qualified all the activities with the same stress level, but the variations of the physiological signals indicate different stress levels. As it can be observed, the answers provided by the subjects are diverse, but in general, the perceived stress is correlated with variations in the physiological signals. Nevertheless, the last activity (hyperventilation) generate many doubts to almost all the subjects, maybe because it is not a real psychological stressing activity.

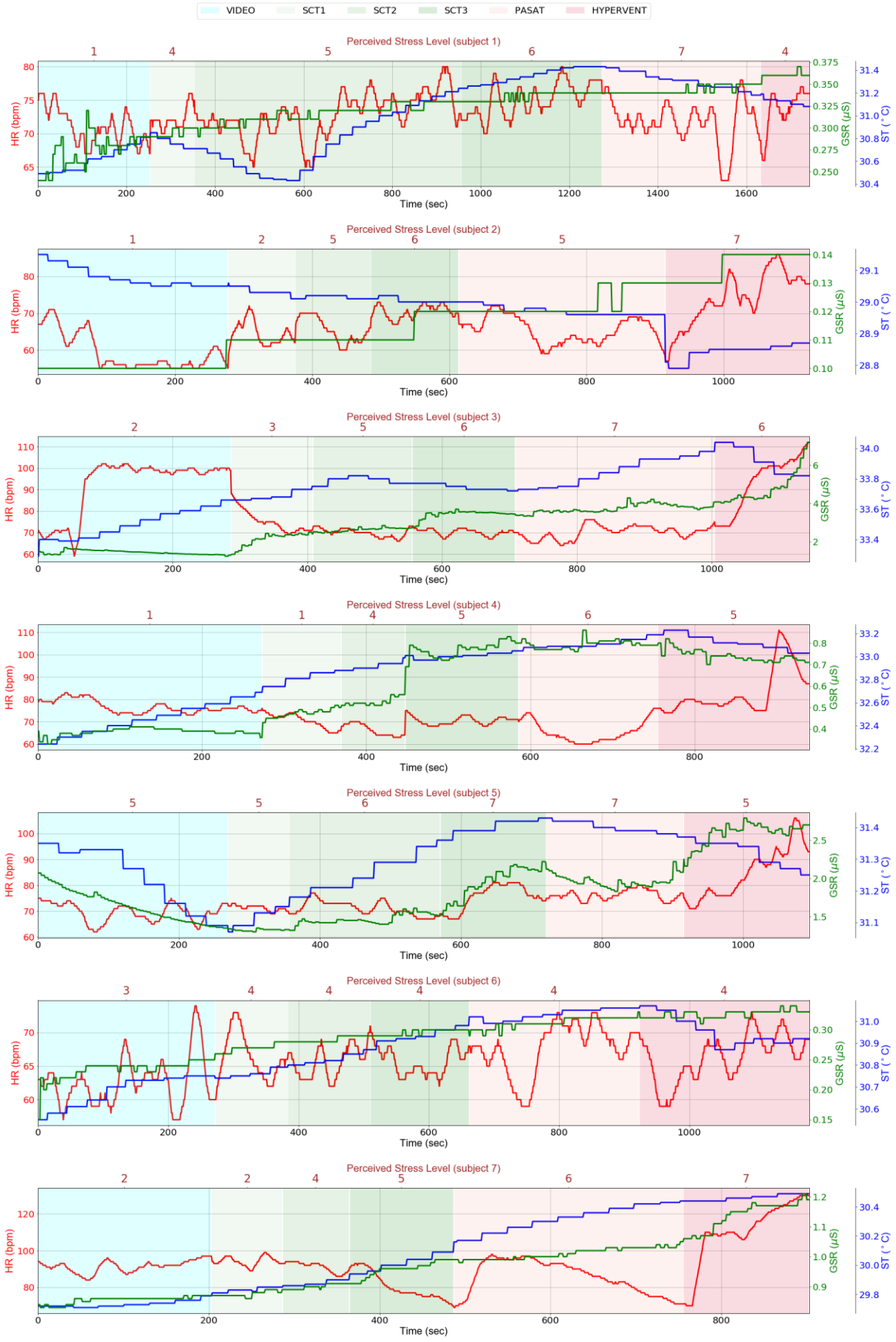

Fig. 7. Physiological signals variations by activity.

3.2.2. Numerical analysis: clustering

Secondly, we analyze the feasibility to estimate stress using machine learning techniques. We begin this process with a clustering analysis of each dataset. The goal is to identify the most similar activities and differentiated groups that share similar values of physiological signals.

To prepare the datasets we have used derived variables in addition to the raw data. Specifically, we have used the following raw variables: HR, ST, Acc and GSR. In addition, for each one of these variables we have used the standard deviation (st), the slope (sl) and the difference between the most remote value and the current one (diff) in the 15 and 30 seconds of previous samples. Moreover, we have only used samples from the last 3 minutes of the activities to avoid non-representative samples of the initial period in which the user is just assuming the difficulty of the task and the physiological signals can vary randomly. Finally, we got 19 datasets with a total of 12200 samples (in mean, 642 samples per subject).

The obtained dataset files are processed by the Expectation Maximisation (EM) clustering algorithm to search 3 clusters for each subject. The clusterization results are shown in Table 4. In this table, it can be observed that activities SCT1 and SCT2 are classified in most cases with the video activity and in other cases with more stressful activities, such as the SCT3 and PASAT. Similarly, the hyperventilation activity shares some samples with stressing activities, such as SCT3 and PASAT (Cluster 2) or it is classified separately (Cluster 3). This can be explained by the abrupt variations in the physiological signals that can be observed. Finally, the relaxing video provides a differentiated cluster (Cluster 1). Only in two cases, this activity introduced samples in a stress cluster. Nevertheless, this event has probably been the result of an initial sense of stress or the unusual physiological behavior of a subject. As a result, we conclude that the relaxing activity and the activities to produce stress provoke differentiated physiological signals. Therefore, we group the samples in three main categories: (i) video tagged as “Relax”; (ii) SCT3 and PASAT tagged as “Medium stress”; and (iii) hyperventilation tagged as “Stress”. Samples of the SCT1 and SCT2 are discarded because they belong to several groups simultaneously. This is coherent if we take into account that both activities serve as a transition between a relaxing activity (VIDEO) and a stressful activity (SCT3).

Table 4. Classification of the individual subject activities into clusters detected.

