Computer Science and Information Systems 15(3):683–703 https://doi.org/10.2298/CSIS180115033G

X3S: A Multi-modal Approach to Monitor and Assess Stress through Human-computer Interaction ⋆

Filipe Gonçalves[1], Davide Carneiro[1,2], José Pêgo[3,4], and Paulo Novais[1]

[1] Algoritmi Research Centre/Department of Informatics
University of Minho, Braga, Portugal
[1]fgoncalves@algoritmi.uminho.pt, {dcarneiro,pjon}@di.uminho.pt
[2] CIICESI, ESTG, Polytechnic Institute of Porto
Felgueiras, Portugal
[3] School of Health Sciences, Life and Health Sciences Research Institute (ICVS)
University of Minho, Braga, Portugal
[3,4]jmpego@ecsaude.uminho.pt
[4] ICVS/3B's - PT Government Associate Laboratory, Braga / Guimarães, Portugal

Abstract. Stress evaluation is nowadays gaining an increasing importance in a time in which inter-individual competition continuously pushes us to be better. Indeed, in the workplace, in the academia or in many other contexts there is increasing pressure for better performance, which pushes us forward but also wears us out. This phenomenon has been studied from many different angles, including psychology, medicine or occupational dynamics. In a medical or biological context, stress is a physical, mental, or emotional factor that causes bodily or mental tension, which can cause or influence the course of many medical conditions including psychological conditions such as depression and anxiety. In these cases, individuals are under an increasing demand for performance, driving them to be under constant pressure, and consequently to present variations in their levels of stress. To mitigate this condition, this paper proposes to add a new dimension in human–computer interaction through the development of a distributed multi-modal framework approach entitled X3S, which aims to monitor and assess the psychological stress of computer users during high-end tasks, in a non-intrusive and non-invasive way, through the access of soft sensors activity (e.g. task performance and human behaviour). This approach presents as its main innovative key the capacity to validate each stress model trained for each individual through the analysis of cortisol and stress assessment survey data. Overall, this paper discusses how groups of medical students can be monitored through their interactions with the computer. Its main aim is to provide a stress marker that can be effectively used in large numbers of users and without inconvenience.

Keywords: Human-computer Interaction, Behavioural & Performance Patterns, Machine Learning, Stress Monitoring & Assessment

1. Introduction

The evolution of brain structures and cognitive abilities in most species has been driven by the need to undertake social interactions. This happens in species with complex social exchanges, including us humans [23]. These evolutive processes are starting to be

⋆ The present work is an extension of a conference paper disclosed in [24].

implemented in machines, with a clear and priority goal: to develop socially intelligent machines that can be integrated into our daily routines [35]. Thus, computers cease to be a simple tool and gain human-like social abilities[41]. While many domains contribute to the eventual achievement of this goal (e.g. psychology, neuroscience, biology, affective computing), Social Signal Processing (SSP) is perhaps one of the most relevant as it focuses on the analysis and modelling of non-verbal behaviour in social interactions by machines, as well as on its production [35]. Thus, machines can take part in human social interactions in a seamless way, and understand behavioral cues (such as facial expressions, gestures or postures) as well as actually producing them to hint attitudes, intentions or beliefs. This is typically carried out using human-like avatars or similar mechanisms, that increase the realism and efficacy of such approaches, making them more beliavable for humans.

When implementing this kind of artificial systems, the key idea is to take advantage of our own cues, which are modelled by physiological and psychological phenomena, and that we use while interacting. These are sometimes hard to detect consciously by Humans, but can be detectable by machines [29]. Tho be able to make this detection, there are typically two complementary paths [41]: the first is to detect the physical characteristics of social signals using a variety of sensors including video cameras and microphones; the second is to interpret such signals in light of the existing knowledge regarding our non-verbal interactions.

From a social psychological point of view, social signals include any behaviour aimed at engaging others in a joint activity, often communication [7], where these signals are defined as "communicative or informative signals which provide information about social facts", i.e. social (inter)actions, social emotions, social evaluations, social attitudes and social relations.

The term *stress* is commonly used to describe a set of physical and physiological responses that emerge as a reaction to a challenging stimulus that alter an organism's environment [34]. Perceiving an individual's physiological stress level is nowadays viewed as an important factor to manage individual performance, in a time when individual and team limits are pushed further, especially in environments such as the workplace or in the academia[40].

Indeed, prolonged exposure to stress-inducing factors is a growing concern, especially in complex activities, which require great responsibility and reliability. This type of activities can lead to states of emotional exhaustion (burnout) and other mental disorders [21] (e.g. depression, chronic stress, chronic diseases), with potential consequences at the personal, professional, family, social and economic levels [17].

Some good examples of stressful environments can be seen everywhere in our life, from workplaces [10] to academia [36]. In this paper we focus on academic stressors, which include the student's perception of the extensive knowledge base required and the perception of an inadequate time to develop it [11]. Students report experiencing academic stress at predictable times each semester with the greatest sources of stress resulting from taking and studying for exams, grade competition, and the large amount of content to master in a small amount of time [1]. This results in a high prevalence of anxiety disorders among higher education students.

Existing stress monitoring approaches rely on the use of complex or expensive hardware, or in the collection of biological variables, all of which require the use of sensors to collect data directly from the body of the individual. One of the limitations of this kind of approaches in these environments is that it alters the routine of the student, which is not desirable, especially in an already potentially stressful situation as an high-stakes exam [16]. Moreover, if dozens or hundreds of individuals are being monitored simultaneously (as is the case), an equivalent number of sensors is required. This type of approaches has an increased cost (due to the hardware) and complexity (due to the processing and analysis of multiple physiological streams of data).

The present work is an extension of the one disclosed in [24] which proposes a distributed system for monitoring human behaviour with the aim of measuring stress based on two different modalities of variables: behavioural and performance. Specifically, the paper discusses how groups of medical students can be monitored through their interaction with the computer and their decision-making during the exam [33], in order to study the effect of stress on performance during high-demand tasks.

The main goal is to provide a non-intrusive and lightweight stress marker that can be effectively used in large numbers of students, without inconveniences. Information on how stress affects each student will make it possible to improve individualised teaching strategies as well as to empower these students with better coping strategies. All this will result in the development of better future professionals.

This article is organised as follows: Section 2 provides related technological works about the existing projects that revolve around the use of behavioural biometrics analysis to solve different problems and the key innovative aspects of our work. The architecture of the system is disclosed in section 3. Section 4 describes the set of precautions considered for obtaining the input from the selected population, taking into account their environment and routines. Section 5 defines the preparation data collection processes required for future data analysis and machine learning. In Section 6 an analysis of the processed data is done regarding the features performance variations of medical students between different academic years. Finally section 7 presents conclusions about the work developed so far and future work considerations.

