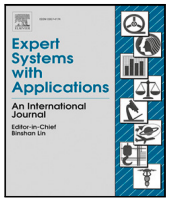
Expert Systems With Applications 242 (2024) 122864

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A multimodal study of the interplay between stress, executive function, and biometrics in game-based assessment

Mariano Albaladejo-González [a,*], Rubén Gaspar-Marco [a], Nancy Tsai [b], Félix Gómez Mármol [a], José A. Ruipérez-Valiente [a]

[a] University of Murcia, Calle Campus Universitario, 30100, Murcia, Spain
[b] McGovern Institute for Brain Research, Massachusetts Institute of Technology, MA 02139, Cambridge, USA

ARTICLE INFO

Keywords:
Stress
Biometrics
Executive functions
Artificial Intelligence

ABSTRACT

Managing stress is a crucial soft skill that affects cognitive performance and health. Stress detection through biometrics can be used to improve and evaluate stress management. However, measuring the effects of stress on biometrics and executive functions is difficult and dependent on the individual. Despite these challenges, this paper presents a case study that collects a comprehensive multimodal dataset with two stress metrics, four biometric signals, and twenty-two executive function metrics from Game-based Assessment (GBA) trace data specifically designed for this purpose. The experiments suggest that biometrics, especially the heart rate and skin temperature, are effective predictors of stress. Additionally, noteworthy correlations were observed between heart rate and certain executive function variables. The levels of GBA that measured shifting and processing speed showed a higher heart rate than the response inhibition levels. This case study, together with the developed stress detectors, enables the detection of persons who struggle to manage stress and measure their executive function performance under stressful situations.

1. Introduction

Soft skills such as stress management, teamwork, and leadership are considered indispensable for modern workers (Vasanthakumari, 2019). This paper focuses on stress management, one critical soft skill due to its relationship to health (O'Connor et al., 2021) and executive functions (Tsai et al., 2019), affecting students' and workers' performance (Pascoe et al., 2019; Pluntke et al., 2019). Nevertheless, stress self-awareness is a complex self-regulated capability that most people do not have (Albaladejo-González & Ruipérez-Valiente, 2022). Therefore, they might not detect their high stress levels until the situation is problematic and difficult to reverse. Stress management is important for all workers, but it is especially essential for those who work in high-pressure environments and make critical decisions, such as emergency professionals. In these situations, inadequate stress management can result in deficient performance, which can have fatal consequences (Pluntke et al., 2019).

While stress detection can help address the problem mentioned above, the most straightforward method for measuring stress is through subjective self-reporting using validated questionnaires. However, this approach has limitations, such as self-biases and the time required for sustained use, making it inconvenient. To overcome these limitations, affective computing aims to develop machine systems that can automatically recognize emotions, including stress, without relying on self-reporting. One approach for automatic stress prediction is affective computing with biometrics data (Mohammadi et al., 2022; Motogna et al., 2021) given that some biometrics are linked to stress, mainly heart rate and heart rate variability (Szakonyi et al., 2021). Stress also affects executive functions, which refer to the high cognitive processes that allow planning, forethought, and goal-directed actions (Shields et al., 2016). We propose to analyze the relationship between executive functions, biometrics, and stress together to explore potential applications for stress detection. However, the differences between individuals can make it challenging to use biometrics for this purpose (Hu et al., 2019). Moreover, previous studies have shown that individual factors play a critical role in how stress affects a person's executive functions (Tsai et al., 2019). Therefore, not all individuals experience changes in their executive functions in the same way during stressful situations, making it difficult to use these functions for stress detection. Another challenge is to measure biometrics and executive functions, especially at the same time, because too invasive and uncomfortable

* Corresponding author.
E-mail addresses: mariano.albaladejog@um.es (M. Albaladejo-González), ruben.gasparm@um.es (R. Gaspar-Marco), ntsai@mit.edu (N. Tsai), felixgm@um.es (F. Gómez Mármol), jruiperez@um.es (J.A. Ruipérez-Valiente).

https://doi.org/10.1016/j.eswa.2023.122864
Received 17 March 2023; Received in revised form 26 November 2023; Accepted 4 December 2023
Available online 6 December 2023

sensors can affect the executive function evaluation itself (Zamkah et al., 2020).

Analyzing the relationship between stress, biometrics, and executive function is highly relevant for today's society. On the one hand, it allows progress in stress prediction through biometrics in order to avoid the harmful consequences of stress on health. On the other hand, this analysis contributes to research on executive functions under stressful conditions, which is especially relevant for professionals who make critical decisions and students who are often under stressful conditions, such as during exams.

Due to the importance of analyzing stress, biometrics, and executive functions, this paper presents a case study to collect and analyze the three of them. The case study records different biometrics under non-stress and stress conditions and also calculated different executive function metrics in the stress conditions. To calculate the executive function metrics, we employed Game-based Assessment (GBA), which refers to the use of games to assess learners' competencies, skills, or knowledge (Gomez et al., 2022). Using a novel GBA of executive functions is one of the unique contributions of this study. Furthermore, the study stands out for including a wide variety of multimodal data from different sources; specifically, the combined analysis of executive functions, biometrics (blood volume pulse, electrodermal activity, temperature, and three-axis acceleration), and trace data from GBA for stress prediction is a novelty in the state-of-the-art. Because no author has previously measured all these data together, the findings of this analysis are highly relevant due to their novelty. Finally, the case study and the developed Artificial Intelligence (AI) stress predictors can be utilized to find subjects who do not manage stress correctly and may be used to train stress management. Since our case study incorporates the assessment of executive function metrics, we can delve deeper into the cognitive abilities of individuals under stressful conditions. Poor executive function performance and elevated stress levels suggest difficulties managing stress effectively, which is essential for professionals who make critical decisions in stressful scenarios.

