Predicting completion time in high-stakes exams

Davide Carneiro[a,d,*], Paulo Novais[d], Dalila Durães[e], José Miguel Pego[b,c], Nuno Sousa[b,c]

[a] *CIICESI, ESTG, Polytechnic Institute of Porto, Portugal*
[b] *Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal*
[c] *ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal*
[d] *Algoritmi Center/Department of Informatics, University of Minho, Braga, Portugal*
[e] *Department of Artificial Intelligence, Technical University of Madrid, Spain*

Abstract

For the majority of students, assessment moments are associated with significant levels of stress and anxiety. While a certain amount of stress motivates the individual and improves performance, too much stress will have the contrary effect. Stress has therefore a fundamental role on student performance. It should be the educational organizations' mission to understand the underlying mechanisms that lead to performance anxiety and provide their students with the best coping tools and strategies. In the present study we analyze student behavior during e-assessment in terms of mouse dynamics. Two major behavioral patterns can be identified, based on ten features that quantify the performance of the student's interaction with the computer: (1) students who are able to sustain performance during the exam and (2) students whose performance varies significantly. Data shows that the behavior of each student during the exam correlates strongly with the time it takes the student to complete it. Several classifiers were trained that predict the completion time of each exam based on the students' interaction patterns. Two of them do it with an average error of around twelve minutes. Results show that there are still mechanisms that can be explored to better understand the complex relationship between stress, performance and human behavior, that

*Corresponding Author
Email addresses: dcarneiro@di.uminho.pt (Davide Carneiro), pjon@di.uminho.pt (Paulo Novais), d.alves@alumnos.upm.es (Dalila Durães), jmpego@ecsaude.uminho.pt (José Miguel Pego), njcsousa@ecsaude.uminho.pt (Nuno Sousa)

can be used for the implementation of better stress detection, monitoring and coping strategies.

Keywords: Online Exams, Stress, Random Decision Forests, Neural Networks

1. Introduction

Higher education is often stressful, especially in courses considered more challenging or demanding such as Medicine in which student stress levels are higher than those of other fields [1]. Stressors such as the overwelming burden of information, the uncertainty about the future or dealing with high expectations put a constant pressure on the student [2].

Each student tends to develop his/her own coping strategies, some counterproductive (e.g. drugs, alcohol, eating) others constructive (e.g. regular exercise, meditation). Indeed, coping styles are related not only to student performance but also to mental health outcomes, especially in high-achieving students. Furthermore, coping styles (specifically, anger and positive appraisal) moderate the influence of stress on global life satisfaction and internalizing symptoms of psychopathology [3].

The educational organization plays an essential role in this matter: that of providing the student community with viable solutions for coping with stress (e.g. guidance, personalized support). However, and despite the existence of successful stress coping initiatives [4], it is still necessary to determine which students need this support as they do not always come forward on their own initiative.

The identification of these students may not always be easy as people often tend to disguise the consequences of stress. Moreover the professor, who might be in the best position to do so, is often dealing with hundreds of students, making this a rather difficult task to achieve.

In this paper we propose an approach to assess the effects of stress on students during high-stakes exams, which constitute particular stressful moments in the academic career. Specifically, we look at behavioral biometrics [5] and mouse dynamics [6] to develop a non-intrusive method to assess the performance of students while interacting with the computer.

The goal is to analyze how performance varies during high-stakes exams for each individual student, eventually distinguishing between students who are able to maintain a steady performance and students whose performance

varies significantly during the duration of the exam. The hypothesis is that each student may be affected differently by stress and that this approach is a valid one for measuring such differences. Indeed, and as detailed in Section 2, stress affects many different aspects of an individual, including physical and behavioral. It is therefore valid to accept that it might affect the way a student interacts with the mouse during a stressful event such as an exam.

In this paper we detail the features extracted from the student's interaction with the computer, which characterize the student's performance. These data are collected and processed in real-time, from each student, allowing to continuously assess the performance of each student throughout the exam. Moreover, results show that performance is related to exam completion time. With the collected data we train four different classifiers, two of which able to predict exam completion time for each student, based on the interaction patterns with an error of approximately 12 minutes.

The rest of the paper is organized as follows. Section 2 briefly addresses stress and its effects on the individual at different levels. Section 3 details the population that participated in the study, the feature extraction process and the dataset used. Afterwards, Section 4 details the methodology followed for analyzing the data and training the classifier, and the main results. Finally, Section 5 discusses the significance of the main findings and presents the conclusions.

2. Stress and its Effects

In nowadays society, stress and its effects at an organizational and individual level have become a significant problem [7]. While stress has undeniable positive effects [8], it is the negative effects that become especially known to society, due to the potential severity of their consequences.

Stress and related concepts can be traced as far back as written science and medicine [9]. In modern science, the study of the physiological effects of stress in the 50's, resulted in the development of a group of reliable physiological indicators, that can be easily acquired through sensors (e.g. skin conductivity, body temperature, heart rate) to assess stress level. Later, in the 70's, research was directed towards the somatic disorders that result from biologic aspects [10].

Simultaneously, Hans Selye, an Austrian-Canadian endocrinologist, provided an accurate and simultaneously accessible definition of stress [11] as a non-specific response of the body to external demands. These demands (the

load or stimulus that triggered a response) are denominated stressors while the internal body changes that they produce constitute the actual stress response. Selye was also the first to document the chemical and hormonal changes that occur in the body due to stress.

Responses to stress are coordinated by a so-called *stress system* and take place when a stressor of any kind exceeds a given threshold. Given that this threshold may vary from person to person (as it depends on individual differences) the study of stress and its causes/effects becomes complex. The composition of this stress system is known to include as main components the corticotropin-releasing hormone and locus ceruleus-norepinephrine/autonomic systems and their peripheral effectors, the pituitary-adrenal axis, and the limbs of the autonomic system [9].

The effects of stress on the nervous system of the individual at an internal level have thus been studied for several decades. In synthesis, the activation of the stress system leads to peripheral changes that improve the ability of the organism to adjust homeostasis and increase its chances for survival.

However, in recent years, attention has been drawn to the effects of stress at an external level. Indeed, when observed externally, stress results in many changes on the individual, namely on behavior, physical response or cognitive performance (e.g. concentration, short-term memory, fine motor control, reasoning), through different physiological mechanisms [12].

This led to the development of non-intrusive systems for stress detection that are based on the individual's behavior. Many different approaches have been proposed. In [13] the authors explore the possibility of detecting cognitive and physical stress by monitoring keyboard interactions with the eventual goal of detecting acute or gradual changes in cognitive and physical function. The researchers analyze keystroke and linguistic features of spontaneously generated text, showing that it is possible to classify cognitive and physical stress conditions relative to non-stress conditions.

Similar approaches exist that use different modalities of our behavior to classify stress, including mouse dynamics, keyboard dynamics and movement patterns [14]. Finally, There is also work on the effect of stress on the interaction with touch screens in mobile devices, showing that it is possible to distinguish between two states of stress from temporal and intensity features of touches [15].

An up-to-date view of stress may thus look at it as a physic-physiologic arousal response occurring in the body as result of stimuli. It should also be added that these stimuli only become stressors by virtue of the cognitive

interpretation of the individual, i.e., the effects of stressors depend on the individual. Regardless of inter-individual differences, it is nowadays clear that the effects of stress can be measured in novel ways, namely those based on Human-Computer Interaction.