UserID	Cluster 1	Cluster 2	Cluster 3
1	VIDEO, SCT2	PASAT, SCT3, HYPERVENT	VIDEO, SCT1, SCT3, PASAT, HYPERVENT
2	VIDEO, SCT1	SCT2, SCT3, PASAT	HYPERVENT
3	VIDEO	SCT1, SCT2, SCT3, PASAT	SCT1, HYPERVENT
4	VIDEO, SCT1, SCT2	SCT3, PASAT, HYPERVENT	HYPERVENT
5	VIDEO, HYPERVENT	SCT1, SCT2	SCT3, PASAT, HYPERVENT
6	VIDEO, SCT1, SCT2	SCT2, SCT3, HYPERVENT	PASAT, SCT3, HYPERVENT
7	VIDEO, SCT1, SCT2	SCT2, SCT3, PASAT	HYPERVENT
8	VIDEO, SCT1, SCT2	SCT3, PASAT	VIDEO, SCT1, SCT2, PASAT, HYPERVENT
9	VIDEO, SCT1, SCT2, SCT3	SCT1, SCT3, PASAT, HYPERVENT	HYPERVENT
10	VIDEO, SCT1, SCT2	SCT2, SCT3, PASAT, HYPERVENT	PASAT, HYPERVENT

11	VIDEO, SCT1, SCT2, SCT3	SCT3, PASAT	PASAT, HYPERVENT
12	VIDEO, SCT1	SCT2, SCT3, PASAT	VIDEO, SCT1, SCT2, SCT3, HYPERVENT
13	VIDEO	VIDEO, SCT1, SCT2, SCT3	PASAT, HYPERVENT
14	VIDEO	SCT1, SCT2, SCT3	SCT3, PASAT, HYPERVENT
15	VIDEO, SCT1	VIDEO, SCT1, SCT2, SCT3, PASAT	HYPERVENT
16	VIDEO, SCT1, SCT2	SCT3, PASAT	VIDEO, SCT1, SCT2, HYPERVENT
17	VIDEO, SCT1, SCT2	SCT1, SCT2, SCT3, PASAT	HYPERVENT
18	VIDEO, SCT1, SCT2	SCT1, SCT2, SCT3, PASAT	SCT1, HYPERVENT
19	VIDEO, SCT1	SCT1, SCT2, SCT3	SCT3, PASAT, HYPERVENT

Our next step was to analyze the features that provide a gain regarding prediction, using the tagged samples. The goal is to select the features that offer more accurate using machine learning classifiers. We used as attribute evaluator an "InfoGainAttributeEval" and as search method a "Ranker" provided by Weka. Table 5 shows the results of this analysis. Columns tagged as Var show the variables: prefix st stands for standard deviation, prefix sl for slope and prefix diff for the difference between the most remote value and the current one in the 15 and 30 seconds of previous samples. In addition, subindexes 15 and 30 are used to differentiate both time windows. For the superficial skin temperature, the Temp abbreviation is used. The results, shown in Table 5, indicate that the most significant features are the raw values and the statistics obtained from a 30 seconds window. The 15 seconds window also offers a significant gain. For this reason, all of them are maintained for further analysis.

Table 5. Mean gain value per attribute.

Var	Gain Ranking	Var	Gain Ranking
Temp	1.0785	$diffTemp_{15}$	0.0805
GSR	0.8124	$stTemp_{30}$	0.0721
$stAcc_{30}$	0.3819	$stGsr_{30}$	0.0594
$stAcc_{15}$	0.3683	$diffHR_{30}$	0.056
Acc	0.1816	$stHR_{15}$	0.0546
$diffTemp_{30}$	0.1723	$slGsr_{30}$	0.0454
HR_{FIR}	0.1672	$slAcc_{30}$	0.0449
$diffAcc_{15}$	0.1343	$stGsr_{15}$	0.0384
$diffAcc_{30}$	0.1229	$slGsr_{15}$	0.0358
$diffGsr_{30}$	0.1128	$stTemp_{15}$	0.0319
$slAcc_{15}$	0.0884	$slHR_{15}$	0.0278
$slHR_{30}$	0.081	$diffHR_{15}$	0.0257
$stHR_{30}$	0.081	$slTemp_{15}$	0.0189
$diffGsr_{15}$	0.0807	$slTemp_{30}$	0.0119

3.2.3. Numerical analysis: Individual data set study

Using the obtained datasets, we analyzed the behavior of several machine learning algorithms to generate individual predictors for stress. Several predictors were evaluated to select the one that performs better following a cross-validation check. In this cross-validation process, the input dataset was divided into 10 parts, using 9 to train and one to test. In the next step, the testing part becomes a part of training, and a part of training becomes testing. This process was performed 10 times moving the training parts. Finally, the mean value of all the obtained results is used to calculate the accuracy and errors.

The results are shown in Table 6. In this table, it can be observed the accuracy for each stress category (Relax, Medium stress, Stress) and the mean accuracy and mean F-measure value of all individual subject's analyses. All the classification algorithms evaluated in this table have been launched with the default parameters defined in the Weka analysis tool. As it can be observed, results are excellent. The accuracy of all the classifiers is over 0.9 (the Zero R is used as the worst result possible). The algorithms offering the best performance and results are the Neuronal Network (a Multilayer Perceptron with 15 hidden layers), Random Forest, C4.5 and IBK. Random Forest and C4.5. These algorithms create decision trees in accordance with the values of each one of the attributes used, and in this way, they can show overfitting problems if the data elements are very similar. For this reason, the algorithms that take into account the similarity among the samples using clusters and neighbor's similitude such as SVM and IBK, or algorithms that try to follow the variations of the signals, such as, the Locally weighted learning can provide most strength. Following a different approach, the use of Neuronal Network also returns very good results. In this case, the drawback is that the computation is very expensive. A main conclusion from Table 6 is that the high degree of the accuracy of several algorithms confirm the existence of a significant physiological difference among the different stress levels.

Table 6. Metric scores of individual subject values with several classifier.

	Accuracy Relax	**Accuracy Medium**	**Accuracy Stress**	**Accuracy Global**	**F-measure**
Neuronal Network	1	0.999	1	1	0.999
SVM	1	0.994	0.994	0.996	0.992
Locally weighted learning	0.991	0.933	0.924	0.95	0.882
C4.5	1	0.998	0.998	0.999	0.998
Random Forest	1	1	1	1	1
IBK	1	0.999	0.999	0.999	0.998
Naive Bayes	0.995	0.959	0.96	0.972	0.953
Zero R	0.725	0.486	0.762	0.658	0.218

3.2.4. Numerical analysis: Global classifier analysis

Beyond the individual per-subject prediction models, we explored a general classifier using all the available samples. The goal is to provide a kind of cold start system that is able to detect stress and relax of a subject without any previous training. To analyze the behavior of this global classifier we carry out different tests that are explained below.

As a first approach, we evaluated the behavior of each classifier applying the same validation system than in the individual analyses case. To do it, we included all the samples in the same single file and applied the cross validation method over the same set of classification algorithms previously shown. As it can be observer in Table 7, the values obtained for Neuronal Network, Random Forest, C4.5 and IBK are very similar to the best results available in literature using specialized devices (Zhai and Barreto 2006; Santos 2012). In this case, the SVM suffers in comparison to the individual solution, but the prediction accuracy is reasonable. Again, Random Forest, C4.5 and IBK offer the best result with a value greater than 0.97. This result indicates that a general model involving all the subjects can be developed, but it does not confirm that new subjects can be assessed with a high success rate.

Table 7. Scores of all user in the same file with several classifiers.