We establish the following research questions (RQs) for this study:

RQ1 What is the relationship between the subjects' executive function metrics and biometrics and self-reported stress? We split this RQ into two sub-RQs.

- RQ1.1 Which executive function and biometric variables have the higher prediction power for stress?
- RQ1.2 What is the potential effectiveness of a stress detector that utilizes a combination of biometric measurements and executive function metrics?

RQ2 What is the relationship between biometrics and executive function metrics?

RQ3 Do the stressors presented in the GBA influence the subjects' heart rate?

The rest of the paper is organized as follows. Section 2 provides a background of stress, biometrics, and executive functions. Section 3 presents the multimodal case study developed. Section 4 introduces the methodology followed to answer through the dataset each of the RQs. Section 5 shows the results obtained, and Section 6 contains the discussion about the outcomes, implications, and limitations. Finally, we present the research conclusions and future work in Section 7.

2. Related works

Stress and mental health are a worldwide concern (Pourmohammadi & Maleki, 2020). Stress is described as either being acute or chronic in nature (Wolff et al., 2020). Acute stress constitutes physiological, psychological, and behavioral responses to demands that exceed an organism's regulatory capacity, mainly in uncontrollable or unexpected situations and lasts only for a short period of time. When these stressful situations are continuous or prolonged, they induce chronic stress that can lead to serious physical and mental diseases (Greene et al., 2016). Proper stress management is necessary to avoid its negative effects on health (O'Connor et al., 2021) and cognitive consequences in today's society (Tsai et al., 2019).

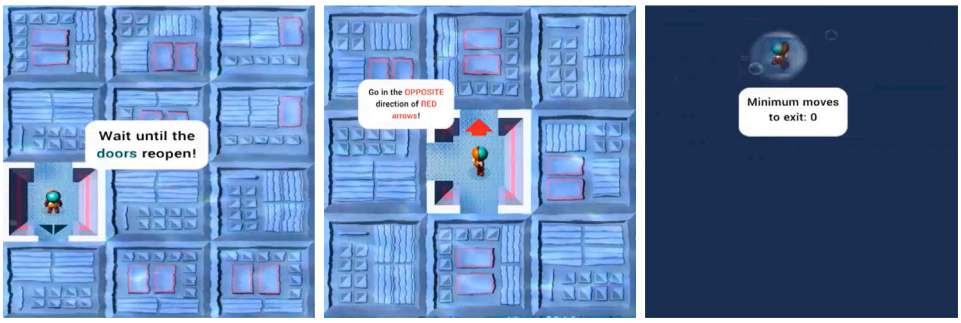
Due to the aforementioned reasons, acute stress prediction is a hot topic (Motogna et al., 2021; Panicker & Gayathri, 2019) and one of the main approaches is to use biometrics because they are linked to stress, mainly heart rate (Motogna et al., 2021; Szakonyi et al., 2021). Typical data processing consists of splitting the biometric signals into time windows and extracting a series of features from each window (Albaladejo-González et al., 2022). Most authors of the state-of-the-art trained AI models with features from windows recorded in stress and non-stress conditions (Albaladejo-González et al., 2022). Therefore, most authors have considered stress prediction a binary classification problem (Panicker & Gayathri, 2019), whereas in this case study, we operationalize it as a regression problem, testing two different stress metrics from self-reported validated instruments. Employing this regression approach enables a higher level of granularity in stress prediction. Instead of simply categorizing individuals as stressed or not stressed, we can provide a more accurate estimation of the stress level. This is particularly relevant to evaluate stress management and its improvement.

Although other authors have analyzed stress prediction through biometrics, there is a lack of datasets to train AI models to predict stress and compare results. As far as we know, these are the three most comprehensive datasets that contain stress and biometrics: AffectiveRoad (Haouij et al., 2018), Wearable Stress and Affect Detection (WESAD) (Schmidt et al., 2018), and Smart Reasoning Systems for Well-being at Work and at Home-Knowledge Work (SWELL-KW) (Koldijk et al., 2014). Like the case study presented in this paper, the case studies from which these datasets were generated enclosed stress and non-stress phases. The stress phase incorporated some stressors that induce high levels of stress. However, none of these datasets included as many stressors as our case study. About the biometric data collected, WESAD includes, in addition to our biometrics, the electromyogram (EMG) and respiration. Nevertheless, none of these datasets also contain any executive function metrics; therefore, there is no information on how stress affects the subjects' cognitive performance. It is also worth noting that these biometrics and stress datasets do not contain a very large number of different subjects because it is difficult to find subjects to record biometrics data, and it is common to lose some subjects due to poor-quality measurements or problems with some sensors. AffectiveRoad, WESAD, SWELL-KW, and our dataset comprise 10, 15, 25, and 20 subjects, respectively. Therefore, the research community still needs more similar studies.

In addition to acute stress prediction, our dataset has also allowed the analysis of stress and executive functions, which is another relevant research field with no definitive conclusion (Plieger & Reuter, 2020; Tsai et al., 2019). Some authors have found that acute stress impairs working memory and cognitive flexibility (Shields et al., 2016); other authors have noticed that the acute stress reaction is adaptive, providing power and helping the organism to deal with stressors in challenging situations (Dhabhar, 2018).

The analysis of executive functions under stressful conditions included in our case study allows for detecting subjects who need to improve their capacity to work in stressful situations. With this purpose, we can find some previous applications to improve stress management for specific professionals such as firefighters (Pluntke et al., 2019) or soldiers (Friedl, 2018). However, these applications focus on specific professionals, while our case study and the AI stress detector developed can evaluate any individual.