Over the last years there has also been a growing interest in studying and measuring stress and its effects in the workplace and similar milieus. In the beginning, interest was mostly from the field of occupational health, i.e., studying the effects of people working with new technological tools and the accompanying new stressors (e.g. inadequate employee training to use new technology, monotonous tasks, electronic performance monitoring) [16]. The interest was thus more on identifying and understanding technology-related job stressors rather than on trying to quantify stress level.

However, the interest of researchers soon shifted to the field of Human-Computer Interaction and to the effects of stress, affective states, fatigue and other factors on the individual's interaction patterns with the technological devices. As when considering interactions between people, interactions between people and technological devices also have two channels: one transmits explicit messages (i.e. the actions we perform on the computer) while the other transmits implicit messages (i.e. how we do it). As research has been demonstrating, we perform actions differently according to our state. The inclusion of this kind of information in next-generation Human-Computer Interaction designs is seen by many experts in the field as the path to produce true human-aware and systems, that are able to understand and adapt to the user's state at each moment [17].

Many different approaches have been followed in this regard, with varying goals. In [18], the authors propose a system based on face analysis and voice recognition to analyze the emotions of computer users. These two methods are actually very common in this field [19], either for stress analysis or for fatigue or emotion classification, along with posture and gaze analysis [20]. Other sensors have also been used by researchers, including Blood Volume Pulse, Galvanic Skin Response, Skin or Facial Temperature and Pupil Diameter, all of which producing features that are strongly correlated to stress level or emotional arousal [21, 22].

The main drawback with these approaches, that are generally very precise and are also nowadays very common, is that they rely on sensors that may be intrusive since they must be placed on the body of the individuals. Few research works exist that acquire relevant data in a non-intrusive way. One of the very few is the work of [23], in which the authors use a pressure sensitive

keyboard and a capacitive mouse to measure user stress level.

In the case of this work, one of the fundamental requirements is that the routines of the students while doing the computer-based exam are not disturbed by the data collection process as, given the potential stressful nature of the exam, any disturbance may have negative consequences on the outcome. Moreover, the use of additional hardware might significantly increase the cost of the data collection process, leading to a probable decrease in the population to control costs. The approach followed in this paper thus relies solely on Mouse Dynamics, in an attempt to put forward a completely non-intrusive method for assessing stress in Human-Computer Interaction.

3. Material and Methods

The purpose of this work is to assess the influence of stress on student behavior during high-stakes exams in the sense that these are exams with important life-changing consequences for the test taker. In the case of the selected population (medical students), passing has significant benefits (e.g. advancing on their academic career or attaining a diploma) while failing has important disadvantages (e.g. delaying or jeopardizing the student's aspirations). This specific population (medical students) was selected to participate in this study given these characteristics as they undergo some of the most rigorous and intense (and consequently stressful) learning and assessment processes there are.

The study took place in the School of Medicine of the University of Minho, where exams take place at the computer. In this kind of exams, when students enter the room, they are indicated their computer. Each computer has a keyboard, a mouse and a screen. At the designated time they log in the exam platform using their personal credentials and the exam begins. During the exam, which consists mostly of single-best-answer multiple choice questions [24], students use mostly the mouse as an interaction means. When the exam ends, students are allowed to leave the room.

The collection of the necessary data is completely transparent from the point of view of the student, i.e., their participation in the study has no effect whatsoever on their routine as all the relevant data is collected in a transparent manner. The population that participated in the study as well as the characteristics of the four exams in which data was collected is detailed in Section 3.1. Section 3.2 details the interaction features that are transparently

Table 1: Summary of the characteristics of each exam.

Exam	Date	Duration (min)			#*students*	#*participants*
		$\bar{x}$	$\tilde{x}$	S		
A	16/10/2015	74.59	47.54	53.28	104	62
B	18/02/2016	102.36	107.87	48.48	107	90
C	19/02/2016	90.80	83.14	49.00	17	17
D	28/10/2015	107.12	120.75	28.83	135	128

extracted from the use of the mouse by the students. Finally, Section 3.3 describes the methodology followed as well as the dataset used.

3.1. Population

Data was collected in four different exams, that took place in different dates, comprising a total of 363 students. Of these, data from 66 students that left the exam in the initial 15 minutes, who were not actually trying to complete the exam, were discarded. The dataset thus includes data from 297 participants out of the initial 363 students (81.82%). Table 1 details the main characteristics of each of the four exams. Figure 1 complements this information graphically by depicting the distribution of the completion time of each exam.

It is important to clarify that in this work, the duration of the exam depicts the completion time, i.e., the time it takes each student to complete the exam as there is no fixed duration for each exam: each student is free to manage her/his time at will.

3.2. Feature Extraction

All the actions of each student during the exam are logged electronically (e.g. moving to a new question, answering a question, changing a previous answer). Moreover, lower-level data is also recorded that describes the interaction of the students with the computer peripherals. Specifically, the following events are recorded:

- MOV, timestamp, posX, posY

  An event describing the movement of the mouse, in a given time, to coordinates (posX, posY) in the screen;

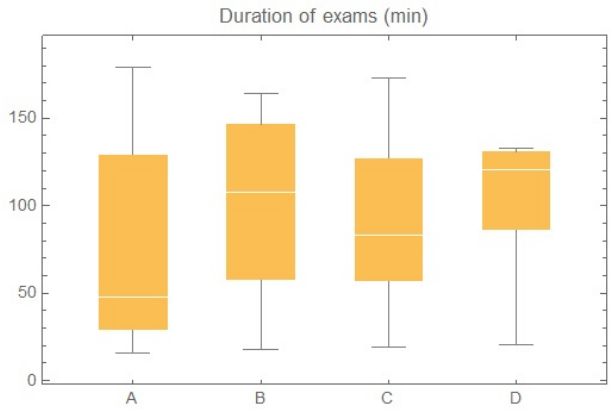

Figure 1: Exam completion time: distribution of the duration of each exam, in minutes, before removing data from exams shorter than 15 minutes.

- MOUSE_DOWN, timestamp, [Left—Right], posX, posY

  This event describes the first half of a click (when the mouse button is pressed down), in a given moment. It also describes which of the buttons was pressed (left or right) and the position of the mouse in that instant;

- MOUSE_UP, 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;

- MOUSE_WHEEL, timestamp, dif

  This event describes a mouse wheel scroll of amount dif, at a given time;

- KEY_DOWN, timestamp, key

  Identifies a given key from the keyboard being pressed down, at a given time;

- KEY_UP, timestamp, key

  Describes the release of a given key from the keyboard, at a given time.

The following example depicts a brief log that starts with some mouse movement (first two lines), contains a click with a little drag (lines 3-5) and ends with some more movement (last two lines).

```
MOV, 635296941683402953, 451, 195
MOV, 635296941684123025, 451, 197
MOUSE_DOWN, 635296941684443057, Left, 451, 199
MOV, 635296941685273140, 452, 200
MOUSE_UP, 635296941685283141, Left, 452, 200
MOV, 635296941685723185, 452, 203
MOV, 635296941685803193, 454, 205
```

On the one hand, this kind of logs allows to fully reconstruct the actions of each student, providing insights to when and were interactions occur. On the other hand, this information can be used to extract features that quantify the performance of the student's interaction with the computer. The process of extracting these features is detailed in [25].