2. Related Work

The current pace of our lifestyle has led to the recognition of stress as a major concern of health organisations around the world.

We sometimes experience stress in the form of relatively short peaks, such as during an exam. In these cases, the response of our body (e.g. increased hearth and breath rhythm, tense muscles) evolved to prepare our body to react to the perceived challenge or threat [38]. It is often, in that sense, a positive process that increases our performance in the task at hand (also known as eustress [31]).

However, we may also experience stress over longer periods of time, such as workplace stress. In these cases, and depending on the intensity and frequency of the stressor agents, the effects may not be so positive. Indeed, when stress extends over long periods of time it tends to wear out the body and the mind, with significant negative consequences [17]. This form of stress is nowadays recognised as one of the major reasons causing health issues, with some authors pointing out stress as being wither the main reason or one of the responsible for 60% of all human ailments or other diseases [28].

The importance of knowing and controlling one's stress level is, therefore, undeniable: stress feedback may help to pinpoint stressors and their intensity, facilitating the design of effective coping strategies. Until recently, stress evaluation was mostly carried out using psychological instruments such as questionnaires, notably the Perceived Stress Scale (PSS) [14]. Some of these instruments where also digitised to be used in the form of mobile applications, such as in [4], a mobile application aimed at the auto-evaluation of stress by students.

Nonetheless, in recent years, a plethora of technology-supported approaches have been proposed that provided a significant contribution, namely towards the continuous analysis of stress and real-time feedback.

This type of systems generally rely on the sampling and processing of multiple streams of physiological signals, acquired from sensors placed on the user's body. The most frequently used sensors include ECG, heart rate, skin conductivity, respiratory rate, blood volume pulse or accelerometers. Not all of these systems are designed to be portable or mobile. That is, sometimes their use is restricted to contexts in which users are not moving, such as in the workplace. Others, however, are designed to be small-factor and to be carried around by the user. In this type of systems, one of the main concerns is to make its architecture low-power and have a low-area footprint, as addressed in [3].

These systems also vary significantly in complexity, cost and requirements/constraints according to the number and type of sensors used. In this regard, there are systems that rely on a single sensor such as [32], in which the authors detect stress remotely (up to a distance of 3 meters) using a five band digital camera that allows for the extraction of heart rate, breathing rate and heart rate variability. In [43], the authors present an energy-efficient system for stress assessment based on features extracted from an electrocardiogram signal, and in [30] the authors analyse stress using galvanic skin response alone.

Other approaches consider a variety of sensors, whether they are placed on a single piece of hardware or not. In [39], the authors use a wristband that provides features regarding galvanic skin response and blood volume pulse signals. On the other hand, in [28], the authors specifically develop a body area network of sensors to be carried by the users and that collects and transmits data from several physiological signals, that is later aggregated for the purpose of stress assessment.

Most or all of these systems share some common characteristics. They are based on one or more physiological signals. This implies the use of at least one physiological sensor per monitored user and possibly the existence of one battery (in the cases of mobile systems). If multiple users are to be monitored simultaneously, the corresponding number of sensors must be acquired and used. This may represent some constraints, namely regarding the cost of the monitoring. Moreover, and depending on the context, users are generally not prone to be monitored and that sensitive data (such as data physiological signals) be collected about them. This is especially true in more sensitive contexts such as the workplace. Finally, these approaches are also uni-modal, i.e. they are based on a single modality: the physiological one.

2.1. Key innovative aspects

In this paper we propose a novel approach in this field, that significantly differs from the general systems existing nowadays. Its key innovative points are:

1. Multi-modal approach - it combines data from two different modalities, to provide a broader view regarding the effects of stress on the individual: task performance and human behaviour;
2. No hard sensors - it uses no sensors in the traditional sense, i.e. specific hardware sensors, but rather relies on soft sensors, which significantly decreases operational and implementation costs;
3. Distributed and Scalable - the proposed system can be scaled to monitor hundreds of people simultaneously, without a significant increase in costs or complexity;
4. Validity - the stress model trained with this system can be validated for each individual using cortisol and perceived stress feedback.

(1) This architecture combines data from two different modalities in order to better understand how the user is being affected by stress in the task at hand. Indeed, while most of the existing systems are generic, the presented one is developed with a strong focus on the user-task binomial. Moreover, it is developed specifically to measure stress on office-like environments, such as many of current workplaces or the academia.

The first group of features, which compose the behavioural dimension of the model, is extracted from Mouse Dynamics. This modality, derived from the concept of behavioural biometrics, essentially quantifies the performance of the user in terms of human-computer interaction which, in previous work, we have shown to be significantly influenced by stress [8]. Behavioural biometrics is a relatively new form of analysis, which defines a field that extracts users' behavioural features from the use of the mouse and the keyboard [42]. These methods are mostly used for user identification and authentication (intelligent security systems) [5], which use multiple techniques for automatic recognition of individuals based on their physiological and/or behavioural characteristics. By using biometrics, it is possible to confirm or establish an individual's identity based on who the individual is, rather than by what the individual possesses (an ID card) or what the individual remembers (a password) [26]. Similarly, for the same individual, interaction patterns change according to situation, context, task or state. Observing these changes allow for detecting potential significant changes in the individual's context or state. This system is based on this precise notion.

The second group of features that compose the performance dimension is obtained by quantifying the user performance in the task being carried out. These performance measures may vary significantly according to the domain and it is the responsibility of a domain expert to define how they are calculated and to feed them to the system. In this paper, and in order to validate the proposed architecture, we conduct a case study in the academic domain, with medical students in computer-based high-stakes exams. Performance measures are obtained, in real-time, from MedQuizz: an e-assessment management system that thoroughly describes each student's actions during the exam (e.g. input of a correct/wrong answer, advancing to the next question, signalling a doubt).

These two groups of features (behavioural and performance-based) are combined based on their timestamps, which allow for an unified analysis of two very important dimensions when it comes to stress analysis.

(2) This approach is also innovative in the sense that it requires no hard sensors which constitute, often, the most expensive component of such systems, especially when the goal is to monitor large groups of people. Moreover, there are domains in which the placement and usage of these sensors is not desirable or ethic as it may interfere with

the task being performed or even with the variables being studies. For the case study presented in this paper we selected one such environment: high-stakes exams. Indeed, it is not desirable to place sensors on students who are in a particularly stressful and marking moment, as doing so might contribute to distract them or to stress them even further. Alternatively, behavioural variables are extracted from each user's computer by a locally installed application that collects system events regarding the interaction with the computer (e.g. mouse usage, keyboard typing). Performance variables, on the other hand, are collected and provided by the MedQuizz software, from the students' actions in the exam platform. These two applications act as soft sensors which provide a continuous stream of data regarding user behaviour and performance, in real-time.