To measure executive functions, there exist validated tests such as the *n*-back task for working memory, the continuous performance



(a) Response inhibition levels (b) Shifting levels (c) Working memory levels

Fig. 1. AquaPressure video game screenshots.

task for inhibitory control, or the Wisconsin Card Sorting Test for shifting (de Assis Faria et al., 2015). Outside the laboratory, there are also a couple of GBA to measure these cognitive skills (Gomez et al., 2022), but neither combine the executive function metrics with biometrics. Furthermore, we also collected two self-reported stress metrics, making our case study stand out for the different sources and modalities analyzed. Therefore, this multimodality represents an advance with respect to the state of the art.

3. Dataset collection

This section introduces the case study developed to obtain the data analyzed in the following sections. To obtain further details regarding the dataset or the code developed, please contact the authors.

3.1. AquaPressure

AquaPressure is a GBA of executive functions developed at the Massachusetts Institute of Technology (MIT). This GBA consists of 50 levels where the player must navigate through maze-like rooms to reach an exit without running out of oxygen. Levels assess the executive functions of inhibitory control, shifting, and working memory and are all designed with slight modifications using basic mechanics of following arrows to navigate through the rooms. For example, in the inhibitory control levels, arrows are accompanied by arrows with special symbols indicating the player should stop rather than pass through the arrow. Levels assessing shifting and working memory are also designed to uniquely test the cognitive construct of interest. Please see Fig. 1 as an example of three levels within AquaPressure.

Two versions of AquaPressure were designed to examine the effects of low and high stress on executive functions. In the present study, only the high-stress version of AquaPressure was used thus, this version will simply be identified as AquaPressure. The first three levels of AquaPressure include tutorial levels where no stressors are present in the game.

3.2. Case study

To answer the aforementioned RQs, we have designed and conducted a case study to record biometric measurements in non-stress and stress conditions and also calculated different executive function metrics in the stress conditions.

3.2.1. Procedure

The developed case study is summarized in Fig. 2. To collect biometrics under non-stress conditions, we created the video phase, in which the volunteer watched an introductory video about executive functions. The stress phase was conducted using AquaPressure, allowing us to collect biometrics and executive functions under stress conditions.

Before starting the experiment, the volunteers watched an introductory video about the case study. This video introduced the case study briefly to the volunteers, informing them about the two phases of the case study and that they had to wear a non-invasive wristband to collect their biometrics. We used a video to give all volunteers the same information about the case study.

Subsequently, the volunteers had to sign a consent form to participate in the experiment. After these previous steps, the volunteers put on the Empatica E4 (a wristband to record biometrics) (Empatica, 2023) in their left hand, and the experiment started with the video phase. In the video phase, we showed a video about executive functions to record the biometric signals under non-stress conditions. We did not include questions about the content of the video, and the video did not contain any relevant information to play AquaPressure. Therefore, the video did not contain any stressors.

Then, the volunteers performed the AquaPressure phase, including the tutorial, test, and game phases, while recording the biometric signals. The next phase was the test phase, which consisted of participants completing a mental arithmetic task out loud. The mental arithmetic task is adapted from the Trier Social Stress Task (Kirschbaum et al., 1993) which aims to induce stress. Participants were told, in order to stress them, that performance on this math task is used to assign them to play at beginner, intermediate, or advanced levels, though all participants are assigned to the beginner level irrespective of their actual performance. Moreover, the researcher told the volunteers that they would be playing against prior players and that their performance, based on accuracy and speed, would be ranked on a live scoreboard, again to stress the volunteers. At the start of the mental arithmetic task, a second observer is welcome to join to monitor the participant, a fictitious "Professor Eli" connected via Zoom, who would analyze their games and screen recordings. The inclusion of "Professor Eli" was intended to increase the volunteers' stress by making them feel evaluated and observed. Finally, during the game phase, the subjects played AquaPressure which includes various additional features designed to enhance feeling stressed, such as exaggerated visual and auditory cues when mistakes are made, a fictitious red-blinking video recording icon, and shortened time to complete levels.

Once the gameplay was over, the participants filled out a questionnaire reporting their stress levels during the video and AquaPressure phases. We collected the stress levels in both phases to calculate the stress difference between the two phases.

3.2.2. Participants

The dataset contains 20 subjects without missing variables (i.e., complete cases), reporting a stress level during the AquaPressure phase equal or higher than in the video phase. All users collected are of legal age and are mostly students or professors at the University of Murcia, Spain. The age of the participants ranged from 18 to 47 years, and 60% of the participants were males.

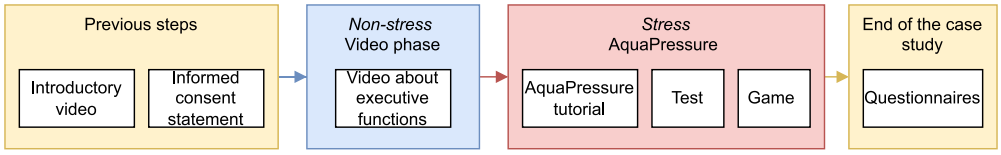


Fig. 2. Phases of the case study.