Given the characteristics of the exams, this study only considers the 10 features that are extracted from the interaction with the mouse:

Absolute Sum of Angles (ASA)
UNITS - degrees
This feature seeks to find how much the mouse "turned", independently of the direction to which it turned (Figure 2 (a)). In that sense, it is computed as the absolute of the value returned by function $degree(x1, y1, x2, y2, x3, y3)$, as depicted in equation 1.

$$rCls\_angle = \sum_{i=0}^{n-2} \mid degree(posx_i, posy_i, posx_{i+1}, posy_{i+1}, posx_{i+2}, posy_{i+2}) \mid \tag{1}$$

Average Distance of the Mouse to the Straight Line (ADMSL)
UNITS - pixels
This feature measures the average distance of the mouse to the straight line defined between two consecutive clicks. Let us assume two consecutive MOUSE_UP and MOUSE_DOWN events, *mup* and *mdo*, respectively in the coordinates $(x1, y1)$ and $(x2, y2)$. Let us also assume two vectors *posx* and *posy*, of size n, holding the coordinates of the consecutive MOUSE_MOV events between *mup* and *mdo*. The sum of the distances between each position and the straight line defined by the points $(x1, y1)$ and $(x2, y2)$ is given by 2, in which *ptLineDist* returns the distance between the specified point

and the closest point on the infinitely-extended line defined by $(x1, y1)$ and $(x2, y2)$. The average distance of the mouse to the straight (Figure 2 (b)) line defined by two consecutive clicks is thus given by s_dists/n.

$$s\_dists = \sum_{i=0}^{n-1} ptLineDist(posx_i, posy_i) \tag{2}$$

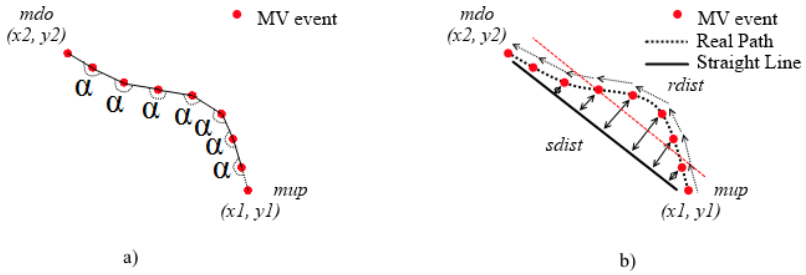

Figure 2: (a) The sum of the angles of the mouse's movement is given by summing all the angles between each two consecutive movement vectors. (b) The average distance at which the mouse is from the shortest line between two clicks is depicted by the straight dashed line.

Average Excess of Distance (AED)

UNITS - pixels

This feature measures the average excess of distance that the mouse travelled between each two consecutive MOUSE_UP and MOUSE_DOWN events. Let us assume two consecutive MOUSE_UP and MOUSE_DOWN events, *mup* and *mdo*, respectively in the coordinates $(x1, y1)$ and $(x2, y2)$. To compute this feature, first it is measured the distance in straight line between the coordinates of *mup* and *mdo* as $s\_dist = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$. Then, it is measured the distance actually travelled by the mouse by summing the distance between each two consecutive MOUSE_MV events. Let us assume two vectors *posx* and *posy*, of size n, holding the coordinates of the consecutive MOUSE_MV events between *mup* and *mdo*. The distance actually travelled by the mouse, $real\_dist$ is given by equation 3. The average excess of distance between the two consecutive clicks (Figure 3 (a)) is thus given by r_dist/s_dist.

Click Duration (CD)
UNITS - milliseconds
Measures the timespan between two consecutive MOUSE_UP and MOUSE_DOWN events.

Distance Between Clicks (DBC)
UNITS - pixels
Represents the total distance travelled by the mouse between two consecutive clicks, i.e., between each two consecutive MOUSE_UP and MOUSE_DOWN events. Let us assume two consecutive MOUSE_UP and MOUSE_DOWN events, *mup* and *mdo*, respectively in the coordinates $(x1, y1)$ and $(x2, y2)$. Let us also assume two vectors *posx* and *posy*, of size n, holding the coordinates of the consecutive MOUSE_MOV events between *mup* and *mdo*. The total distance travelled by the mouse is given by equation 3.

$$r\_dist = \sum_{i=0}^{n-1} \sqrt{(posx_{i+1} - posx_i)^2 + (posy_{i+1} - posy_i)^2} \tag{3}$$

Distance of the Mouse to the Straight Line (DMSL)
UNITS - pixels
This feature is similar to the previous one in the sense that it will compute the *s_dists* between two consecutive MOUSE_UP and MOUSE_DOWN events, *mup* and *mdo*, according to equation 2. However, it returns this sum rather than the average value during the path.

Excess of Distance (ED)
UNITS - pixels
This feature measures the excess of distance that the mouse travelled between each two consecutive MOUSE_UP and MOUSE_DOWN events. *r_dist* and *s_dist* are computed as for the AED feature. However, ED is given by $r\_dist - s\_dist$

Mouse Acceleration (MA)
UNITS - pixels/milliseconds^2
The velocity of the mouse (in pixels/milliseconds) over the time (in milliseconds). A value of acceleration is computed for each interval defined by two consecutive MOUSE_UP and MOUSE_DOWN events, using the intervals and data computed for the Velocity.

Mouse Velocity (MV)
UNITS - pixels/milliseconds
The distance travelled by the mouse (in pixels) over the time (in milliseconds). The velocity is computed for each interval defined by two consecutive MOUSE_UP and MOUSE_DOWN events. Let us assume two consecutive MOUSE_UP and MOUSE_DOWN events, *mup* and *mdo*, respectively in the coordinates $(x1, y1)$ and $(x2, y2)$, that took place respectively in the instants $time_1$ and $time_2$. Let us also assume two vectors *posx* and *posy*, of size n, holding the coordinates of the consecutive MOUSE_MOV events between *mup* and *mdo*. The velocity between the two clicks is given by $r\_dist/(time_2 - time_1)$, in which r_dist represents the distance travelled by the mouse and is given by equation 3.

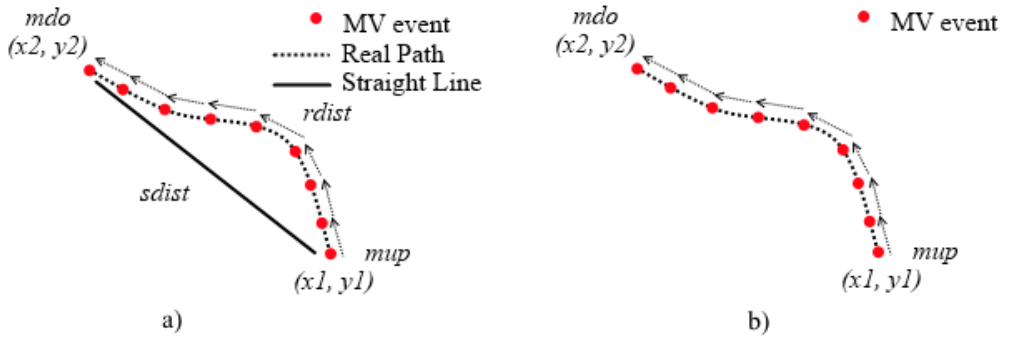

Figure 3: (a) A series of MOV events, between two consecutive clicks of the mouse. The difference between the shortest distance (sdist) and distance actually travelled by the mouse (rdist) is depicted. (b) The real distance travelled by the mouse between each two consecutive clicks is given by summing the distances between each two consecutive MOV events.