(3) For the same reasons mentioned above, the proposed system can easily be scaled to hundreds or thousands of users. Moreover, the streams of data generated are significantly less complex than those generally generated by physiological sensors, which make its processing, analysis and storage significantly less complex and costly.

(4) Existing systems usually perform some form of unsupervised classification on the data as there is no actual way of validating physiological readings against an actual level of stress. That is, these system generally try to divide the observed data into two different groups, labelling them as "stressed" and "not stressed". The proposed system, on the other hand, accepts as input a validator that can be used to perform supervised classification techniques on the collected data, which increases the reliability and validity of the developed models. As an example, in the case study described in this paper, we use the difference of cortisol in the students' saliva between the beginning and the end of the exam and their perceived stress score (through the use of PSS) of students as markers of stress.

In conclusion, the proposed system entitled X3S system (previously called EUStress system in [24]) may constitute a very interesting tool for the contextualised analysis of stress in real-time, especially in environments such as workplaces or academia. It is contextualised in the sense that the collected data is inherently related to the task being performed and to *how* the task is being performed by the user. The acronym X3S represents the combination of three dimensional features analysed (behavioural, performance and stress markers variation patterns) and processed by the system to predict the user's stress state.

In the academic environment, as in the case-study detailed in this paper, this may be very important to properly contextualise each student's academic results. For example, is a given bad result due to the student's lack of knowledge or was she/he unable to cope with a high level of stress? This system may thus be used to point out those students who cope with stress more poorly, allowing to develop personalised stress coping strategies that can be beneficial not only for academic performance but also later in their professional careers. This is especially important in the medical domain, in which professionals frequently deal with peak levels of stress and in which good decision-making skill under stress are paramount.

3. Architecture

In order for the X3S framework to assess an individual's stress levels, it requires the analysis of two types of information: mouse interaction behaviour and decision-making behaviour. As such, the X3S system can be decomposed into the following components:

- *MedQuizz:* An e-assessment management system that enables trainers, educators and testing professionals to author, schedule, deliver, and report on surveys, quizzes, tests and exams, and an useful tool to create item banks. It allows the management of information about the quality of the items supporting the individual in the decision to design his/her assessments. It also has fail-safe features in the case of network failure and has functions capable of creating a log of the individuals actions[5]. This component is used to study the cognitive performance and the behaviour decision making patterns of the individual;
- *MouseDynamics:* A MedQuizz's module framework that includes not only the sheer acquisition and classification of the mouse input data, based on the biometric behaviour, but also a presentation tier that supports the human-based or autonomous decision-making mechanisms. [9, 8].

In this system, MedQuizz is the core management platform for the execution of all behaviour/task analysis modules. It has the ability to work in a SaaS (Software as a Service) environment, where the system is fully scalable and modular (features are turned on and off according to the users' permissions). Although task performance and human behaviour were the analysis modules applied in this system, other modules can be implemented into the framework. MedQuizz system works in all major web browsers and software packages/versions are distributed with native clients for Windows and Mac OS.

The use of saliva is an important biomarker of exam stress and a predictor of exam performance. In the study published in the "journal of psychosomatic research", done by Miri Cohen and Rabia Khalaila, it is shown that pH levels of saliva may serve as a reliable, accessible and inexpensive means by which to assess the degree of physiological reactions to exams and other naturalistic stressors [13]. Also, salivary cortisol is routinely used as a biomarker of psychological stress and related mental or physical diseases [25].

In our case study, a sample of the individual's saliva is taken before and after each exam, as a mean to analyse his/her levels of cortisol in their biological system, and to predict the levels of stress of the individual. The variance of the levels of cortisol is calculated, using formula (1), where α represents the identification of the individual, and β represents the identification of the exam.

$$DeltaCortisol_{\alpha,\beta} = PreCortisol_{\alpha,\beta} - PosCortisol_{\alpha,\beta} \tag{1}$$

The assessment of stress is further complemented through the Perceived Stress Scale tool (PSS-52), where each student provide their feedback each month. PSS-52 is a 52-item scale that assesses the perception of stressful experiences by asking the respondent to rate the frequency of his/her feelings and thoughts related to events and situations that occurred over the previous month [14, 15]. Notably, high PSS scores have been correlated with higher biomarkers of stress.

[5] The website of MedQuizz can be assessed at http://www.medquizz.com/

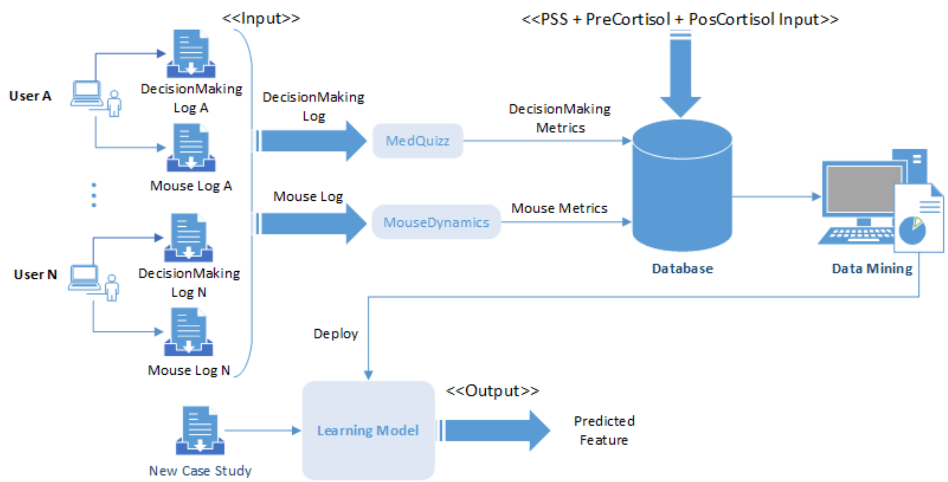

Fig. 1: X3S Dataflow System.

Since the study of behavioural features are mostly related to the individual's conduct habits, the calculation of stress level considers the individual's ID, actions of mouse and decisions made during the exams (CyberPsychological computation methods), which are monitored and acquired by the computation system [44]. After pre-processing, all data are stored in a database for machine learning. With these techniques, it allows us to study the relationships between the ubiquitous individual's psychological reactions and their behavioural patterns in cyber space for the psychological assessment of their situations in the learning process [18]. Through this system, it is intended in the future to develop a learning model capable of predicting the user's stress biomarkers values, based on their behavioural and performance features during on-line exams. The explained system is shown in Fig. 1.