3.2.3. Instruments and collected data

Our study used three sources to collect physiological, behavioral, and self-report data: the Empatica E4 (Empatica, 2023), AquaPressure executive function GBA, and questionnaires. The Empatica E4 is a non-invasive wristband that measures BVP, electrodermal activity (EDA), temperature, and three-axis acceleration. We chose the Empatica E4 because it is the non-invasive device that measures most biometric signals with high precision (McCarthy et al., 2016). This wristband recorded biometrics from subjects under stress and non-stress conditions. To split the signals in each phase, we also noted the beginning and end of each phase through the Empatica by adding an event through the Empatica and manually by taking notes. On the other hand, AquaPressure executive function GBA recorded behavioral executive function performance metrics under the stress phases. The AquaPressure logs contain the volunteer's interactions with the video game for each level. These AquaPressure logs and biometrics required previous preprocessing and feature engineering to obtain the final variables needed for the RQs. Finally, the questionnaires include the State-Trait Anxiety Inventory (STAI; Marteau & Bekker, 1992), a 5-point Likert scale, and a Visual Analog Scale (VAS) from 0 to 100 to register self-reported stress levels during the video and AquaPressure protocol. The questionnaires also required previous steps to compute the scores.

4. Methodology

The methodology followed to answer the aforementioned RQs is depicted in Fig. 3. This figure shows the data sources for the case study and their use in each RQ.

4.1. Preprocessing and feature engineering

The preprocessing and the feature extraction applied to generate the variables employed are summarized in Fig. 4, including the prefixes used in the variable names according to their origin and the suffixes indicating the function applied in the final aggregation.

To extract the variables from the biometric signals, we split the signal of each phase into sliding time windows with a duration of 180 seconds and a shift of 15 seconds with the previous signal. The window and shift sizes were selected after several trials from window sizes ranging from 120 to 390 seconds and shift sizes between 0 and 30 seconds, checking the performance achieved in RQ1.2 and the robustness features extracted. Then we extracted different variables for each window and signal. We employed the library Neurokit 2 (Makowski et al., 2021) to extract the heart rate and EDA variables. Finally, we aggregated the data of each subject's windows, obtaining six variables for each of the variables extracted during the previous step. The application of sliding windows for biometric signal processing is common in the state-of-the-art; however, we had to summarize these data to generate one single input per subject since the executive function metrics calculated by GBA through the AquaPressure video game are only obtained per subject. Even so, we tested the RQ1.2 experiments without aggregating the data in windows and observed that aggregating the data in windows mainly improved the AI models' performance; one possible reason is that aggregating the data in windows reduced the impact of noise. The aggregation functions used were the maximum, minimum, mean, median, standard deviation, and variance.

To measure executive functions, we used the log file provided by AquaPressure, which contains the video game events in a JSON format, including the room and level where the event occurred. Another file stored the content of the rooms of each level which allowed us to identify the peculiarities of the rooms and levels of the player's events as the presence of inverted controllers. For inhibitory control, shifting, and working memory, we calculated different metrics summarized in Table A.5 in Appendix A.

To calculate stress, we computed from the STAI and VAS the stress variation between the video and the AquaPressure protocol (including the tutorial, the math test, and the video game). We calculated the difference between the reported stress on AquaPressure and the video phase, named as change, following the formula: $(stress\_aquapressure - stress\_video)$, being $stress\_aquapressure$ the stress reported in AquaPressure, and $stress\_video$ the stress watching the video. Therefore, we obtained the stress metrics: *STAI_change* and *VAS_change*.

4.2. Final dataset

Finally, from the biometrics, we obtained 582 variables (see Fig. 4). From AquaPressure logs, we obtained 22 executive function metrics of response inhibition, processing speed, shifting, cognitive inhibition, and working memory (see Table A.5). Finally, the *STAI_change* and *VAS_change* variables, used as response variables in RQ1, were calculated from the questionnaires.

4.3. RQ-specific analysis details

This subsection introduces the methodology followed specifically for each research question.

4.3.1. RQ1.1. Which executive function and biometric variables have the higher prediction power for stress?

To answer this RQ, we calculated the Pearson correlation coefficient of the biometric and executive function variables. Later, the two variables most correlated with each stress metric were used to model the stress change as a linear regression. We selected only two variables to comply with dummy rules, such as 10 cases per predictor variable. To avoid the collinearity of these two variables, we employed the Variance Inflation Factor (VIF) metric. If these variables showed VIF values above 5, the second variable was switched to the next most correlated variable until the VIF was lower than or equal to 5.

4.3.2. RQ1.2. What is the potential effectiveness of a stress detector that utilizes a combination of biometric measurements and executive function metrics?

Stress prediction using AI is a relevant field of research in today's society due to the numerous benefits of detecting high-stress levels in different environments and contexts. RQ1.2 focus on developing a stress detector using our dataset. In contrast to RQ1.1, we applied for RQ.1.2 a leave-one-subject-out (LOSO) cross-validation to obtain a more realistic performance evaluation. In LOSO cross-validation, one subject is used as the test set and the rest as the training set; this process is repeated until all subjects have been employed as a test set. The LOSO cross-validation was repeated with the two variables most correlated with each stress metric with a VIF lower or equal to 5 (obtained in RQ1.1), the two variables selected by Recursive Feature Elimination (RFE), and all the variables. Besides, in RQ1.2, we evaluated more AI models: Regression Tree, Random Forest (RF), Support