	Accuracy Relax	**Accuracy Medium**	**Accuracy Stress**	**Accuracy Global**	**F-measure**
Neuronal Network	0.929	0.894	0.916	0.913	0.857
SVM	0.776	0.696	0.786	0.753	0.552
Locally weighted learning	0.766	0.75	0.76	0.759	0.464
C4.5	0.986	0.981	0.983	0.983	0.973
Random Forest	0.998	0.994	0.995	0.996	0.993
IBK	0.997	0.995	0.995	0.996	0.993
Naive Bayes	0.58	0.621	0.761	0.654	0.458
Zero R	0.725	0.486	0.762	0.658	0.218

Our next goal was to analyze the behavior of the system in a more real scenario. To do it, a file was generated including the samples of 18 subjects to train the classifiers. Another file was generated containing the samples of the remaining subject, to be used as a test. This procedure is performed 19 times, once for each subject. In this case, as it can be seen in Table 8, the results obtained are much worse. The main difference is obtained with the Neural Network, C4.5 and IBK, providing accuracy values around 0.6. This is not the case of the Locally weighted learning and Random Forest, that continue to offer a reasonable percentage of detection, greater than 0.7. However, in both classifiers the value of F-measure is reduced dramatically.

Table 8. Scores with all samples of subjects except one in a file as training and the remaining one samples in a different file as validation.

	Accuracy Relax	**Accuracy Medium**	**Accuracy Stress**	**Accuracy Global**	**F-measure**
Neuronal Network	0.64	0.569	0.611	0.606	0.353
SVM	0.726	0.614	0.728	0.689	0.434
Locally weighted learning	0.761	0.749	0.762	0.757	0.454
C4.5	0.661	0.631	0.607	0.633	0.398

Random Forest	0.727	0.718	0.688	0.711	0.478
IBK	0.638	0.561	0.624	0.608	0.368
Naive Bayes	0.569	0.609	0.75	0.643	0.406
Zero R	0.725	0.486	0.762	0.658	0.218

Taking into account the results shown in Table 8, a new analysis was carried out. This time working with just two stress levels and discarding the hyperventilation samples. Table 9 shows the results obtained. A slight improvement in the accuracy level of the Random Forest and Locally weighted learning classifiers can be observed, with F-measure values above 0.7. These results are acceptable and they are very close to the ones shown in the literature using specialized devices (Hernandez et al. 2011; Mokhayeri et al. 2011).

Analyzing separately the 19 iterations presented in Table 10, we noticed the existence of subjects whose values of accuracy and F-measure presented were very estrange. These subjects (marked in bold in Table 10) have certain peculiarities in their physiological signals, which include extremely high GSR values, physiological signals almost unchanged during the performance of the tests, skin surface temperature with sharp drops and an HR excessively high during video viewing. Table 11 shows the results when the samples of these subjects were omitted. In this case, the classifiers SVM, C4.5, Random Forest and Locally weighted learning improved their values of accuracy and F-measure significantly. This improvement was especially important for the last two classifiers, which place both metrics in a range of 0.7 to 0.85.

Table 9. Scores with all samples of the subjects except one in a file as training and the remaining one samples in a different file as validation. Two-stage analysis.

	Accuracy	**F-measure**
Neuronal Network	0.574	0.523
SVM	0.67	0.631
Locally weighted learning	0.785	0.76
C4.5	0.678	0.648
Random Forest	0.724	0.70
IBK	0.567	0.538
Naive Bayes	0.623	0.554
Zero R	0.495	0.331

Table 10. Accuracy and F-measures of each one of the experiment iterations using a single file with samples of all subjects except one as training, and a different file with the remaining one samples as validation.

Interaction	**Accuracy**	**F-measure**
1	**0.555**	**0.443**
2	0.918	0.918
3	0.728	0.701

4	0.654	0.574
5	0.828	0.824
6	0.922	0.922
7	**0.549**	**0.522**
8	0.734	0.713
9	**0.576**	**0.483**
10	0.817	0.815
11	0.972	0.972
12	0.8	0.792
13	**0.528**	**0.432**
14	0.879	0.878
15	0.845	0.841
16	0.972	0.972
17	0.857	0.857
18	0.972	0.972
19	0.813	0.812

Table 11. Scores using a single file with samples of all subjects except one as training, and a different file with the samples of the remaining one as validation, omitting atypical subjects

	Accuracy	**F-measure**
Neuronal Network	0.587	0.535
SVM	0.7	0.671
Locally weighted learning	0.847	0.838
C4.5	0.71	0.688
Random Forest	0.765	0.758
IBK	0.558	0.532
Naive Bayes	0.642	0.575
Zero R	0.496	0.332

Finally, we wanted to see how this stress estimation system behaves in a realistic scenario. The Locally Weighted Learning (LWL) classifier was trained with the initial 19 subjects and evaluated with 5 new subjects. The results, shown in Table 12, offered values of accuracy and F-measure higher than 0.8 and mostly close to 0.9 in all cases. In addition, we wanted to analyze which variables are the most relevant for the two classifiers with the best results (LWL and Random Forest). We used Weka's ClassifierAtrributeEval algorithm to obtain the data shown in Table 13. As it can be observed, the GSR is the most representative signal of stress, as it appears in the top positions for both classifiers. This conclusion is aligned with works such as (Setz et al., 2010; Hernandez et al., 2011; Sano and Eng, 2016), where stress is estimated using only this variable.

Table 12. Metric scores of LWL in a real case study.

	Accuracy	**F-measure**
1	0.944	0.944
2	0.894	0.894
3	0.958	0.958
4	0.81	0.803
5	0.892	0.892

Table 13. Rankings of most important attributes for the LWL and Random Forest classifiers.

Rank	**Random Forest**	**Locally weighted learning**
1	Temp	$stAcc_{15}$
2	GSR	$stAcc_{30}$
3	$stAcc_{15}$	Acc
4	$stAcc_{30}$	GSR
5	Acc	$diffTemp_{30}$
6	$diffAcc_{15}$	$diffAcc_{30}$
7	$diffAcc_{30}$	$diffAcc_{15}$
8	$slAcc_{15}$	$slAcc_{15}$
9	$diffTemp_{30}$	$stHR_{15}$
10	HR_{FIR}	$stTemp_{30}$
11	$stTemp_{30}$	Temp
12	$slAcc_{30}$	$stGsr_{30}$
13	$diffGsr_{30}$	$stGsr_{15}$
14	$diffGsr_{15}$	HR_{FIR}

In conclusion, the obtained results provide a validation of the capability of COTS wrist wearables to estimate stress levels. In any case, such good results are possible because of the control of the experimental conditions. The models offer slightly worse results than the ones reported in the literature, although well aligned with them.

4. Stress related indicators for educational purposes

In the previous section, it is shown how COTS wrist wearable can be used to provide an estimation of stress level from the values of physiological variables. This can be considered as an indicator of the stress experienced at a particular time. Nevertheless, other indicators based on accumulated stress can be conceived to manage stress mainly for educational purposes. From the review of subjective tests included in section 2.1, it is clear the variety of testing methods to provide different stress indicators, considering different situations and stress-related features. Nevertheless, at this initial stage in our piece of research, we consider that the automatic identification of such conditions is not feasible in a wearable-based approach. Therefore, we propose some specific indicators taking into account the accumulation or variations of the stress level in the long time. This type of information can be especially useful in educational settings, where the burnout syndrome is present. In addition, prolonged stress, known by the name of chronic stress, has been shown to affect the health conditions and can be the source of pathologies (Dallman et al. 2003; Chandola et al. 2006; Mariotti 2015; Mayo Clinic Staff 2016) such as: heart problems, obesity, etc. The measurement of

accumulated stress does not pretend to be something new. As we have shown, the subjective tests, such as PSS-14 or STAI, study the stress conditions during a period of time and not the instantaneous stress, as we have been working on. For this reason, in this section we present five new indicators that are of potential utility for the characterization of stress in educational contexts. In a final stage, these indicators could support the development of new services (recommendation systems, alert generators) and enable the development of extended student profiles.