3.1. MouseDynamics Performance Analysis

MouseDynamics Data Output: For this particular study, and given the aforementioned objectives, MouseDynamics model focused on the data collected from the mouse of the user's computers. The interaction of the user with the computer is monitored in terms of specific Operating System events fired from the use of the computer's mouse:

- *MOV, timestamp, posX, posY*: an event describing the movement of the mouse, in a given time, to coordinates (posX, posY) in the screen;
- *MOUSEDOWN, timestamp, [Left—Right], posX, posY*: this event describes the first half of a click (when the mouse button is pressed down), in a given time. It also describes which of the buttons was pressed (left or right) and the position of the mouse in that instant;
- *MOUSEUP, timestamp, [Left—Right], posX, posY*: an event similar to the previous one but describing the second part of the click, when the mouse button is released;

– *MOUSEWHEEL, timestamp, dif*: this event describes a mouse wheel scroll *dif*, in a given time;

MouseDynamics Analysed Features: MouseDynamics data collection output analyses the individual's mouse behaviour, and calculates his/her behavioural biometrics. These features aim at quantifying the individual mouse performance. Taking as example the movement of the mouse, one never moves it in a straight line between two points, there is always some degree of curve. The larger the curve, the less efficient the movement is [9, 8]. Some of the most important calculated metrics are presented in Fig. 2 and detailed in the following list:

– *Absolute Sum of Degrees (ASD):* Seeks to find how much the mouse turned, independently of the direction to which it turned (in degrees unit). The angle between the first line (defined by (x1,y1) and (x2,y2)) and the second line (defined by (x2,y2) and (x3,y3)) is given by equation $degree(x1, y1, x2, y2, x3, y3) = tan(y3 - y2, x3 - x2) - tan(y2 - y1, x2 - x1)$, where the absolute sum of degrees is depicted by equation (2);

$$ASD = \sum_{i=0}^{n-2} |degree(posx_i, posy_i, posx_{i+1}, posy_{i+1}, posx_{i+2}, posy_{i+3})|; \qquad (2)$$

– *Average Distance of the Mouse to the Straight Line (ADMSL):* Quantifies the average sum of the successive distances of the mouse to the straight line defined by two consecutive MOUSEUP and MOUSEDOWN events (in pixels);
– *Average Excess of Distance Between Clicks (AED):* Measures the average excess of distance that the mouse travelled between each two consecutive MOUSEUP and MOUSEDOWN events (in pixels);
– *Click Duration (CD):* Measures the timespan between two consecutive MOUSEUP and MOUSEDOWN events (in milliseconds). The longer the clicks, the less efficient the interaction is;
– *Distance Between Clicks (DBC):* Measures the total distance travelled by the mouse between two consecutive mouse clicks (in pixels), i.e. the distance of mouse movement between each two consecutive MOUSEUP and MOUSEDOWN events;
– *Mouse Velocity:* Distance travelled by the mouse (in pixels) over time (in milliseconds). The velocity is computed for each interval defined by two consecutive MOUSEUP and MOUSEDOWN events;
– *Mouse Acceleration:* The velocity of the mouse (in pixels/milliseconds) over time (in milliseconds). A value of acceleration is computed for each interval defined by two consecutive MOUSEUP and MOUSEDOWN events, using the intervals and data computed for the Mouse Velocity;
– *Time Between Clicks:* the timespan between two consecutive MOUSEUP and MOUSEDOWN events, i.e. how long did it took the individual to perform another click (in milliseconds).

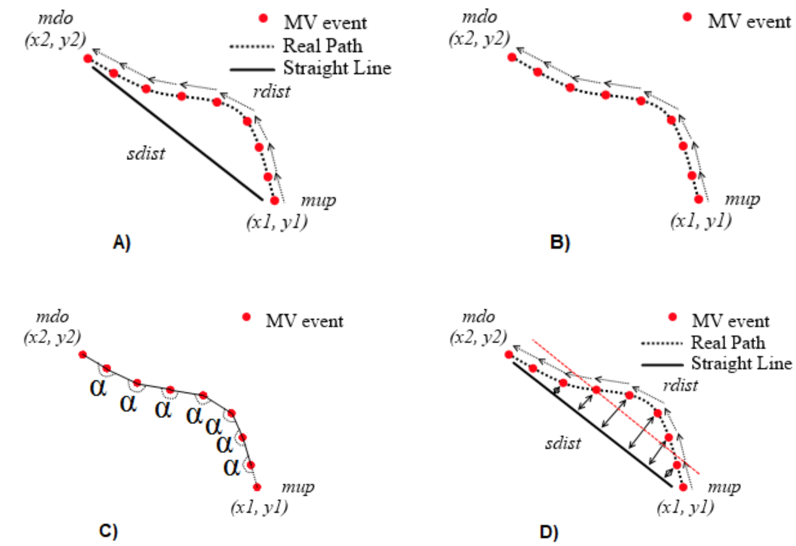

Fig. 2: (A) A series of mouse movement events (MV), between two consecutive clicks of the mouse. The difference between the shortest distance (sdist) and distance actually traveled by the mouse (rdist) is depicted; (B) The real distance traveled by the mouse between each two consecutive clicks is given by summing the distances between each two consecutive MV events; (C) The sum of the angles of the mouse's movement is given by summing all the angles between each two consecutive movement vectors; (D) The average distance at which the mouse is from the shortest line between two clicks is depicted by the straight dashed line.

3.2. MedQuizz Performance Analysis

MedQuizz Data Output: MedQuizz data collection output shows the individuals actions recorded during the exam into a log file. This file shows a list of actions, sorted by student and action time, making it possible to study and analyse the decision making behaviour of an individual. Some of the most important variables of this log are:

- *Exam ID:* Unique identification of the exam;
- *Student ID:* Unique identification of the individual/student;
- *Exam Question Number:* Unique identification of the question;
- *Action Timestamp:* Presents the action time of the decision making, written in date and hour format;
- *Action Description:* Defines the type of the decision made by the individual. For example, action description can show when the individual entered or left a question, an answer was inserted or removed, among other actions;
- *Action Result:* When an answer is inserted or changed, shows if the decision made was right or wrong, based on the correct answer for the question.

MedQuizz Analysed Features: By monitoring the actions log file of each individual during the execution of an exam, it is possible to analyse his/her decision making behaviour. Everything an individual does, consciously or unconsciously, is the result of some decision. The information we gather is to help us understand occurrences, in order to develop good judgements to make decisions about these occurrences. To make a decision, an individual needs to know the problem, the need and purpose of the decision, the criteria of the decision, and the alternative actions to take. Then there's the need to determine the best alternative [37]. Decision making, for which we gather most of our information, has become a mathematical science today. It formalises the thinking we use so that, what we have to do to make better decisions is transparent in all its aspects.