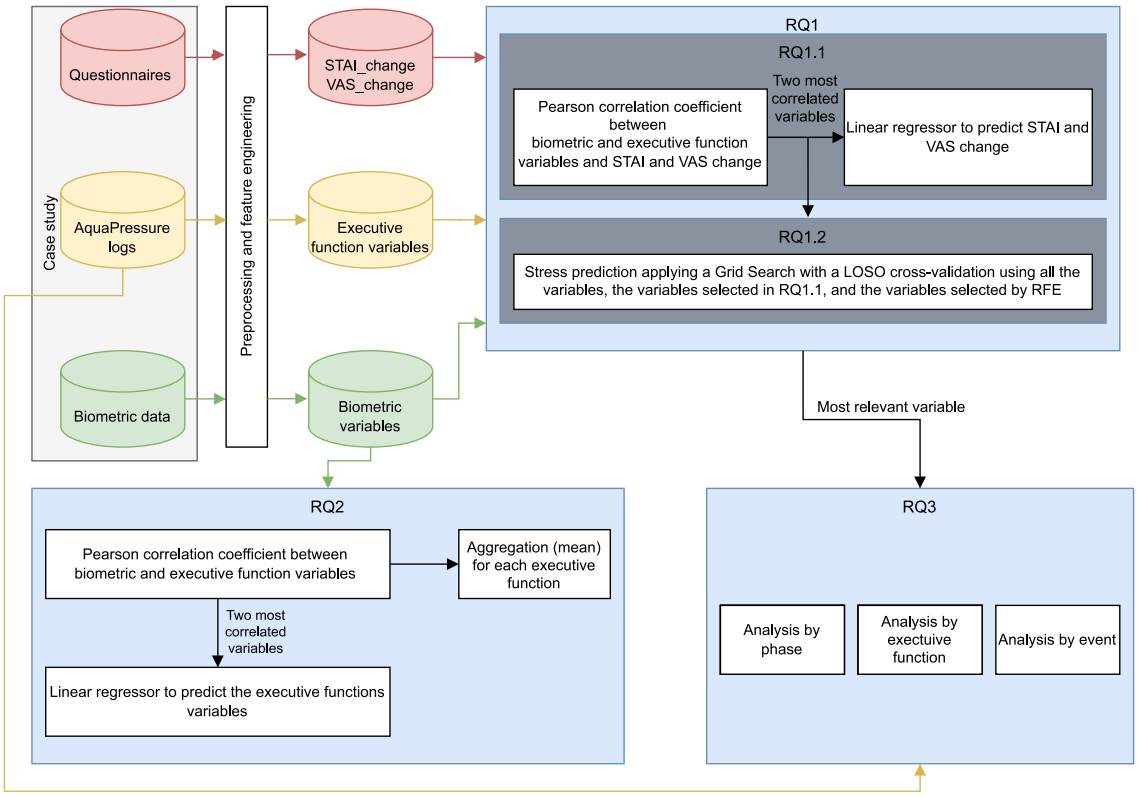


Fig. 3. Methodology applied to answer the RQs.

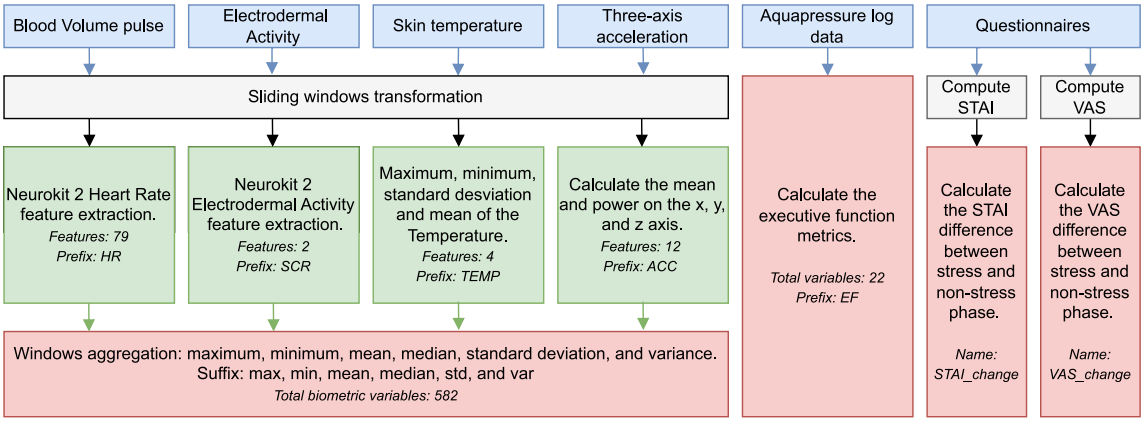


Fig. 4. Preprocessing and feature engineering applied to the collected data.

Vector Regression (SVR), Elastic Net (a linear regression that combined L1 and L2 regularizers), and a dummy model that predicts the mean of the target variable to have a baseline performance. To configure the hyper-parameters of these models, we included a grid search with LOSO cross-validation; therefore, we applied LOSO cross-validation to each configuration.

4.3.3. RQ2. What is the relationship between biometrics and executive function metrics?

The methodology applied in RQ2 is similar to the first part of RQ1.1. We calculated the Pearson correlation coefficient between biometric variables and the executive functions metrics. Then, we calculated the average correlation between the biometric variables and each type of executive function (working memory, response inhibition, processing speed, shifting, and cognitive inhibition). We applied this aggregation to summarize the results due to the high amount of executive function metrics. Finally, we utilized the two biometric variables most correlated to each executive function metric with a VIF lower or equal to 5 to model the executive function metrics as linear regression.

4.3.4. RQ3. Do the stressors presented in the GBA influence the subjects' heart rate?

The first step was to analyze the mean heart rate of all the users in the different phases of the case study. We decided to use the mean heart rate because heart rate variables were the most common in the rankings of correlations with both stress metrics in RQ1, and among the heart rate variables, the mean heart rate is the easiest to interpret and measure with external devices. Heart rate values due to measurement error were removed. Then we evaluated the mean heart rate at each level to detect patterns such as heart rate increases in specific types of levels. Finally, we looked for differences in the heart rate after

Table 1
Most correlated variables with the stress metrics (RQ1.1).