Time Between Clicks (TBC)
UNITS - milliseconds
The timespan between two consecutive MOUSE_UP and MOUSE_DOWN events, i.e., how long did it took the individual to perform another click.
These features quantify the performance of the students' interaction with the computer. As an example, longer clicks as well as more excessive distance traveled by the mouse reveal a decreasing performance.

For each student, the collected data is aggregated and summarized at 5-minute intervals. The average value of the feature in the interval is used.

Figure 4 depicts the type of information that these features provide. It shows the evolution of the performance of a specific student during an exam through two features: Click Duration and Mouse Velocity. The duration of each click decreases until roughly the middle of the exam and then increases up to a global maximum. The velocity of the mouse increases until approximately the same point in time and then it starts decreasing. Both features point out an initial improvement of performance (faster clicks and increasing mouse velocity), followed by a degrading.

Figure 4 actually reveals a classical effect of stress: performance tends to improve for some time after the beginning of the stressor stimulus (eustress), with a drop off in performance after some time performing above average (distress)[26].

These features allow for an individualized view on how stress affects each particular student, potentially devising each one's breaking points, behavior or overall performance under stress.

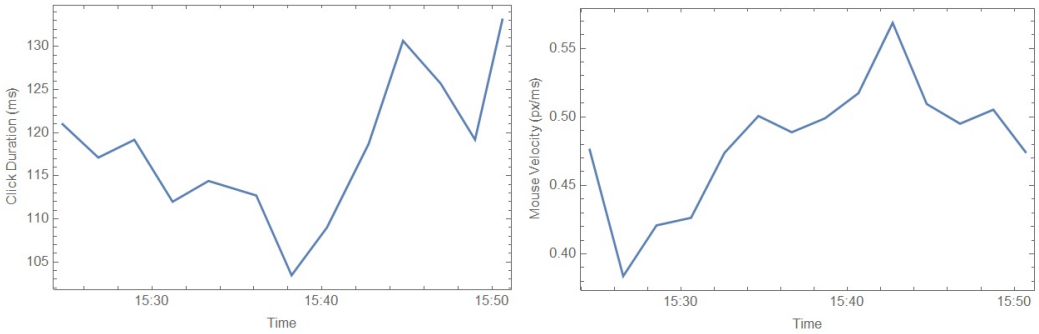

Figure 4: Real-time performance: evolution of one student's interaction performance during the exam. Left: Click Duration. Right: Mouse Velocity.

3.3. Dataset

In order to analyze the students' behavior throughout the exam and, more precisely, the evolution of students' performance, we performed a least-squares fit to a quadratic curve of each student/feature. The hypothesis is that the way the performance of the student changes through the exam, and consequently the shape of the quadratic curve, may be related to certain characteristics of the student.

A univariate quadratic function can be represented in the standard form as $f(x) = ax^2 + bx + c$, where a, b and c represent the coefficients. The coefficient a controls the degree of curvature of the graph: a larger magnitude of a gives the graph a more closed (sharply curved) appearance. Moreover, a positive value of a results in an parabola open upwards, and vice-versa. The coefficients b and a together control the location of the axis of symmetry of the parabola. Finally, the coefficient c controls the height of the parabola where it intercepts the y-axis.

In the context of this study, a larger magnitude of a indicates a larger variation of performance throughout the exam while a smaller one indicates that the performance was more constant. Moreover, a positive value of a indicates that the performance drops at the beginning of the exam and then improves, and vice-versa.

Figure 5 depicts a least-squares fit of a quadratic curve to the data depicted in Figure 4. It makes the temporal evolution of performance (improving and then degrading) more clearly visible. In this specific case, the corresponding quadratic functions are $f(x) = 2.44 * 10^{-11}x^2 - 70.45x + 5.09 * 10^{13}$ (Figure 5, left) and $f(x) = -7.88 * 10^{-14}x^2 - 0.23x - 1.64 * 10^{11}$ (Figure 5, right).

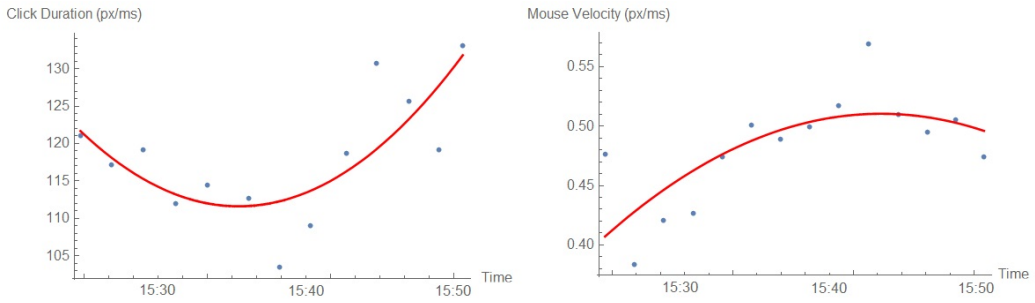

Figure 5: Modeling student performance: the temporal evolution of performance is made clearer trough the least-squares fit of a quadratic curve (same data as depicted in Figure 4).

Given the importance of the coefficient a in shaping the temporal evolution of performance, this study focuses on this coefficient. Moreover, given that we are interested in distinguishing between students who are able to maintain a steady performance and students whose performance varies significantly, we consider its modulus $|a|$.

Consequently, the dataset used in this study has the following structure:

one attribute identifying the student/exam, one attribute that quantifies the completion time (in minutes) of that exam for that student and ten more attributes, each one quantifying the modulus of the coefficient a of the quadratic function of each of the features described in Section 3.2. The dataset is thus composed of twelve columns and 297 instances, one for each participating student. An excerpt of the dataset is presented:

```
#id,duration,ASA,ADMSL,AED,CD,DBC,DMSL,ED,MA,MV,TBC
a73332,36.10,15.17,0.049,0.003,0.06,0.02,9.12,...
a65778,62.79,5.07,0.01,0.0003,0.006,0.12,3.05,...
a71033,25.52,6.91,0.04,0.0004,0.09,0.20,8.14,...
a71216,58.75,3.12,0.004,0.0009,0.006,0.11,6.12,...
a71061,26.45,4.19,0.05,0.0004,0.05,0.004,13.16,...
```

4. Methodology and Results

The analysis of the collected data started with a visual analysis. Figure 6 depicts the scatter diagrams of the modulus of the coefficient a of the quadratic functions that fit each feature against the completion time of the exams. Some interesting preliminary insights are revealed. Specifically, the scatter diagrams show that longer completion times seem to be associated to smaller magnitudes of $|a|$. In practical terms, this means that students who stay longer at the exams are also those whose performance varies less during the exam.