With that in mind, decision making analysis aims to evaluate the performance of the individual, based on the time between decisions, the correctness of the selected decisions, if those decisions serve the objectives of the decision maker, number of times a question was visualised, among other features. Some of the most important calculated behavioural features are:

- *Average Time Between Decision (ATBD):* This feature seeks to find the average time it takes each individual to take a decision. All decisions are taken into account. The decisions analysed vary from entering or leaving a question, inserting, changing or removing answers, marking or unmarking questions for review, among others (in milliseconds);
- *Median Time Between Decision (MTBD):* This feature quantifies the median of the time it takes each individual to take a decision (in milliseconds). The average is a measure greatly influenced by large or small number of values, even if these values appear in small numbers in the sample. These values are responsible for the misuse of the average in many situations where it would be more meaningful to use the median;
- *Standard Deviation / Variance Time Between Decision:* This feature measures the standard deviation/variance of the time between decisions for each individual (in milliseconds);
- *Average Time Between Questions (ATBQ):* This feature measures the average time the individual spent between all visualised question (in milliseconds);
- *Decision Making Ratio (DMR):* Verifies the ratio between the number of answers inserted, changed and removed and the total number of actions recorded (in percentage);
- *Correct Decision Making Ratio (CDMR):* Verifies the ratio between the number of decisions considered correct and the number of answers inserted, changed and removed (in percentage). A decision is considered correct when an individual inserts or changes an answer into a correct option or when he/she removes an incorrect answer from the question;
- *Final Grade:* This feature measures the total percentage of correct answers of an individual once the user confirms the completion of the task.

4. Study Design

In order to determine the stress levels of a group of individuals, based on their mouse behaviour performance and decision making behaviour performance, data was collected from the participation of a group of medical students in computer-based high stake exams. Through the evaluation of exams, these students test their academic knowledge in a monthly period. In these exams, students are indicated to their seats, and at the designated time they log in the exam platform using their personal credentials and the exam begins. The participation in the data-collection process does not imply any change in the student's routine, and all monitored metrics are calculated through background processes (using MedQuizz and MouseDynamics), making the collection data process completely transparent from the student's point of view, just like a normal routine exam. These exams consists mostly of single-best-answer multiple choice questions, where the students only use the mouse as an interaction means. When the exams end, students are allowed to leave the room. As explained in section 3, a sample of the individual's saliva is taken before and after the execution of each exam. Additionally, each month students provide information regarding their perceived stress through the use of PSS survey. From these samples, it will be later used to compare with the predicted stress levels computed by the developed approach, as a mean to validate its results and conclusions.

Specifically, the case study considers a group of 270 medical students (102 from the 1st year, 87 from the 2nd year and 81 from the 3rd year), which are monitored to study the effect of stress/anxiety in the performance of high demand tasks, such as the execution of exams. The methods used for data collection were taken into account, since any external factor can influence variations in the decision making and mouse performance behaviour of the individuals. As such, it is important to include non-intrusive and non-invasive measures as essential requirements during the execution of the exams.

5. Data Collection and Preparation

In order to analyse the data received from both analysis modules (MedQuizz and MouseDynamics), some conditions were required in the study.

5.1. Temporal Approaches

The first step of the preparation process is the verification of the variations in the decision making of the individual during an exam. In order to study these variations, the set of events of each individual are ordered by action event timestamp, cloned and prepared into five different datasets:

- *Chronological Time:* All actions collected are aggregated into intervals of five minutes (e.g. action events from minute 0-5, 5-10, 10-15, etc.) during the execution of an on-line exam by the individual;
- *Percentage Time:* All actions collected are aggregated into intervals, each one comprising 5% of the total duration of the exam spent by the individual (e.g. action events from 0-5%, 5-10%, 10-15%, etc. of total exam time);
- *Quarter Time:* All actions collected are aggregated into four different intervals, each containing 25 percent of the total duration of the exam spent by the individual (e.g. 0-25%, 25-50%, 50-75%, 75-100% of total exam time);
- *Sliding Time:* All actions collected are aggregated into intervals of five minutes. The main difference between *Chronological Time* and *Sliding Time* is the transition of time between the different groups (e.g. action events from minute 0-5, 1-6, 2-7, etc.);

– *Complete Time:* The data collected is not divided in intervals, presenting the processed features set of the complete exam.

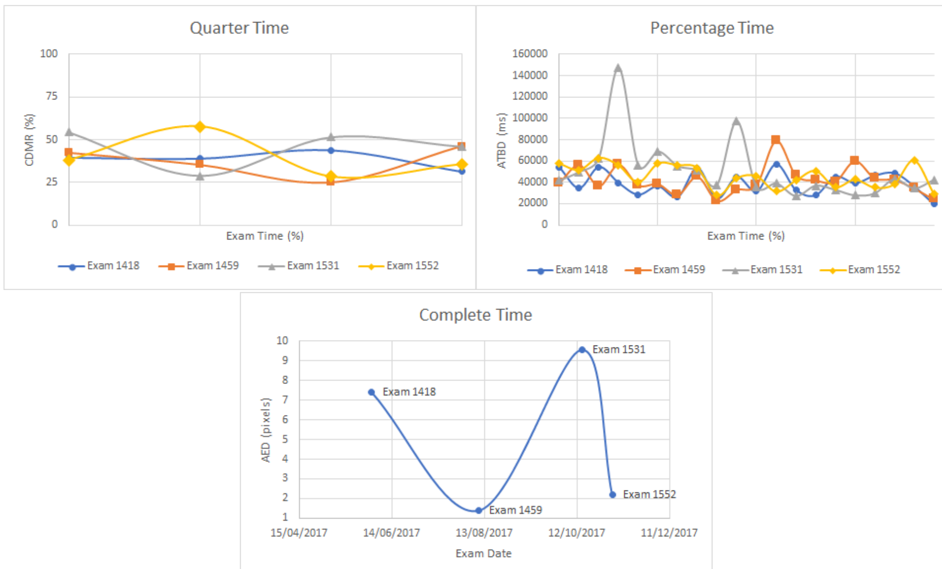

Fig. 3: Different temporal approaches addressed in X3S framework, presenting the CDMR in the Quarter Time approach (top left subfigure), the ATBD in the Percentage Time approach (top right subfigure) and the AED in the Complete Time approach (bottom subfigure) of one medical student during four different exams.

Fig. 3 shows some examples of the *Quarter*, *Percentage* and *Complete* temporal approaches, through the analysis of one medical student's variation features during four different exams. This preparation process is used as a way of presenting different approaches for our case study, and consequently to take advantage of the different conclusions for future data mining.

5.2. Data Uncertainty Management & Dimension Reduction

The second step in the preparation of the data is the implementation of data transformation processes that can provide additional insights. Moreover, when dealing with real-world data, it is often necessary to deal with missing/ambiguous information and to reduce its dimensionality to improve big data management, with minimum users' involvement.