#	STAI_change	VAS_change
1	*TEMP_Mean_var* (0.75***)	*HR_CD_std* (0.55**)
2	*HR_AI_max* (0.75***)	*SCR_Peaks_Per_Second_var* (−0.52**)
3	*TEMP_Mean_std* (0.72***)	*SCR_Peaks_Per_Second_std* (−0.49**)
4	*TEMP_Max_std* (0.71***)	*HR_CD_var* (0.49**)
5	*HR_AI_std* (0.7***)	*HR_PAS_median* (−0.48**)
6	*TEMP_Max_var* (0.69***)	*HR_AI_max* (0.47**)
7	*HR_CVI_std* (0.69***)	*TEMP_Mean_std* (0.46**)
8	*TEMP_Min_var* (0.68***)	*TEMP_Min_std* (0.45**)
9	*HR_HFn_std* (0.68***)	*HR_GI_max* (0.45**)
10	*HR_MFDFA_alpha2_ Fluctuation_std* (0.67***)	*HR_PAS_mean* (−0.45**)

the subjects' AquaPressure errors, which showed a jarring sound and lowered the subject's oxygen bar.

For this purpose, we obtained the heart rate variation in a normal movement, in the collision with a wall, and in the collision with a door that opens and closes automatically. The 30 seconds before and the 3 seconds after the event were utilized to calculate the heart rate variation. To select the time before the event, we carried out different tests looking for the size of time that would return the heart rate in a time interval large enough to average the effect of previous events but trying to reduce the number of previous events. Again, the time after the event was selected by different tests whose objective was to use the shortest time possible to avoid losing the event's influence on the heart rate and also overlapping with the influence of the next events, still being long enough to process the heart rate. Furthermore, in the three analyses performed for RQ3, we looked for significant differences employing Friedman's non-parametric test (Friedman, 1937) since none of the distributions satisfied the normality conditions for applying ANOVA. In the case of finding significant differences, the Nemenyi test (Nemenyi, 1963) was applied to locate them.

5. Results

5.1. RQ1.1. Which executive function and biometric variables have the higher prediction power for stress?

The Pearson correlation coefficient between the top ten biometric and executive function variables most correlated to stress metrics and the stress metrics are shown in Table 1. In this table, we appreciate that all the top ten correlated features in absolute value are biometric variables, mainly from heart rate and skin temperature. In contrast, EDA features only appear two times in all the rankings, and accelerometer variables and executive function metrics do not appear in the rankings. These top ten correlated features have high correlation values with significant p-value (lower than 0.05), showing that heart rate and temperature features have high explanation power of the stress.

The two variables most correlated to each stress metric with a VIF lower or equal to 5 and the performance of the deployed linear regressor with these variables are summarized in Table 2. Between the two stress metrics, *STAI_change* achieved the highest R^2 and adjusted R^2 with 0.67 and 0.63, respectively. Furthermore, this linear regression has the independence of errors tested with a Durbin and Watson (1950) of 2.13. In contrast to the promising results achieved with the STAI metric, the *STAI_change* linear regression only achieved a R^2 of 0.42 and an adjusted R^2 of 0.35. These values are reasonable considering that we only used two variables, but they are significantly lower than the results achieved with *STAI_change*. However, the independence of errors is not clear, as we got a Durbin–Watson statistic of 3.14 in *VAS_change*. Finally, Fig. 5 shows a regression plot of the most correlated biometric variable with each stress metric to display the apparent relationship.

Table 2
Performance of stress linear regressors using two variables (RQ1.1).

	STAI_change	VAS_change
Var. 1	*TEMP_Mean_var* (43.6**)	*HR_CD_std* (100.4*)
Var. 2	*HR_AI_max* (1.01**)	*SCR_Peaks_Per_Second_var* (−26598.8*)
Const.	−48.24**	26.53*
R^2	0.67	0.42
Adj R^2	0.63	0.35
Durbin-Watson	2.13	3.14
VIF	2.13	1.33

Table 3
Performance of the stress models applying a LOSO cross-validation (RQ1.2).

Stress metric	Model	Linear		RFE		All	
		MSE	MAE	MSE	MAE	MSE	MAE
STAI_Change	RF	9.19	2.37	14.66	3.01	17.11	3.56
	Tree	11.95	2.76	18.75	3.49	17.91	3.11
	ElasticNet	10.80	2.69	16.49	3.22	283.78	10.04
	SVR	10.62	2.38	23.70	3.85	19.83	3.53
	Dummy	23.72	4.25	23.72	4.25	23.72	4.25
VAS_Change	RF	472.44	16.07	602.40	19.82	605.16	19.94
	Tree	400.72	14.93	651.18	20.03	851.84	23.47
	ElasticNet	463.05	17.22	633.53	21.11	6609.39	55.66
	SVR	420.45	15.97	1038.72	26.22	503.85	18.61
	Dummy	571.59	20.02	571.59	20.02	571.59	20.02

5.2. RQ1.2. What is the potential effectiveness of a stress detector that utilizes a combination of biometric measurements and executive function metrics?

The results of the applied cross-validation employing the two variables obtained in RQ1.1 and with all the variables are summarized in Table 3. For these results, we calculated the error metrics of Mean Square Error (MSE) and Mean Absolute Error (MAE). In this experiment, the best configuration of all models analyzed outperformed the dummy model; therefore, the predictors include information that enhances stress prediction. Besides, it is worth noting that all the models achieved the highest performance using only the two variables selected in RQ1.1.