4.1. *Smoothed Stress* (SS)

This is the primary reference stress indicator providing the stress level value at each time point. This is based on the predictions of a trained classifier. The classifier provides an instantaneous stress level (IS) of the user from the new sensors' raw data received. One issue of this indicator is that the results provided for each sample can produce oscillations for short time periods (cf. Fig. 8). These small oscillations are not relevant for the user and do not show a clear stress level. To manage this problem, we follow the following process (1):

- A 60 seconds temporal window is applied to the results taking the all the values estimated during that period.
- The state that is more repeated in such a period (mode) is taken. In other words, we use the mode statistic to detect the most repeated value in the period.
- Such a state is provided as the estimation for such period.

$$SS(t) = Mo(\{IS(t-60), IS(t-59) \dots IS(t)\}) \qquad (1)$$

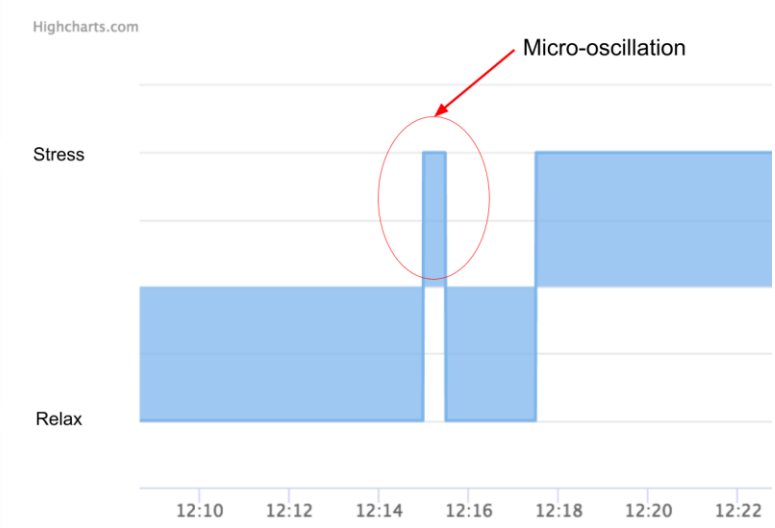

Fig. 8. Micro-oscillation in predicted stress.

4.2. Pattern of Stress Regularity (PSR)

In this case, we followed an approximation similar to sleep regularity indicator (Sano and Eng 2016). Variations in a basic index are compared during a certain period of time on different days to check if the same pattern is repeated. In our case, the Pattern of Stress Regularity (PSR) is calculated by measuring the similarity of the vector $\overrightarrow{(\mathrm{SS}_{norm})}$, that contains all the normalized stress measurements in a period of time (T) against the medium vector of stress in the same period of time ($\overrightarrow{MeanSS}$). Each element of the vector $\overrightarrow{(\mathrm{SS}_{norm})}$ takes a value 0 if the subject's state is relaxed, 0.5 in

case of stress and 1 in case of high stress. This indicator will allow us to check if stress varies along days. Next, the mathematical formula that calculates this indicator is presented (2).

$$PSR = \frac{\int_0^T \overrightarrow{\mathrm{SS}_{norm}}[t]\overrightarrow{MeanSS}[t]\, dt}{T} \quad (2)$$

4.3. Aggregated Stress (AS)

This indicator is calculated as the aggregation of SS values for a period of time, e.g. a whole day. As a result, we get a number that represents the percentage of stress experienced during such a period. While the SS offers an instantaneous value of the stress level at a certain time point, the AS offers the percentage of time in which the user has been stressed. One of the most significant periods is one that covers 24 hours a day. This measures the percentage of time a student feels stressed during a whole day. If the AS is high, it could indicate a possible burnout syndrome. Using this indicator, we could check if the AS from Monday to Friday, school days, is higher or lower than on Saturday and Sunday. We propose a graphic representation related to this indicator (cf. Fig. 9). This kind of representation shows not just the AS value, but also the percentages for each level and the number of periods and duration of each period for each level.

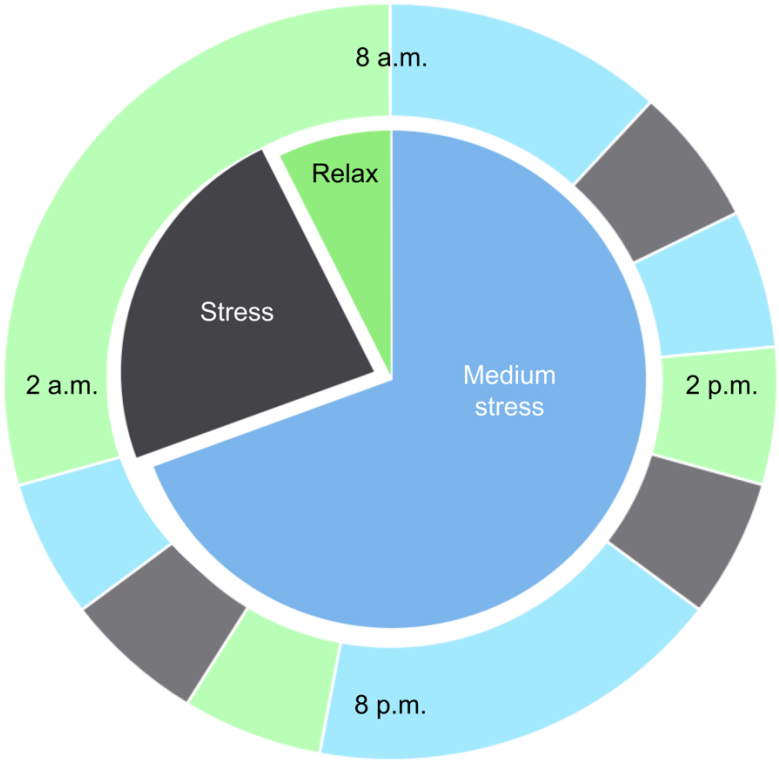

Fig. 9. Aggregated stress representation.

4.4. Stress Variability (SV)

Another factor that can influence the perception of stress is its variability. A scattered stress with periods of relaxation followed by periods of high stress can adversely affect the person. This indicator calculates the standard deviation (Std) of accumulated stress at different times, in periods of time similar to the duration of a class (60/120 minutes) $\overrightarrow{\mathrm{AS}_{subject}}$.

In case a student shows a high SV along a certain weekday, it is feasible that the student suffers during the performance of some activities and it is relaxed during other ones. This could be taken into account to schedule a better school activities and subjects programs. The calculation of this indicator is given by the following formula (3), where we have used 60 minute time periods.

$$SV = Std\left(\overrightarrow{\mathrm{AS}_{subject}}\right) \qquad (3)$$

4.5. Latent Stress (LS)

If the stress experienced by the user is extended on time, it can cause a stress-related condition, such as the Burnout syndrome (Maslach and Jackson 1981). This is a not desirable situation. It can produce discourage and lack of interest. Therefore, it is important to detect and manage lengthy stress situations. The LS indicator is proposed for this purpose. Meanwhile, the previous two indicators provide just information, the goal of this indicator is to alert to potentially dangerous situations.