Missing data is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data [27]. Also, sensor measurements inherently incorporate varying degrees of uncertainty and are, occasionally, spurious and incorrect, presenting ambiguous information into the data. In order to solve this problem, several data uncertainty management techniques are available [6, 2].

One of the most used techniques is the conditional mean imputation method. The objective of missing data imputation is to estimate the missing part of the data given the observed part, exploiting the statistical relationship between the two [20]. In other words, the process is accomplished by first removing ambiguous feature values from the data, followed by regressing the respective variable with missing data (restricted from cases of students during the same exam). The estimated regression equation is then used to generate predicted values for the cases with missing data.

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality [12]. Ideally, the reduced representation should have a dimensionality that corresponds to the intrinsic dimensionality of the data. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [22]. As a result, dimensionality reduction facilitates, among others, classification, visualisation, prediction, and compression of high-dimensional data.

Moreover, the data collected from Mouse Dynamics comprises a significantly large amount. To reduce its dimensionality and given the usual shape of the data series (exemplified in Fig. 4), we construct a linear fit and use the resulting quadratic function to represent the raw data. The model build of the linear and the quadratic model are represented as $f(x) = \alpha + \beta x$ for the linear model and as $g(x) = \alpha + \beta x + \gamma x^2$ for the quadratic model.

After calculating the coefficients of both models, a mean squared error (MSE) calculation is performed. This mathematical formula measures the quality of an estimator, that is, the difference between the estimator and what is estimated. In other words, values closer to zero are better. By comparing the MSE of both functions, we can choose which model is more accurate for the set of values. The MSE of the predictor can be estimated by:

$$MSE = \tfrac{1}{n}\sum_{i=1}^{n}(\hat{Y_i} - Y_i)^2$$

in which $\hat{Y_i}$ is a vector of n predictions, and Y is the vector of observed values corresponding to the inputs to the function which generated the predictions.

Fig. 4 partially depicts the outcome of this process. In this figure, the dots represent the aggregated raw data at regular intervals while the lines represent the resulting quadratic model. After this reduction process, it is possible to use the parameters of the quadratic function instead of the raw data, which simplifies the posterior use of machine learning techniques.

6. Data Analysis

An analysis was conducted to compare the data collected in medical school exams over the three first years. To achieve this, the dataset was divided in three groups, each containing information collected during 1st, 2nd and 3rd-year medical exams.

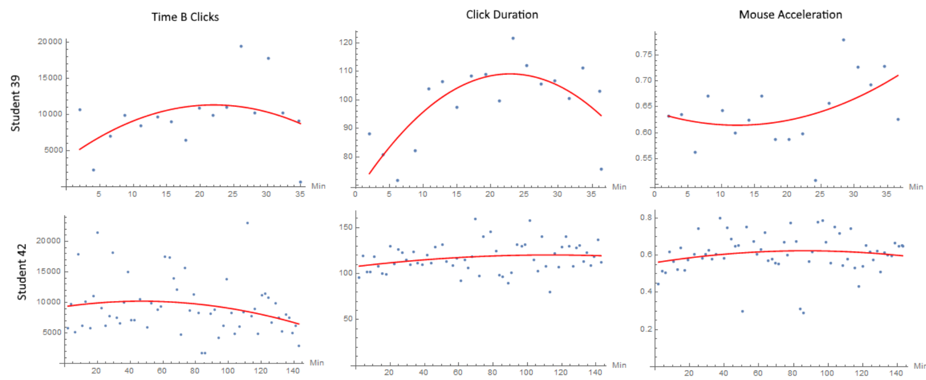

Fig. 4: Example of dimensionality reduction for three features in two students.

After this, an analysis was done to measure the variability of each performance feature (Table 1). In this study, a pearson's correlation method and random forest model were used to provide an evaluation of the feature's importance. Pearson's correlation method measures linear correlation between two variables, where the resulting value lies between [-1;1], with -1 meaning perfect negative correlation (as one variable increases, the other decreases), +1 meaning perfect positive correlation and 0 meaning no linear correlation between the two variables [19]. As for random forests, this model are among the most popular machine learning methods thanks to their relatively good accuracy, robustness and ease of use. Additionally, they also provide two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy [19]. The following features were selected based on their relevance (in both methods) to determine an individual's stress: mouse velocity (MV), absolute sum of degrees (ASD), average excess of distance between clicks (AED), average distance of the mouse to the straight line (ADMSL), decision making ratio (DMR), correct decision making ratio (CDMR), median time between decision (MTBD) and average time between decision (ATBD). Furthermore, the biological markers of stress reaction (PreCortisol, PosCortisol and DeltaCortisol) and student's perceived stress (PSS score) were added to this analysis.

The features depicted in Table 1 are intrinsically related to student performance. Taking as example the movement of the mouse, one never moves it in a straight line between two points, there is always some degree of curve. The larger the curve, the less efficient the movement is. An interesting property of these features is that, except for mouse velocity and acceleration (for which the relationship is not so clear), an increasing value denotes a decreasing performance. Also, longer clicks and larger ADMSL are associated to poorer performance, which shows that the user is in a relaxed state.

It is interesting to note how the values of the features vary from year to year, especially when comparing the 1st and the 3rd years: in all of them there is a tendency to increase.

Concerning the analysis of the students' cortisol levels over the three years (more specifically the PreCortisol and PosCortisol), the average values rise from 0.3717 and 0.1558 to 0.4394 and 1.1827, respectively. This indicates that students in the third year have an overall higher level of cortisol. Despite this increase, the students' perception of

Table 1: Comparison of the mean, median and variance for each feature and each of the first three years of medical school, analysed according to the *Complete Time* approach. The dataset contains data from 102 students in the 1st-year, 87 students in the 2nd-year and 81 students in the 3rd-year.