These differences are visible when comparing graphically the variation of performance of students who spent different times to complete the same exam. As an example, Figure 7 shows the evolution of click duration for four different students taking the same exam. The students depicted in the top row took more than 2 hours to complete the exam. The values of the coefficient a are, respectively, -0.0003, -0.001 and -0.002. On the other hand, the students in the lower row completed the exam in around 30 minutes. The values of the coefficient a in this case are, respectively, 0.056, 0.088 and -0.81.

Indeed, this inverse relationship between the two variables is confirmed by the figures detailed in Table 2, that shows the correlation between the completion time of each exam (A - D) and the magnitude of $|a|$ in each feature. A negative correlation between the variables exists in all features/exams,

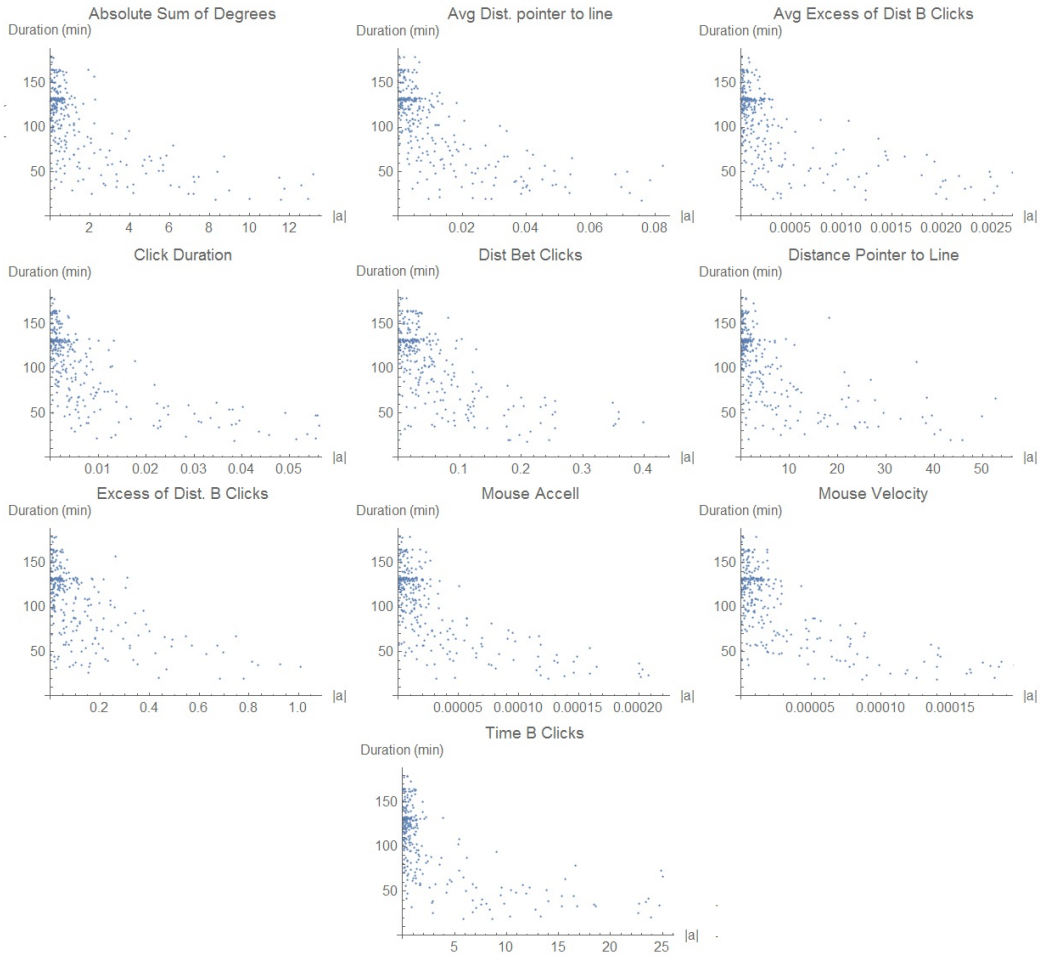

Figure 6: Exam completion times vs. performance variation: scatter diagrams plotting the modulus of the coefficient a of the quadratic functions that fit each feature against the completion time of the exams.

with values ranging between -0.69 and -0.35 which, considering that it is a biological phenomena being studied, constitute very interesting values.

The table also shows that, in average, the features that show the strongest correlation are CD, MA and MV (-0.58, -0.61 and -0.57, respectively).

This constitutes, in itself, a rather interesting fact that will be investigated further in the future, namely in search for correlation with other characteristics of the student including academic performance, perceived stress effects of objective measures of stress. It could be possible that the behavior of the

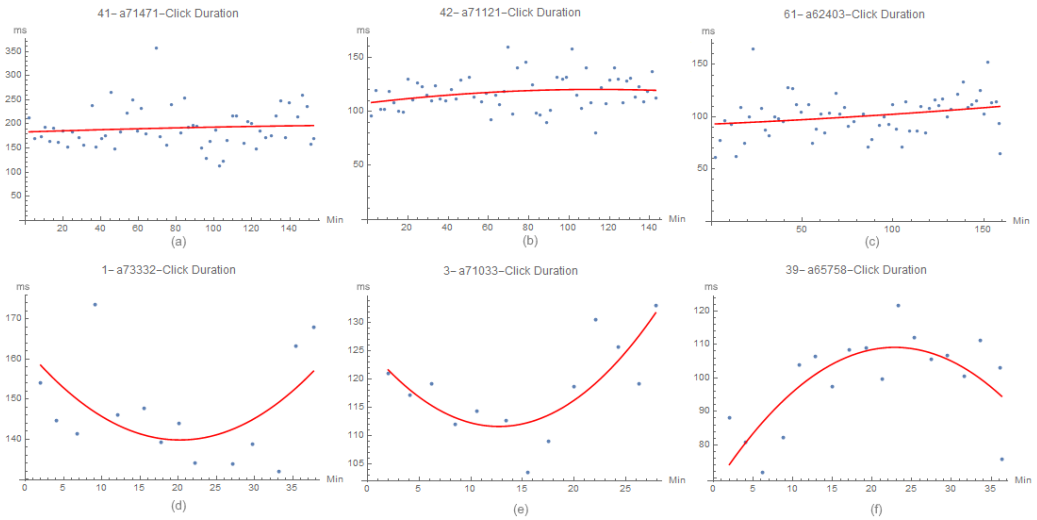

Figure 7: Performance over time: differences in the shape of the linear regression (for click duration) when comparing students who spent around 2 hours to complete the exam (first row) and students who spent around 30 minutes (second row).

student is related to stress coping strategies or stress response.

However, given the relatively strong negative correlation between the completion time of the exam and $|a|$ in all students in general and in all four exams, the next step was to verify the possibility of predicting the completion time of the exam based on the student's interaction patterns.

In this process there was also an interest in determining which, of the ten features considered, were the most relevant for the problem as well as the most appropriate classifier. To this end, four different classifiers were used: Random Forests, Neural Networks, Linear Regression and Gaussian Process.

Table 2: Correlation, for each exam and each feature, between the completion time of the exam and $|a|$.