5.3. RQ2. What is the relationship between biometrics and executive function metrics?

In the top four most correlated variables between the biometric variables and the executive functions, there are only heart rate variables in all the executive functions, and variables from the temperature and the accelerometer are very rare. Response inhibition and processing speed have one and two correlations equal to or higher than 0.5, showing an interesting correlation with the heart rate variables. In contrast, shifting stands out as the executive function with the lowest correlation value. The top ten average absolute values of Pearson correlation coefficients between the biometric and the executive functions are summarized in Table A.6 included in Appendix A

Employing the above results, we deployed a linear regressor using the biometric variables most correlated with a VIF lower or equal to 5 to predict each executive function metric; the performance of these regressors is summarized in Table 4. In this table, the maximum R^2 is at least 0.48 in all executive functions and 0.42 in the adjusted R^2. Besides, in cognitive inhibition and processing speed, the minimum R^2 is 0.41, and the adjusted R^2 is 0.34, showing that biometric variables, especially heart rate variables, correlate with both executive functions. In contrast, in working memory, the median adjusted R^2 is 0.3, and the minimum is 0.15, indicating that some working memory metrics are more difficult to predict using biometric variables. To visualize this relationship, we also display in Fig. 6 the regression plot of the

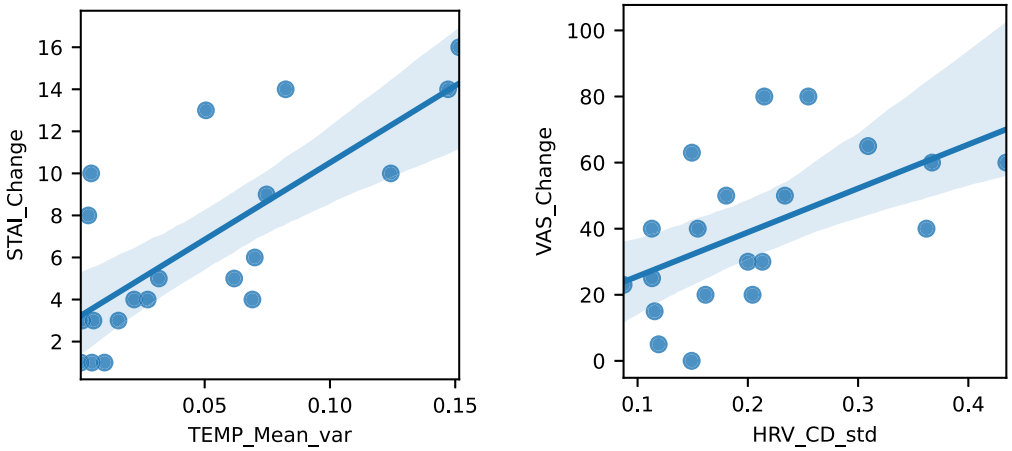


Fig. 5. Regression plot illustrating the apparent relationship between stress metrics and biometric variables.

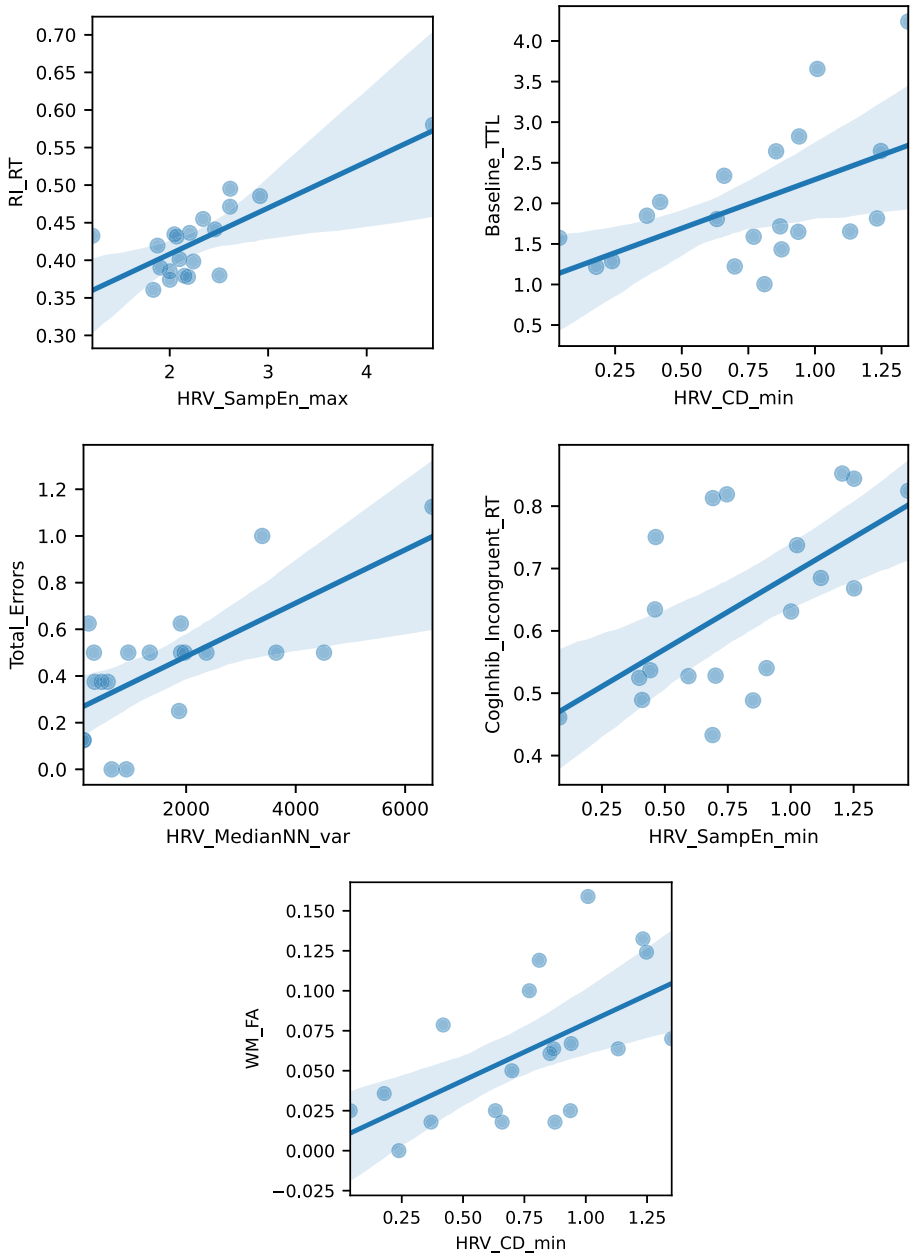


Fig. 6. Regression plot illustrating the apparent relationship between executive function metrics and biometric variables.