This indicator is based on an idea used in other scientific areas, such as the declining curves using a roll-off rate α. Some examples of these systems can be found in the educational field as the Ebbinghaus forgetting curves (Ebbinghaus 2013; wranx 2016) or in training, as recovery status curves (Polar 2017). The Ebbinghaus curves represent the loss of memory retention in time and how the review of the learning material contributes to extend it. The recovery status curves indicate that the recovering time after performing a physical exercise is extended on time. They are used by vendors, such as Polar, to help users find the perfect balance between training and rest. Similarly, we consider the application of this idea to the stress, representing that it is extended on time in accordance with a declining curve. This provides a new indicator known as latent stress (LS). A low value of this indicator represents low stress over time, while high value indicates that the user has been subject to a high pressure for a long time.

The goal is to provide a stress indicator that varies in accordance with the duration of stress periods proportionally, and that is decreased by a K-factor. For example, is the subject experiences stress during half an hour then the latent stress will take a certain value. In the case K-factor=1, the declining curve will take the latent stress to 0 after half an hour without any stress.

To calculate the LS at a certain time, the following formulas are used (4) and (5):

$$LS(t) = LS(t-1) - K \text{ if } K \leq LS(t-1) \quad (4)$$

$$LS(t) = 0 \text{ if } K > LS(t-1) \quad (5)$$

Where:
K is the stress recovery factor.

In case a new stress level is produced after a relax or medium stress state, the value of LS is updated each second in accordance with the following formula (6):

$$LS(t) = LS(t-1) + 1 \quad (6)$$

An example of the LS variations on time with K=1 can be seen at Fig. 10.

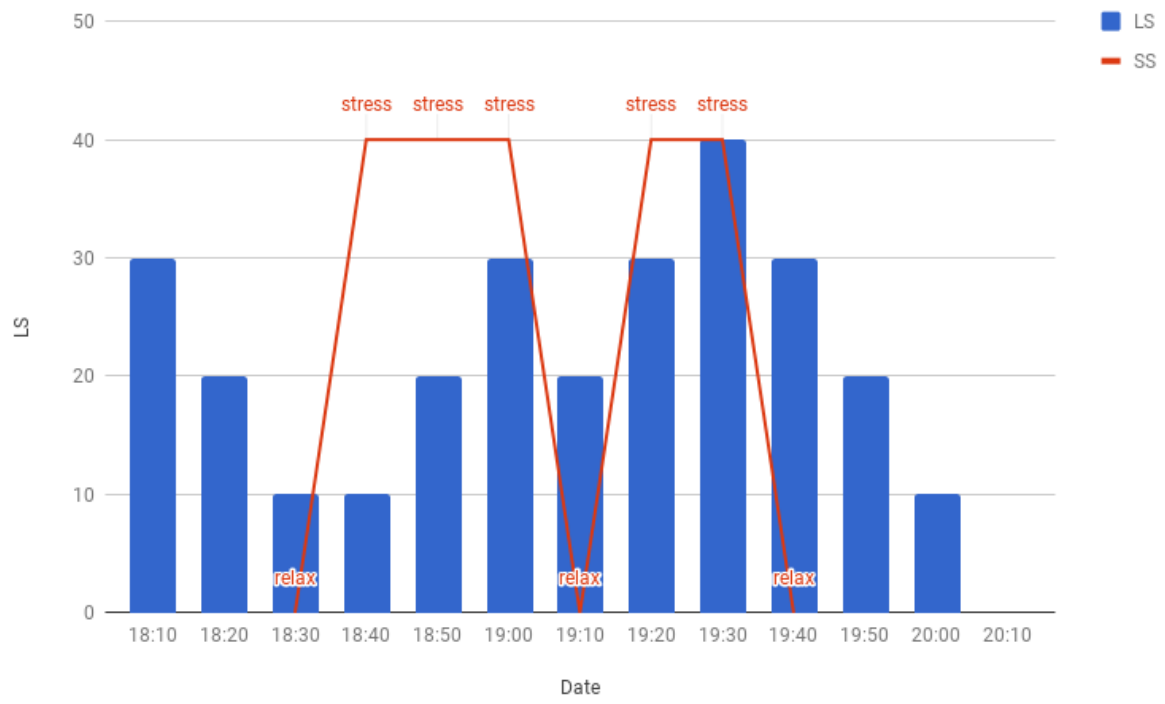

Fig. 10. Variation of latent stress.

5. Conclusions

Despite the significant amount of scientific literature about stress, there are not many solutions to measure it in a non-intrusive and simple way. The use of subjective tests only allows to estimate stress at certain time points, but not to measure it in a continuous basis, following its evolution. There exist clinical devices that allow to detect stress from physiological signals, but clinical procedures need to be supervised in hospitals, managed by experts and in some situations introduce psychological biases (Mario et al. 2009; Ibáñez et al. 2018). New COTS wrist wearables offer significant advantages related to the comfortability and automatization to collect physiological signals and detect information of interest such as stress states. The issue that we have tried to answer in this paper is if these devices are good enough to provide valid measures of stress.

After a thorough analysis of the available sensor's variables, we show COTS wrist wearables can be used to obtain classifiers with a high accuracy in stress estimation. This conclusion is provided taking into account the results obtained in several analysis. Firstly, results shown a 0.90 accuracy and F-measure in practically all classifiers analyzed in an individual subject model and a similar good result in the first global model analyses. Secondly, the good results were confirmed by a second experiment, where a classifier was trained with 18 subjects and tested with another one. The accuracy and F-measure results were well below the individual model. Nevertheless, when the samples of subjects who presented anomalous behavior were removed, the metrics of the two best classifiers approached 0.8. Finally, a new experiment was performed training an LWL classifier, the one which had provided the best results, with the samples of 19 subjects and tested with 5 new subjects. This time, the accuracy and F-measure values were greater than 0.8 and on average close to 0.9. Notice the classifiers are evaluated with the parameters by default defined in the library of Weka

functions. The improvement of these parameters would allow to offer a general performance much greater than the provided one, especially in the case of the Neural Network.

Several developments have been needed to make possible these results. Particularly, the app to induce stress can be an interesting resource, because it enables to reproduce the experiments in the same conditions. Despite the good results obtained and the possible applications, it is important to notice that there is a long way to walk until they can be used in everyday basics. One important issue is related to the reduced number of COTS wrist wearables providing the sensors needed to collect the required physiological signals. This type of wearables is oriented to support physical training activities, and they do not currently include sensors such as GSR or skin temperature. In another way, the preparation of the experiment also revealed some problems, particularly related to the GSR sensor. This sensor is sensitive to position changes, and when the device moves over the wrist, the GSR measure can be reset. Therefore, during the experiments, special care was taken to fix the wearable to the wrist. In general, these devices also present issues related to the battery-life duration, the memory available or the interoperability issues to enable third-party systems to collect data from them. Despite these problems, it is important to notice that the wearable domain is in continuous evolution. Therefore, as in the case of the HR sensor (it was almost inexistent two years ago but has been included by many vendors in their new products), we hope that sensors such as the GSR and the ST meters would become more common in future COTS wearables.