	ASA	ADMSL	AED	CD	DBC	DMSL	ED	MA	MV	TBC
A	-0.47	-0.42	-0.45	-0.64	-0.43	-0.45	-0.35	-0.64	-0.60	-0.52
B	-0.35	0.46	-0.37	-0.40	-0.45	-0.50	-0.50	-0.61	-0.57	-0.45
C	-0.58	-0.56	-0.58	-0.59	-0.57	-0.42	-0.65	-0.54	-0.50	-0.61
D	-0.62	-0.69	-0.66	-0.69	-0.51	-0.56	-0.64	-0.67	-0.59	-0.57
$\overline{x}$	-0.50	-0.53	-0.51	-0.58	-0.49	-0.48	-0.54	-0.61	-0.57	-0.54

Random Forest predictors use an ensemble of decision trees to predict the intended value. Each decision tree is trained on a random subset of the training set and only uses a random subset of the features. A Neural Network consists of several layers of computing neurons, with information being processed in each layer and passed on from the input to the output layer. The Neural Network is trained to minimize a loss function on the training set using gradient descent. The Linear Regression predicts a numerical output using a linear combination of numerical features. The combination of these numerical features is also obtained by minimizing a loss function. Finally, the Gaussian Process method assumes that the function to be modeled has been generated from a Gaussian process, defined by a so-called covariance function. In the training phase, the method estimates the parameters of this covariance function and is then conditioned on the training data and used to infer the value of a new example using a Bayesian inference.

Each of these algorithms were used to train a different classifier for each subset of features of the original dataset. Since there were ten features in the dataset, 1023 classifiers were trained for each of the four algorithms. This corresponds to the number of subsets in a set of size 10 (2^{10}) excluding the empty subset.

Random forest predictors were composed by 200 trees, a leaf size of 5 and a variable sample size of 2. Neural Networks were configured with two hidden layers and 15 nodes in each layer. Concerning the numerical covariance function of the Gaussian Process, the Squared Exponential function was used.

Given the size of the dataset, the holdout method was employed to build the training and testing samples [27]. This method suggests that the available data should be randomly split into two disjoint subsets for a single train-test experiment. One group is used for the task of training while the other is held for the task of testing. In this specific work, the training set holds two thirds of the data while the test set holds the remaining third.

One limitation of this method is that, since the results are so dependent on the choice of the training/testing set of samples, they will be misleading in the event of an unfortunate division. For instance, it might be too easy/difficult to classify certain examples of data in the testing set, leading to biased results.

In order to avoid this drawback, this process of dividing the dataset and training a classifier was repeated twenty times for each subset of features and each classifier, and its results averaged. The measure of fit of each of the predictors is thus the average error of the prediction (in minutes) over the twenty rounds.

Figure 8 illustrates how the average error of the predictors varies with the different subsets of features considered. The first subsets include a single feature each. As the number of the subset grows, the number of features considered also grows until all ten features are considered in the last subset. Figure 8 depicts how the average error decreases from around 20 minutes when only one feature is considered to around 14 minutes when an increased number of features are considered.

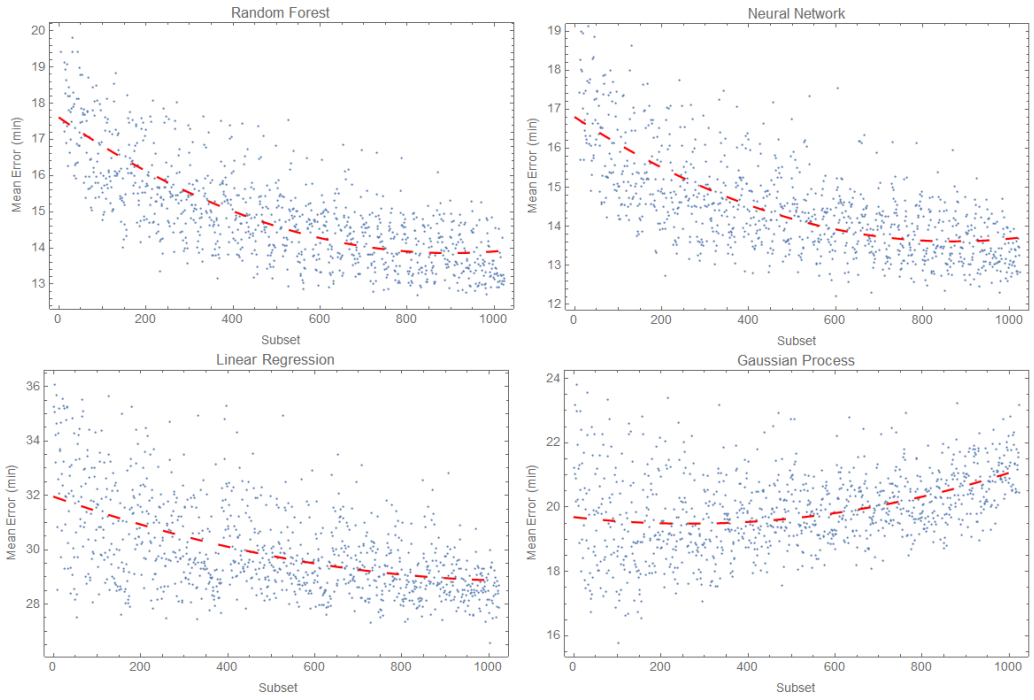

Figure 8: Mean error of each subset of features and each algorithm: evolution of the average error of the predictors trained as the number of features considered increases.

Table 3 shows the results of each predictor when using all features, in terms of the average, standard deviation, minimum and maximum value of error (in minutes).

Tables 4, 6, 5 and 7 present some more details about these results. Tables 4 and 5 show the five subsets with the best performance for the Random Forest and Neural Network algorithms, respectively. The top average performance is achieved by the subset of features {3, 4, 5, 6, 8, 10} using Random Forests and by the subset {3, 4, 7, 9, 10} using Neural Networks, with an

Table 3: Performance of each predictor when using all features (units: minutes).

Subset	$\bar{x}$	σ	Min	Max
Random Forests	13.15	0.80	11.48	14.72
Neural Networks	12.82	0.88	11.39	14.14
Gaussian Process	20.44	2.02	17.59	23.54
Linear Regression	28.68	1.76	26.24	31.61

Table 4: Five best subsets in terms of predictor performance using Random Forests (units: minutes).

Subset	$\bar{x}$	σ	Min	Max
{3, 4, 5, 6, 8, 10}	12.70	0.80	11.14	14.45
{1, 2, 3, 4, 7, 8, 9, 10}	12.73	0.64	11.90	14.02
{2, 3, 4, 5, 6, 8, 10}	12.77	0.94	11.37	14.39
{1, 3, 4, 6, 8, 10}	12.78	0.89	11.71	15.02
{1, 3, 4, 6, 8, 9, 10}	12.81	0.63	11.70	13.90

average error of 12.7 and 12.2 minutes, respectively, over the 20 runs. In both algorithms, minimum errors of 10 to 11 minutes occur. The low magnitude of the standard deviation (< 1 minute in both algorithms) in all the predictors shows the consistency of each one over the 20 rounds. The other two algorithms are left out of this more detailed analysis as they perform significantly worse.