Feature	School Year	Variance	Median	Mean
MV	1st	0,0189 px/s	0,7288 px/s	0,7316 px/s
	2nd	0,0274 px/s	0,7350 px/s	0,7497 px/s
	3rd	0,0296 px/s	0,7679 px/s	0,8003 px/s
ASD	1st	54898486 ^{o}	11028 ^{o}	12151 ^{o}
	2nd	20235912 ^{o}	9418 ^{o}	10424 ^{o}
	3rd	49429299 ^{o}	11376 ^{o}	13334 ^{o}
AED	1st	1,8638 px	7,5614 px	7,4055 px
	2nd	1,3693 px	7,5614 px	7,3115 px
	3rd	9,7929 px	7,5614 px	8,0036 px
ADMSL	1st	163,3815 px	92,9642 px	90,8413 px
	2nd	158,9729 px	92,9642 px	90,8770 px
	3rd	142,4232 px	92,9642 px	94,3740 px
DMR	1st	0,55 %	31,72 %	31,69 %
	2nd	0,42 %	32,97 %	34,03 %
	3rd	0,77 %	32,99 %	32,83 %
CDMR	1st	1,02%	51,63%	51,54%
	2nd	1,21%	50,78%	51,68%
	3rd	0,49%	61,29%	61,11%
MTBD	1st	13441980 ms	7000 ms	7411 ms
	2nd	14207011 ms	6000 ms	7087 ms
	3rd	22547068 ms	9000 ms	9377 ms
ATBD	1st	34899381 ms	19212 ms	19460 ms
	2nd	37464048 ms	20456 ms	20926 ms
	3rd	37584397 ms	20940 ms	21113 ms
PreCortisol	1st	0,0871 nmol/L	0,2845 nmol/L	0,3717 nmol/L
	2nd	0,0327 nmol/L	0,4167 nmol/L	0,3954 nmol/L
	3rd	0,0623 nmol/L	0,3820 nmol/L	0,4394 nmol/L
PosCortisol	1st	0,0037 nmol/L	0,1530 nmol/L	0,1558 nmol/L
	2nd	0,0056 nmol/L	0,1739 nmol/L	0,1736 nmol/L
	3rd	0,0121 nmol/L	0,1739 nmol/L	0,1827 nmol/L
DeltaCortisol	1st	0,0796 nmol/L	0,1905 nmol/L	0,2303 nmol/L
	2nd	0,0218 nmol/L	0,2514 nmol/L	0,2616 nmol/L
	3rd	0,0565 nmol/L	0,2514 nmol/L	0,2393 nmol/L
PSS Score	1st	44,16	27	27,46
	2nd	47,24	23	23,41
	3rd	58,15	24	23,66

stress seem to decrease, where the average values decreases from 27.76 to 23.66. However, PosCortisol and stress perception score variance tends to increase each year, while PreCortisol varies widely.

A similar trend is observed for all the interaction features, which indicates a decrease in performance over the years. The joint analysis of these two groups of variables puts forward some interesting hypotheses that will be tested in the future. Namely, the students' level of stress increases as they progress in their course; this is accompanied by a drop in performance. Moreover, this shows that these two groups of variables are potentially related and that it is possible that one can predict the other.

However, and despite the observed differences, these are general conclusions and not every student is expected to behave the same. While this may reveal the overall behaviour, we are aware of the importance of developing individual models, trained with data from each user, that may be used to more accurately identify those students who have poorer stress coping strategies.

The main conclusions of this data is, nonetheless, that when the students exhibit higher levels of stress and are probably closer to a state of burnout, their performance decreases. This is evidenced by less efficient interaction patterns (e.g. longer mouse clicks, larger distances travelled by the mouse, longer key down times, faster and less efficient decisions made, etc.). This conclusion, which comes as no surprise, is nonetheless important in the sense that it allows, for the first time, to quantify and study this relationship between an objective stress measure (i.e. cortisol and PSS score) and the features selected.

As analysed previously, there are some potential drawbacks associated to this kind of performance-based approaches, namely the difficulty in precisely accounting for changes in performance (e.g. they may not be entirely due to mental stress). Moreover, measuring performance is not always easy. Different students present different interaction patterns which may significantly influence the prediction of their stress state. The workload is another issue to consider.

To address these issues, it is our intention to work towards a classifier of behavioural pattern (e.g. mouse and decision-making behaviours) to be used to classify each student, identifying workload and quantifying the level of stress during high-end tasks. This kind of information, which to some extent describes the user's context, will enable the development of more accurate classifiers.

7. Conclusions and Future Work

In this paper we present a technological approach for a non-intrusive analysis of performance in groups of people. This approach is implemented in the form of a distributed system, that constantly collects, processes, stores and monitors data describing the behaviour of multiple individuals simultaneously, during the execution of on-line high-stakes exams in real-time. Through the metrics monitored during the execution of a set of exams, the platform will be able to correlate data between the mouse performance metrics (using MouseDynamics output) and decision making performance metrics (analysing MedQuizz features), in order to quantify the stress levels of an individual.

The main conclusion, as expected, is that when the users are in a state of burnout, their performance decreases. This is evidenced by less efficient interaction patterns (e.g. longer mouse clicks, larger distances travelled by the mouse, longer key down times, faster and less efficient decisions made, etc.). This conclusion is nonetheless important in the sense that it allows, for the first time, to quantify and study these differences.

As analysed previously, there are some potential drawbacks associated to this kind of performance-based approaches, namely the difficulty in precisely accounting for changes in performance (e.g. they may not be entirely due to mental stress). Moreover, measuring performance is not always easy. Different students present different interaction patterns which may significantly influence the prediction of their stress state. The workload is another issue to consider.

By monitoring human behaviour, our aim is to study the effect of stress in the performance of high demand tasks and to point out how each individual is affected by stress. This will allow the educational institution to act on each student, through personalised teaching and coping strategies, and thus improve the quality of the future professionals that are being trained. This approach could also be used realistically in a common workplace environment, especially in workplaces where individuals spend long hours interacting with a computer.

The current approach still presents some limitations. Specifically, the proposed solution is currently high dependent on MedQuizz tool, where the set of services provided are for the time being limited to this framework. However, this technological approach was designed aiming to simplify the application of new behaviour analysis modules (e.g. keystroke, ambient luminosity, noise, etc.) and to be used, not only in learning/exam setting, but also in decision-making tasks of other domains (e.g. Psychology Assessment, Science Quizzes, User's Feedback, etc.).

As future work, it is planned to apply data mining algorithms to analyse data from different perspectives and summarise it into useful information, as a mean to find correlations or patterns among dozens of features in our databases. Yet, it is required to define which of the mining methods and fields can influence positively the results to quantify the stress levels of an individual, during the execution of an exam. This kind of information, which to some extent describes the user's pattern interaction, will enable the development of more accurate classifiers for non-intrusive user's stress monitoring.

Acknowledgements This work is part-funded by ERDF–European Regional Development Fund and by National Funds through the FCT–Portuguese Foundation for Science and Technology within project NORTE-01-0247-FEDER-017832. The work of Filipe Gonçalves is supported by a FCT grant with the reference ICVS-BI-2016-005.