The app to induce stress reproduced some broadly-used exercises already described in the literature. However, recent studies indicate that stress can be the product not only of the performance of a specific activity, but also the result of an emotional aspect and possible environmental factors. That is, a test may induce stress in a subject not just because the direct interactions, but also by the emotions that it evokes. Some authors (Rincon et al. 2018; Costa et al. 2019) use cameras or ad-hoc devices in order to estimate these emotions. The identification of these emotions can be used to create a model of enriched stress. However, in this paper we have not been able to approach this analysis, although we consider it is really interesting to propose an experiment that, in addition to evaluating stress, evaluates the emotions of the individual and therefore generates a more "real" measure of stress.

Besides, we propose other stress-related indicators that can be of interest. They are based on the availability of continuous values of the stress level. This is a new research area still at an initial stage of development. From this point of view, it is still not clear how solutions and indicators can be integrated successfully in educational scenarios. Previously, we have been working in the development of sleep indicators (de Arriba-Pérez et al. 2017). Together with the stress indicators proposed in this paper, they could be used to enhance the learner models used in many educational systems: recommendation systems, cognitive tutors, affective tutors, etc. Beyond this, we consider other possible areas of application:

- Self-regulated learning. The availability of the stress indicators can be used to make learners more aware of their own features and conditions. They could get better self-knowledge. If their study habits and learning strategies are provoking high-stress levels and event latent stress, they should consider some changes.
- Improve learning activities. Anyone can easily understand that examinations produce a high-stress level in learners. Nevertheless, it is not so easy to know what learners feel more stressed performing group activities or the stress level during a particular lecture. Therefore, the stress level could be used to classify the learning activities in accordance with the stress experienced by learners. In another way, it is possible to measure the stress experienced by learners during an exam, and we can get the stress accumulated during the days previous

to the exam. Also, if we take student stress values during the performance of assessment tasks in accordance to different assessment strategies (e.g. continuous assessment vs final assessment), we could know the stress level of each one of the tasks and strategies. For otherwise, this can suggest the need of a certain amount of relaxing activities during the next days to an exam.

All this information can be of high value in the educational environments. The development of solutions to provide this kind of indicators and their use in combination with other ones, opens the window to new opportunities that need to be conceived, yet. These new research fields using these devices offer a new opportunities in the education domains and others important areas such as health(Din and Paul 2018; Cola and Vecchio 2018).

6. Acknowledge

This work is supported by the Spanish State Research Agency, the European Regional Development Fund (ERDF) under the PALLAS (TIN2016-80515-R AEI/EFRD, EU) project and the employment contract granted by the University of Vigo in July 2016 for the performance of PhD studies.

7. Conflicts of Interest

Conflicts of Interest: The authors declare no conflict of interest.

8. References

Beck A, Steer R (1990) Manual for the Beck anxiety inventory. San Antonio, TX Psychol Corp

Ben-Zeev D, Scherer EA, Wang R, et al (2015) Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 38:218–226. doi: 10.1037/prj0000130

Brown TA, Chorpita BF, Korotitsch W, Barlow DH (1997) Psychometric properties of the Depression Anxiety Stress Scales (DASS) in clinical samples

Burns A, Greene BR, McGrath MJ, et al (2010) SHIMMER™ – A Wireless Sensor Platform for Noninvasive Biomedical Research. IEEE Sens J 10:1527–1534. doi: 10.1109/JSEN.2010.2045498

Caddy B (2018) Stress tracking tech: Heart rate monitoring and guided breathing devices. https://www.wareable.com/wearable-tech/stress-beating-tech-to-keep-you-sane. Accessed 15 Nov 2018

Cano-Vindel A, Miguel-Tobal JJ (1999) Evaluación de la ansiedad desde un enfoque interactivo y multidimensional: el Inventario de Situaciones y Respuestas de Ansiedad (ISRA). Psicol Contemp 6:14–21

Cano A, Miguel-Tobal JJ, González H, Iruarrizaga I (2007) Hiperventilación y experiencia de ansiedad. Ansiedad y Estrés 13:291–302

Carroll B, Davidson J (2000) Screening Scale for DSM-IV GAD

Chandola T, Brunner E, Marmot M (2006) Chronic stress at work and the metabolic syndrome: prospective study. BMJ 332:521–5. doi: 10.1136/bmj.38693.435301.80

Cohen S, Kamarck T, Mermelstein R (1983) A global measure of perceived stress. J Health Soc Behav 385–396

Cohen S, Williamson G (1988) Perceived stress in a probability sample of the United States. S Spacapan S Oskamp (Eds), The Soc Psychol Heal Claremont Symp Appl Soc Psychol Newbury Park CA Sage

Cola G, Vecchio A (2018) Wearable systems for e-health and wellbeing. Pers Ubiquitous Comput 22:225–225. doi: 10.1007/s00779-017-1041-1

Colligan TW, Higgins EM (2006) Workplace Stress: Etiology and consequences. J Workplace Behav Health 21:89–97. doi: 10.1300/J490v21n02_07

Cooper CL, Cartwright S (1997) An intervention strategy for workplace stress. J Psychosom Res 43:7–16. doi: 10.1016/S0022-3999(96)00392-3

Costa A, Rincon JA, Carrascosa C, et al (2019) Emotions detection on an ambient intelligent system using wearable devices. Futur Gener Comput Syst 92:479–489. doi: 10.1016/J.FUTURE.2018.03.038

Dallman MF, Pecoraro N, Akana SF, et al (2003) Chronic stress and obesity: A new view of "comfort food." PNAS 97:325–330. doi: 10.1073/pnas.97.1.325

de Arriba-Pérez F, Caeiro-Rodríguez M, Santos-Gago JM (2016) Collection and Processing of Data from Wrist Wearable Devices in Heterogeneous and Multiple-User Scenarios. Sensors 16:1538. doi: 10.3390/s16091538

de Arriba-Pérez F, Caeiro-Rodríguez M, Santos-Gago JM (2017) How do you sleep? Using off the shelf wrist wearables to estimate sleep quality, sleepiness level, chronotype and sleep regularity indicators. J Ambient Intell Humaniz Comput 1–21. doi: 10.1007/s12652-017-0477-5

de Arriba-Pérez F, Santos-Gago JM, Caeiro-Rodríguez M, Fernández-Iglesias MJ (2018) Evaluation of Commercial-Off-The-Shelf Wrist Wearables to Estimate Stress on Students. JoVE 9. doi: doi:10.3791/57590

Deberard C, Scott M, Glen I, et al (2004) Predictors of academic achievement and retention among college freshmen: a longitudinal study. Coll Stud J 381:66–80

Din S, Paul A (2018) Smart health monitoring and management system: Toward autonomous wearable sensing for internet of things using big data analytics. Futur Gener Comput Syst. doi: 10.1016/J.FUTURE.2017.12.059