Tables 6 and 7 detail the five subsets with the worst performance for the Random Forests and Neural Networks algorithms, respectively, all of which composed by a single feature. In these subsets, the error averages 22 minutes. Three of the five features of each group are common for both algorithms.

Given these results, the predictor trained using Neural Networks and with the feature set {3, 4, 7, 9, 10} can be selected as the best one. Figure 9 depicts a plot of the actual values versus the predicted ones, for the best group of features of each algorithm. The dashed line depicts the perfect correlation. The correlation value between the actual and the predicted values for each of the algorithms is as follows: 0.938 for Random Forests, 0.898 for Neural Networks, 0.605 for Linear Regression and 0.877 for Gaussian Process.

Finally, we also evaluated the Random Forests predictor on the training

Table 5: Five best subsets in terms of predictor performance using Neural Networks (units: minutes).

Subset	$\bar{x}$	σ	Min	Max
{3, 4, 7, 9, 10}	12.22	0.94	10.96	13.88
{1, 2, 3, 4, 5, 8, 9, 10}	12.30	0.91	10.91	13.64
{1, 2, 4, 5, 9, 10}	12.31	1.01	10.82	13.85
{1, 2, 3, 4, 5, 6, 9}	12.36	0.68	11.44	13.34
{1, 4, 6, 7, 8, 9, 10}	12.47	1.02	10.86	14.24

Table 6: Five worst subsets in terms of predictor performance using Random Forests (units: minutes).

Subset	$\bar{x}$	σ	Min	Max
{3}	22.35	1.84	19.46	25.15
{7}	21.88	1.73	18.70	24.80
{6}	21.597	1.23	19.25	23.66
{1}	21.50	1.57	18.70	25.03
{10}	21.31	1.46	18.80	23.52

Table 7: Five worst subsets in terms of predictor performance using Neural Networks (units: minutes).

Subset	$\bar{x}$	σ	Min	Max
{3}	21.95	1.20	20.37	24.05
{5}	21.58	3.42	17.72	29.27
{7}	21.22	1.80	17.47	24.42
{1}	20.76	1.90	18.08	24.35
{9}	20.53	1.18	19.10	23.21

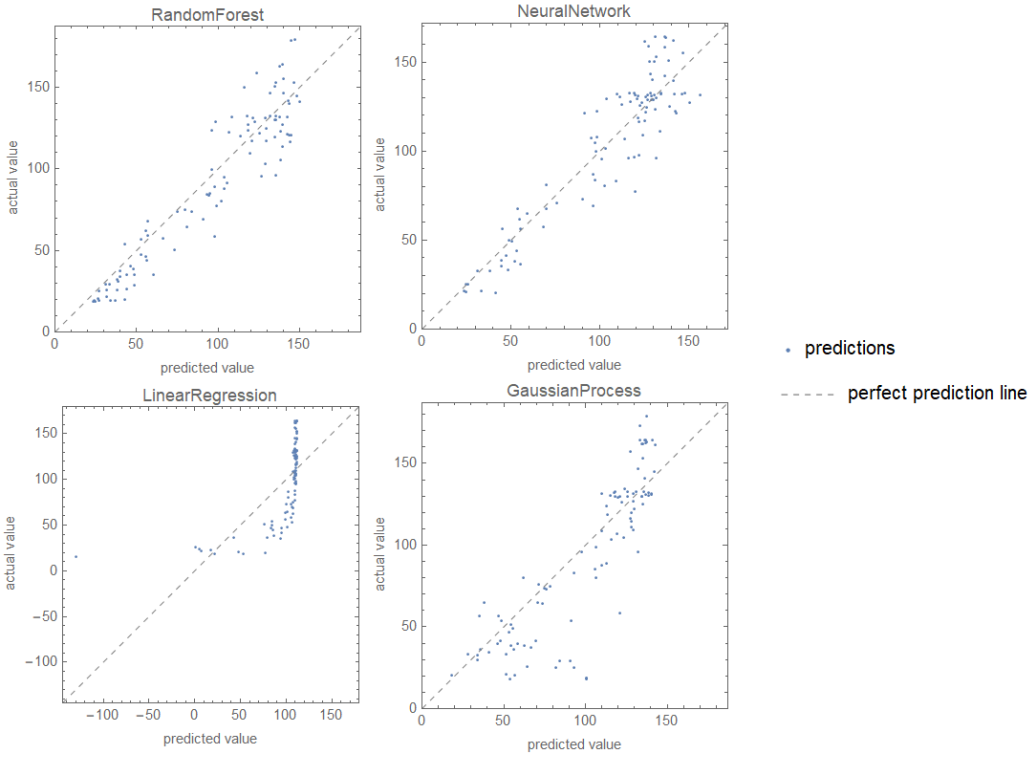

Figure 9: Validation: plot of the actual values against the values predicted by the random forest for the test set.

set. The results are, as expected, better than on the test set, as depicted in Table 8

5. Discussion and Conclusion

Stress is a complex phenomena, with effects at many different levels on the individual. It is our belief that some of these effects are still to be discovered, especially in cases in which they are measured through recent technological developments. That is the case of the approach studied in this work.

It is a known fact that, among others, stress affects our behavior. In the past, an individual's behavior would be characterized mostly in terms of social behavior (e.g. social withdrawal, mood, irritability, sadness/depression) or physical behavior (e.g. body language, muscle tension or pain). In recent years, our interaction with technological devices became another potential source of information to characterize our state or our behavior, with

Table 8: Performance of the Decision Forests on the training set, using the best subsets of features (units: minutes).

Subset	$\bar{x}$	σ	Min	Max
{1, 3, 4, 6, 8, 9, 10}	5.36	0.33	4.71	5.72
{2, 3, 4, 5, 6, 8, 10}	5.43	0.19	5.12	5.77
{1, 3, 4, 6, 8, 10}	5.50	0.20	5.28	5.76
{1, 2, 3, 4, 7, 8, 9, 10}	5.52	0.21	5.14	5.69
{3, 4, 5, 6, 8, 10}	5.57	0.22	5.15	5.95

the added advantage that data can be collected continuously, autonomously, and transparently.

In this paper we focus on such an approach. Specifically, we look at mouse dynamics to characterize user behavior during high-stakes exams. One important aspect worth highlighting, especially concerning this domain of application, is the relationship between the features used and performance. Indeed, all the features can be used to quantify the performance of the student's interaction with the computer: longer clicks or larger distances traveled with the mouse to complete the same task are signs of lower performance. Given that there are differences in the average values of performance between students, one future research direction will be to assess if this measure of performance correlates with academic performance.

Moreover, and as the data points out, it is also interesting to note that there are two main groups of students in what concerns behavior during an exam.

The first group is constituted by those students who take longer to complete the exam. Indeed, there are students who take more than 2 hours to do so. Interestingly enough, these are also the students who maintain a more steady performance throughout the exam.