References

1. Abouserie, R.: Sources and levels of stress in relation to locus of control and self esteem in university students. Educational psychology 14(3), 323–330 (1994)
2. Allison, P.D.: Missing data: Quantitative applications in the social sciences. British Journal of Mathematical and Statistical Psychology 55(1), 193–196 (2002)
3. Attaran, N., Brooks, J., Mohsenin, T.: A low-power multi-physiological monitoring processor for stress detection. In: SENSORS, 2016 IEEE. pp. 1–3. IEEE (2016)
4. Baharum, A., Wee, L.X., Tanalol, S.H., Hanapi, R.: Stress monitoring with mobile application: Stress catcher 2.0. Annals of advanced sciences 1(2) (2017)
5. Bailey, K.O., Okolica, J.S., Peterson, G.L.: User identification and authentication using multimodal behavioral biometrics. Computers & Security 43, 77–89 (2014)
6. Bobek, S., Nalepa, G.J.: Uncertain context data management in dynamic mobile environments. Future Generation Computer Systems 66, 110–124 (2017)
7. Brunet, P.M., Cowie, R.: Towards a conceptual framework of research on social signal processing. Journal on Multimodal User Interfaces 6(3-4), 101–115 (2012)

8. Carneiro, D., Novais, P.: Quantifying the effects of external factors on individual performance. Future Generation Computer Systems 66, 171–186 (2017)
9. Carneiro, D., Novais, P., Pêgo, J.M., Sousa, N., Neves, J.: Using mouse dynamics to assess stress during online exams. In: International Conference on Hybrid Artificial Intelligence Systems. pp. 345–356. Springer (2015)
10. Cartwright, S., Cooper, C.L.: Managing workplace stress, vol. 1. Sage (1997)
11. Carveth, J.A., Gesse, T., Moss, N.: Survival strategies for nurse-midwifery students. Journal of Nurse-Midwifery 41(1), 50–54 (1996)
12. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: From big data to big impact. MIS quarterly 36(4), 1165–1188 (2012)
13. Cohen, M., Khalaila, R.: Saliva ph as a biomarker of exam stress and a predictor of exam performance. Journal of psychosomatic research 77(5), 420–425 (2014)
14. Cohen, S., Kamarck, T., Mermelstein, R.: A global measure of perceived stress. Journal of Health and Social Behavior 24(4), 385–396 (1983)
15. Cohen, S., Kamarck, T., Mermelstein, R., Others: Perceived stress scale. Measuring stress: A guide for health and social scientists (1994)
16. Cohen, S., Kessler, R.C., Gordon, L.U.: Measuring stress: A guide for health and social scientists. Oxford University Press on Demand (1997)
17. Colligan, T.W., Higgins, E.M.: Workplace stress: Etiology and consequences. Journal of workplace behavioral health 21(2), 89–97 (2006)
18. Dai, W., Duch, W., Abdullah, A.H., Xu, D., Chen, Y.S.: Recent advances in learning theory. Computational intelligence and neuroscience 2015, 14 (2015)
19. Egghe, L., Leydesdorff, L.: The relation between Pearson's correlation coefficient r and Salton's cosine measure. Journal of the Association for Information Science and Technology 60(5), 1027–1036 (2009)
20. Faubel, F., McDonough, J., Klakow, D.: Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. pp. 3869–3872. IEEE (2009)
21. Fink, G.: Stress: definition and history. Stress science: neuroendocrinology pp. 3–9 (2010)
22. Fukunaga, K.: Introduction to statistical pattern recognition. Academic press (2013)
23. Gallese, V.: Intentional attunement: A neurophysiological perspective on social cognition and its disruption in autism. Brain research 1079(1), 15–24 (2006)
24. Gonçalves, F., Carneiro, D., Novais, P., Pêgo, J.: Eustress: A human behaviour analysis system for monitoring and assessing stress during exams. In: International Symposium on Intelligent and Distributed Computing. pp. 137–147. Springer (2017)
25. Hellhammer, D.H., Wüst, S., Kudielka, B.M.: Salivary cortisol as a biomarker in stress research. Psychoneuroendocrinology 34(2), 163–171 (2009)
26. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology 14(1), 4–20 (2004)
27. Kang, H.: The prevention and handling of the missing data. Korean journal of anesthesiology 64(5), 402–406 (2013)
28. Khan, M., Rizvi, Z., Shaikh, M.Z., Kazmi, W., Shaikh, A.: Design and implementation of intelligent human stress monitoring system. International Journal of Innovation and Scientific Research, ISSN pp. 2351–8014 (2014)
29. Kim, S., Filippone, M., Valente, F., Vinciarelli, A.: Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes. In: Proceedings of the 20th ACM international conference on Multimedia. pp. 793–796. ACM (2012)
30. Kocielnik, R., Sidorova, N.: Personalized stress management: enabling stress monitoring with lifelogexplorer. KI-Künstliche Intelligenz 29(2), 115–122 (2015)

31. Le Fevre, M., Matheny, J., Kolt, G.S.: Eustress, distress, and interpretation in occupational stress. Journal of Managerial psychology 18(7), 726–744 (2003)
32. McDuff, D., Gontarek, S., Picard, R.: Remote measurement of cognitive stress via heart rate variability. In: Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE. pp. 2957–2960. IEEE (2014)
33. Novais, P., Carneiro, D.: The role of non-intrusive approaches in the development of people-aware systems. Progress in Artificial Intelligence 5(3), 215–220 (2016), http://dx.doi.org/10.1007/s13748-016-0085-1
34. O'Sullivan, G.: The relationship between hope, eustress, self-efficacy, and life satisfaction among undergraduates. Social indicators research 101(1), 155–172 (2011)
35. Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroeder, M., Vinciarelli, A.: Social signal processing: the research agenda. In: Visual analysis of humans, pp. 511–538. Springer (2011)
36. Ross, S.E., Niebling, B.C., Heckert, T.M.: Sources of stress among college students. Social psychology 61(5), 841–846 (1999)
37. Saaty, T.L.: Decision making with the analytic hierarchy process. International journal of services sciences 1(1), 83–98 (2008)
38. Sanders, A.: Towards a model of stress and human performance. Acta psychologica 53(1), 61–97 (1983)
39. Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., Mozos, O.M.: Stress detection using wearable physiological sensors. In: International Work-Conference on the Interplay Between Natural and Artificial Computation. pp. 526–532. Springer (2015)
40. Smith, M.J., Conway, F.T., Karsh, B.T.: Occupational stress in human computer interaction. Industrial health 37(2), 157–173 (1999)
41. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and vision computing 27(12), 1743–1759 (2009)
42. Wang, L.: Behavioral Biometrics for Human Identification: Intelligent Applications: Intelligent Applications. IGI Global (2009)
43. Xu, H., Hua, K., Wang, W., Lu, M., Jiang, T.: Energy efficient ecg monitoring system for human emotional stress assessment. Computer Science and Engineering 5(1A), 8–14 (2015)
44. Zhou, X., Dai, G., Huang, S., Sun, X., Hu, F., Hu, H., Ivanović, M.: Cyberpsychological computation on social community of ubiquitous learning. Computational intelligence and neuroscience 2015, 12 (2015)