Dishman RK, Nakamura Y, Garcia ME, et al (2000) Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. Int J Psychophysiol 37:121–133. doi: 10.1016/S0167-8760(00)00085-4

Ebbinghaus H (2013) Memory: a contribution to experimental psychology. Ann Neurosci 20:155–6. doi: 10.5214/ans.0972.7531.200408

empatica (2016) E4 wristband. https://www.empatica.com/e4-wristband. Accessed 26 Jun 2017

Espinosa HG, Lee J, Keogh J, et al (2015) On the Use of Inertial Sensors in Educational Engagement Activities. Procedia Eng 112:262–266. doi: 10.1016/j.proeng.2015.07.242

Extremera N, Durán A, Rey L (2007) Inteligencia emocional y su relación con los niveles de burnout, engagement y estrés en estudiantes universitarios. 342:239–256

Fan Q, Wang Y (2010) The real-time realization of filtering of speech with DSP TMS320VC5416 Chip. In: 2010 International Conference on Educational and Information Technology. IEEE

García-Ros R, Pérez-González F, Pérez-Blasco J, Natividad LA (2012) Evaluación del estrés académico en estudiantes de nueva incorporación a la universidad Academic stress in first-year college students. 143–154

Gonzáles–Romá V, Schaufeli W, Bakker A (2002) The measurement of burnout and engagement: A confirmatory factor analytic approach. Jou Happ Stu

Grös DF, Antony MM, Simms LJ, Mccabe RE (2007) Psychometric Properties of the State–Trait Inventory for Cognitive and Somatic Anxiety (STICSA): Comparison to the State–Trait Anxiety Inventory (STAI). Psychol Assess 19:369. doi: 10.1037/1040-3590.19.4.369

Guo F, Li Y, Kankanhalli MS, Brown MS (2013) An evaluation of wearable activity monitoring devices. In: Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia - PDM '13. ACM Press, New York, New York, USA, pp 31–34

Hamilton M (1959) The assessment of anxiety states by rating. Br J Med Psychol 32:50–55. doi: 10.1111/j.2044-8341.1959.tb00467.x

Harari GM, Gosling SD, Wang R, et al (2017) Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. doi: 10.1016/j.chb.2016.10.027

Harari GM, Lane ND, Wang R, et al (2016) Using Smartphones to Collect Behavioral Data in Psychological Science. Perspect Psychol Sci 11:838–854. doi: 10.1177/1745691616650285

Healey JA (2000) Wearable and automotive systems for affect recognition from physiology

Healey JA, Picard RW (2005) Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. IEEE Trans Intell Transp Syst 6:156–166. doi: 10.1109/TITS.2005.848368

Hernandez J, Morris RR, Picard RW (2011) Call Center Stress Recognition with Person-Specific Models. Springer, Berlin, Heidelberg, pp 125–134

Highcharts (2017) Interactive JavaScript charts for your webpage | Highcharts. https://www.highcharts.com/. Accessed 11 Jan 2018

Ibáñez V, Silva J, Cauli O (2018) A survey on sleep assessment methods. PeerJ 6:e4849. doi: 10.7717/peerj.4849

IDC (2017) IDC Forecasts Shipments of Wearable Devices to Nearly Double by 2021 as Smart Watches and New Product Categories Gain Traction. https://www.idc.com/getdoc.jsp?containerId=prUS43408517. Accessed 7 May 2018

IDC (2016a) The Worldwide Wearables in 2015, According to IDC. http://www.idc.com/getdoc.jsp?containerId=prUS41037416. Accessed 26 Jun 2017

IDC (2016b) Worldwide Wearables Market Increases 67.2% Amid Seasonal Retrenchment, According to IDC. http://www.idc.com/getdoc.jsp?containerId=prUS41284516. Accessed 26 Jun 2017

IDC (2016c) Basic Wearables Soar and Smart Wearables Stall as Worldwide Wearables Market Climbs 26.1% in the Second Quarter. http://www.idc.com/getdoc.jsp?containerId=prUS41718216. Accessed 26 Jun 2017

IDC (2016d) IDC Forecasts Wearables Shipments to Reach 213.6 Million Units Worldwide in 2020 with Watches and Wristbands Driving Volume While Clothing and Eyewear Gain Traction. http://www.idc.com/getdoc.jsp?containerId=prUS41530816. Accessed 26 Jun 2017

Jersey (2016) Jersey. https://jersey.java.net/. Accessed 26 Jun 2017

Karthikeyan P, Murugappan M, Yaacob S (2012) Descriptive Analysis of Skin Temperature Variability of Sympathetic Nervous System Activity in Stress. J Phys Ther Sci 24:1341–1344. doi: 10.1589/jpts.24.1341

Kikhia B, Stavropoulos TG, Meditskos G, et al (2015) Utilizing ambient and wearable sensors to monitor sleep and stress for people with BPSD in nursing homes. J Ambient Intell Humaniz Comput 1–13. doi: 10.1007/s12652-015-0331-6

Kitsantas A, Winsler A, Huie F (2008) Self-Regulation and Ability Predictors of Academic Success During College: A Predictive Validity Study. J Adv Acad 20:. doi: 10.4219/jaa-2008-867

Kompier M, Cooper C (1999) Preventing stress, improving productivity: European case studies in the workplace

Koskimäki H, Mönttinen H, Siirtola P, et al (2017) Early detection of migraine attacks based on wearable sensors. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers on - UbiComp '17. ACM Press, New York, New York, USA, pp 506–511

Kothgassner OD, Felnhofer A, Hlavacs H, et al (2016) Salivary cortisol and cardiovascular reactivity to a public speaking task in a virtual and real-life environment. Comput Human Behav 62:124–135. doi:

10.1016/J.CHB.2016.03.081

Lin T, Omata M, Hu W, Imamiya A (2005) Do physiological data relate to traditional usability indexes? Proc 17th Aust Conf Comput Interact Citizens Online Considerations Today Futur Comput Interact Spec Interes Gr Aust 1–10

Lovibond S, Lovibond P (1995) Manual for the depression anxiety stress scales. Hum Reprod

Lu L (1994) University transition: major and minor life stressors, personality characteristics and mental health. Psychol Med 24:81. doi: 10.1017/S0033291700026854

Lundberg U, Kadefors R, Melin B, et al (1994) Psychophysiological stress and emg activity of the trapezius muscle. Int J Behav Med 1:354–370. doi: 10.1207/s15327558ijbm0104_5

Mario B, Massimiliano M, Chiara M, et al (2009) White-coat effect among older patients with suspected cognitive impairment: prevalence and clinical implications. Int J Geriatr Psychiatry 24:509–517. doi: 10.1002/gps.2145

Mariotti A (2015) The effects of chronic stress on health: new insights into the molecular mechanisms of brain–body communication. Futur Sci OA 1:fso.15.21. doi: 10.4155/fso.15.21

Mark H, Ian W, Eibe F (2011) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers

Maslach C, Jackson SE, Schwab RL (1986a) The MBI-Educators Survey. The Maslach

Maslach, Jackson S, Leiter M (1986b) Maslach Burnout Inventory. Palo Alto

Maslach, Jackson SE (1981) The measurement of experienced burnout*. J Occup Behav 2:99–113