Then, there is the group of students who complete the same exam significantly faster, in as low as 30 minutes. When comparing this group with the previous one, the main difference is that these students show a more marked variation of performance throughout the exam. Moreover, in this group there are students who improve performance until roughly the middle of the exam and then start degrading, and vice-versa. See for example Figure 7 (g): the student starts the exam with clicks of around 80 milliseconds, the duration then increases until reaching a maximum of 120 milliseconds (an increase of

40%), and at the end of the exam it decreases back to around 80 milliseconds. And all this happens in the span of 35 minutes, which is this student's completion time for this exam. The mechanisms underlying this behavior are still to be explored. Namely, what makes this student behave like this and students (e) and (f) in Figure 7 behave exactly the opposite (i.e. improving performance and then degrading)? Is this related to some characteristic of the student such as stress coping strategies or the way they perceive stress? This is something that will be pursued in the future.

While many new questions are raised by this research work, some concrete practical advances are also put forward. First of all, it is shown that the proposed approach can effectively be used to characterize the behavior of students in a non-intrusive way, in real time. It can collect data in a distributed way, from dozens or hundreds of students simultaneously without interfering with the routine of the exam.

Secondly, a random decision forest and a neural network to predict exam completion time were trained and evaluated. The encouraging results show, on the one hand, the significance and consistency of the behaviors observed. On the other hand, they open the door to a possibility not yet explored, to the extent of our knowledge: the one of predicting, in real-time, the exam completion time. Indeed, this may constitute a valuable tool for a professor to assess the state of the students while taking an exam, eventually intervening to calm them if necessary and possible.

If intervening is not possible, which is common in high-stakes exams, the professor still has a valuable tool to assess the magnitude of the effects of this stressful experience on the students. This is fundamental, especially when the educational organization assumes the mission of providing the students with the best possible environment, including the provision of stress coping guidance and initiatives: such an approach could point out those students who might be more affected by stress.

We believe that these interesting results can still be further improved, namely by using additional modalities. In the present study, we were limited by the hardware available in the exam rooms (i.e. mouse, keyboard and screen).In future work we will consider the inclusion of other sources of information such as web-cams, pressure sensitive keyboards or mice equipped with galvanic skin response sensors, to create a multi-modal approach on the problem which we expect will improve its accuracy.

Concluding, the presented approach unveils interesting future research directions that may lead to a better understanding of the behavioral effects

of stress, namely on Human-Computer Interaction. Specifically, it may allow to better understand how stress influences student behavior during high-stakes exams and eventually predict other related and important student characteristics.

Acknowledgments

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013. This work was funded by "EUSTRESS – Sistema de Informação para a monitorização e avaliação dos níveis do stress e previsão de stress crónico", N°2015/017832 P2020 SI I&DT, (NUP, NORTE-01-0247-FEDER-017832) in co-promotion between Optimizer-Lda and ICVS/3B's-Uminho.

Bibliography

[1] J. Firth, Levels and sources of stress in medical students., Br Med J (Clin Res Ed) 292 (6529) (1986) 1177–1180.

[2] M. Dahlin, N. Joneborg, B. Runeson, Stress and depression among medical students: A cross-sectional study, Medical education 39 (6) (2005) 594–604.

[3] S. M. Suldo, E. Shaunessy, R. Hardesty, Relationships among stress, coping, and mental health in high-achieving high school students, Psychology in the Schools 45 (4) (2008) 273–290.

[4] S. L. Shapiro, G. E. Schwartz, G. Bonner, Effects of mindfulness-based stress reduction on medical and premedical students, Journal of behavioral medicine 21 (6) (1998) 581–599.

[5] M. Bhatnagar, R. K. Jain, S. K. Nilam, A survey on behavioral biometric techniques: mouse vs. keyboard dynamics, in: IJCA Proceedings on International Conference on Recent Trends in Engineering and Technology, 2013, pp. 27–30.

[6] K. Revett, H. Jahankhani, S. T. de Magalhães, H. M. Santos, A survey of user authentication based on mouse dynamics, in: Global E-Security, Springer, 2008, pp. 210–219.

[7] I. Donald, P. Taylor, S. Johnson, C. Cooper, S. Cartwright, S. Robertson, Work environments, stress, and productivity: An examination using asset., International Journal of Stress Management 12 (4) (2005) 409.

[8] M. Le Fevre, J. Matheny, G. S. Kolt, Eustress, distress, and interpretation in occupational stress, Journal of Managerial psychology 18 (7) (2003) 726–744.

[9] G. P. Chrousos, P. W. Gold, The concepts of stress and stress system disorders: overview of physical and behavioral homeostasis, Jama 267 (9) (1992) 1244–1252.

[10] J. M. Weiss, Somatic effects of predictable and unpredictable shock., Psychosomatic Medicine 32 (4) (1970) 397–408.

[11] H. Selye, The stress of life. edição revista (1978).

[12] J. E. Dimsdale, J. Moss, Plasma catecholamines in stress and exercise, Jama 243 (4) (1980) 340–342.

[13] L. M. Vizer, L. Zhou, A. Sears, Automated stress detection using keystroke and linguistic features: An exploratory study, International Journal of Human-Computer Studies 67 (10) (2009) 870–886.

[14] D. Carneiro, J. C. Castillo, P. Novais, A. Fernández-Caballero, J. Neves, Multimodal behavioral analysis for non-invasive stress detection, Expert Systems with Applications 39 (18) (2012) 13376–13389.

[15] D. Carneiro, P. Novais, M. Gomes, P. M. Oliveira, J. Neves, A statistical classifier for assessing the level of stress from the analysis of interaction patterns in a touch screen, in: Soft Computing Models in Industrial and Environmental Applications, Springer, 2013, pp. 257–266.

[16] M. J. SMITH, F. T. CONWAY, B.-T. KARSH, Occupational stress in human computer interaction, INDUSTRIAL HEALTH 37 (2) (1999) 157–173. doi:10.2486/indhealth.37.157.

[17] M. Pantic, L. J. Rothkrantz, Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE 91 (9) (2003) 1370–1390.

[18] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. G. Taylor, Emotion recognition in human-computer interaction, IEEE Signal processing magazine 18 (1) (2001) 32–80.

[19] R. Jacob, K. S. Karn, Eye tracking in human-computer interaction and usability research: Ready to deliver the promises, Mind 2 (3) (2003) 4.

[20] A. Jaimes, N. Sebe, Multimodal human–computer interaction: A survey, Computer vision and image understanding 108 (1) (2007) 116–134.

[21] A. Barreto, J. Zhai, M. Adjouadi, Non-intrusive physiological monitoring for automated stress detection in human-computer interaction, Human–Computer Interaction (2007) 29–38.

[22] S. Baltaci, D. Gokcay, Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features, International Journal of Human–Computer Interaction 32 (12) (2016) 956–966.

[23] J. Hernandez, P. Paredes, A. Roseway, M. Czerwinski, Under pressure: sensing stress of computer users, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 2014, pp. 51–60.

[24] S. M. Case, D. B. Swanson, Constructing written test questions for the basic and clinical sciences, National Board of Medical Examiners Philadelphia, 1998.

[25] D. Carneiro, P. Novais, J. M. Pêgo, N. Sousa, J. Neves, Using mouse dynamics to assess stress during online exams, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2015, pp. 345–356.

[26] I. Rodríguez, M. W. Kozusznik, J. M. Peiró, Development and validation of the valencia eustress-distress appraisal scale., International Journal of Stress Management 20 (4) (2013) 279.

[27] A. K. Jain, Advances in statistical pattern recognition, in: Pattern recognition theory and applications, Springer, 1987, pp. 1–19.