Mastrandrea R, Fournet J, Barrat A (2015) Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys. PLoS One 10:e0136497. doi: 10.1371/journal.pone.0136497

Mayo Clinic Staff (2016) Chronic stress puts your health at risk - Mayo Clinic. https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/stress/art-20046037. Accessed 24 Oct 2018

Microsoft (2015) Microsoft Band SDK. https://developer.microsoftband.com/Content/docs/Microsoft Band SDK.pdf. Accessed 26 Jun 2017

Mohr DC, Jorm A, Saeb S, et al (2016) The relationship between mobile phone location sensor data and depressive symptom severity. doi: 10.7717/peerj.2537

Mokhayeri F, Akbarzadeh-T M-R, Toosizadeh S (2011) Mental stress detection using physiological signals based on soft computing techniques. In: 2011 18th Iranian Conference of Biomedical Engineering (ICBME). IEEE, pp 232–237

MongoDB (2017) MongoDB for GIANT Ideas | MongoDB. https://www.mongodb.com/. Accessed 26 Jun 2017

Norton PJ (2007) Depression Anxiety and Stress Scales (DASS-21): Psychometric analysis across four racial groups. Anxiety, Stress Coping 20:253–265. doi: 10.1080/10615800701309279

Pedrotti M, Mirzaei MA, Tedesco A, et al (2014) Automatic Stress Classification With Pupil Diameter Analysis. Int J Hum Comput Interact 30:220–236. doi: 10.1080/10447318.2013.848320

Polar (2017) Recovery status | Polar Global. https://www.polar.com/en/smart_coaching/features/recovery_status. Accessed 26 Jun 2017

Prieto LP, Sharma K, Dillenbourg P, Rodríguez-Triana MJ (2016) Teaching Analytics: Towards Automatic Extraction of Orchestration Graphs Using Wearable Sensors. Proc Sixth Int Conf Learn Anal Knowl 148–157. doi: 10.1145/2883851.2883927

Rashkova MR, Ribagin LS, Toneva NG (2012) Correlation between salivary α-amylase and stress-related anxiety. Folia Med (Plovdiv) 54:46–51. doi: 10.2478/v10153-011-0088-4

Reiss S, Peterson RA, Gursky DM, McNally RJ (1986) Anxiety sensitivity, anxiety frequency and the prediction of

fearfulness. Behav Res Ther 24:1–8. doi: 10.1016/0005-7967(86)90143-9

Rincon JA, Costa A, Villarrubia G, et al (2018) Introducing dynamism in emotional agent societies. Neurocomputing 272:27–39. doi: 10.1016/J.NEUCOM.2017.03.091

Sandhu MM, Javaid N, Jamil M, et al (2015) Modeling mobility and psychological stress based human postural changes in wireless body area networks. Comput Human Behav 51:1042–1053. doi: 10.1016/J.CHB.2014.09.032

Sano A, Eng B (2016) Measuring College Students' Sleep, Stress, Mental Health and Wellbeing with Wearable Sensors and Mobile Phones. Massachusetts Institute of Technology

Santos A de (2012) Design, implementation and evaluation of an unconstrained and contactless biometric system based on hand geometry and stress detection. E.T.S.I. Telecomunicación (UPM)

Schaufeli W, Leiter M (1996) Maslach burnout inventory-general survey. Maslach Burn Invent Man 1:19–26

Selye H (1973) The Evolution of the Stress Concept: The originator of the concept traces its development from the discovery in 1936 of the alarm reaction to modern therapeutic applications of syntoxic and catatoxic hormones. Am Sci 61:692–699

Setz C, Arnrich B, Schumm J, et al (2010) Discriminating Stress From Cognitive Load Using a Wearable EDA Device. IEEE Trans Inf Technol Biomed 14:410–417. doi: 10.1109/TITB.2009.2036164

Shimmer Shimmer Galvanic Skin Response Sensor | EDA sensor. http://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor. Accessed 18 Dec 2018

Spielberger CD, Gorsuch RL, Lushene RE (1970) Manual for the State-Trait Anxiety Inventory

Stahl SE, An H-S, Dinkel DM, et al (2016) How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? BMJ Open Sport Exerc Med 2:e000106. doi: 10.1136/bmjsem-2015-000106

Statista (2017) Fitbit Leads Global Wearables Market. https://www.statista.com/chart/8420/wearable-device-shipments/. Accessed 16 May 2018

Statista (2018) Apple Jumps to Top of the Global Wearables Market. https://www.statista.com/chart/13115/worldwide-wearable-device-shipments/. Accessed 16 May 2018

Stroop JR (1935) Studies of interference in serial verbal reactions. J Exp Psychol 18:643–662. doi: 10.1037/h0054651

Taylor S, Zvolensky MJ, Cox BJ, et al (2007) Robust dimensions of anxiety sensitivity: Development and initial validation of the Anxiety Sensitivity Index-3. Psychol Assess 19:176–188. doi: 10.1037/1040-3590.19.2.176

Tombaugh TN (2006) A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). Arch Clin Neuropsychol 21:53–76. doi: 10.1016/j.acn.2005.07.006

Travers CJ, Cooper CL (1997) El Estrés de los profesores : la presión en la actividad docente. Paidós

Vizer LM, Zhou L, Sears A (2009) Automated stress detection using keystroke and linguistic features: An exploratory study. Int J Hum Comput Stud 67:870–886. doi: 10.1016/j.ijhcs.2009.07.005

Vrijkotte TGM, Van-Doornen LJP, De-Geus EJC (2000) Effects of Work Stress on Ambulatory Blood Pressure, Heart Rate, and Heart Rate Variability. Hypertension 35:

Wallen MP, Gomersall SR, Keating SE, et al (2016) Accuracy of Heart Rate Watches: Implications for Weight Management. PLoS One 11:e0154420. doi: 10.1371/journal.pone.0154420

Wang R, Blackburn G, Desai M, et al (2017) Accuracy of Wrist-Worn Heart Rate Monitors. JAMA Cardiol 2:104. doi: 10.1001/jamacardio.2016.3340

Wang R, Chen F, Chen Z, et al (2014) StudentLife. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct. ACM Press, New York, New York, USA, pp 3–14

Wang R, Harari G, Hao P, et al (2015) SmartGPA: How Smartphones Can Assess and Predict Academic Performance

of College Students. doi: 10.1145/2750858.2804251

Wittchen H-U, Boyer P (1998) Screening for anxiety disorders: Sensitivity and specificity of the Anxiety Screening Questionnaire (ASQ—15). Br J Psychiatry

wranx (2016) Ebbinghaus and the forgetting curve. http://www.wranx.com/ebbinghaus-and-the-forgetting-curve/. Accessed 26 Jun 2017

Xu J, Zhong B (2018) Review on portable EEG technology in educational research. Comput Human Behav 81:340–349. doi: 10.1016/J.CHB.2017.12.037

Zhai J, Barreto A (2006) Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp 1355–1358

Zhai J, Barreto AB, Craig-Chin, Chao-Li (2005) Realization of Stress Detection using Psychophysiological Signals for Improvement of Human-Computer Interaction. In: Proceedings. IEEE SoutheastCon. IEEE, pp 415–420