Table 4
Performance of executive function metrics regressors (RQ2).

Executive function	Best regressed variable	Max. R^2	Med. R^2	Min. R^2
Response Inhibition	*EF_RI_RT*	0.62 (adj. 0.57)	0.5 (adj. 0.44)	0.38 (adj. 0.31)
Cognitive Inhibition	*EF_Conflict_Resolution_Accuracy*	0.48 (adj. 0.42)	0.46 (adj. 0.39)	0.41 (adj. 0.34)
Processing Speed	*EF_Baseline_RT*	0.59 (adj. 0.54)	0.5 (adj. 0.44)	0.41 (adj. 0.34)
Shifting	*EF_Shifting_Congruent_Accuracy*	0.7 (adj. 0.67)	0.46 (adj. 0.4)	0.36 (adj. 0.29)
Working Memory	*EF_WM_FA*	0.59 (adj. 0.54)	0.38 (adj. 0.3)	0.24 (adj. 0.15)

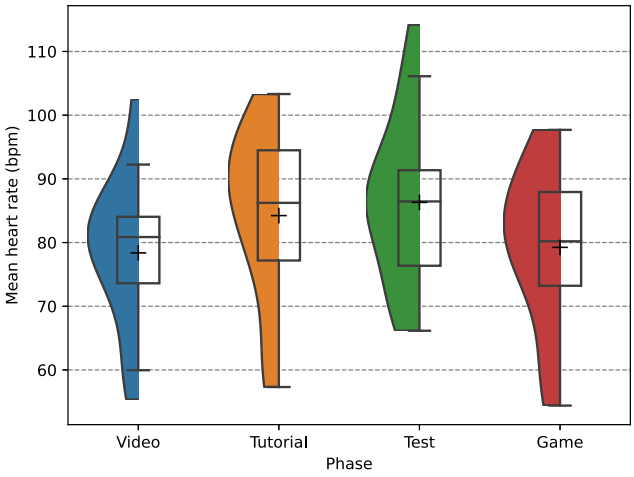


Fig. 7. Boxplot of the average heart rate in each phase.

most correlated metric between a biometric variable for each executive function.

5.4. RQ3. Do the stressors presented in the GBA influence the subjects' heart rate?

The average heart rate in each phase of the case study is summarized in Fig. 7. In this figure, we can notice that during the tutorial phase, the volunteers' heart rates increased compared to the video phase. Another observation is that the volunteers' heart rate is more dispersed during the tutorial, noting that the distance between the first and the third quartile is the largest of all the phases. In the test phase, the minimum, maximum, mean, and median increased their values compared with the tutorial; in contrast, the third quartile is lower. During the game phase, the heart rate decreased, with a minimum, maximum, and all the quartiles lower than in the test phase.

Friedman's test found significant differences in distributions of the mean heart rate of the subjects in each phase. The Nemenyi test located these differences between the video and the test phases, where we can appreciate that the heart rate during the video is lower than in the test phase. These findings are easy to visualize in some subjects, as Fig. 8 illustrates. This figure shows how the heart rate is the lowest during the video phase, increases during the tutorial, reaches the maximum during the test and decreases again during the game.

Later, we focused on heart rate during the game phase, looking for differences between the levels according to the executive function analyzed. These results are summarized in Fig. 9, where we see that the processing speed and shifting levels have a slightly higher third quartile than the rest of the levels. Friedman's test found significant differences in distributions, and the Nemenyi test located these differences in the subjects' heart rate between the levels of shifting and the levels of cognitive and response inhibition, and the levels of processing speed and response inhibition. Therefore, we detect higher heart rate distribution in processing speed and shifting, especially noticeable compared with response inhibition.

Finally, we examined the heart rate variation after the subjects' errors in AquaPressure. We obtained the heart rate variation in normal movements, hit with a wall, and slamming the door. Fig. 10 shows the heart rate variations 30 seconds before and 3 seconds after the movement. We cannot appreciate big differences in the heart rate variation between the types of movements analyzed, and Friedman's test did not find significant differences. In Fig. 8, we also do not observe that collisions with walls or doors change the subject's heart rate. As we have explained in the methodology, the time interval analyzed before and after the movements come from different tests previously performed.

6. Discussion

6.1. Experiments

The experiments developed for RQ1.1 show that biometrics, especially the heart rate and the skin temperature, have a high prediction power of stress metrics, especially the metrics calculated through the STAI questionnaire (RQ1.1). In addition, a prediction of both stress metrics is feasible in our case study, emphasizing a higher performance with the stress metric extracted from the STAI questionnaire (RQ1.2). It is worth noting that the best configurations of all the AI models used only two variables rather than all variables during the RQ1.2 experiments, indicating that using all variables is less recommendable in our database. This finding is interesting to consider for other studies and applications since obtaining and computing some of these variables can be tricky.

It is difficult to compare these results with the performance obtained by other authors because they considered stress prediction as a classification problem and reported classification error metrics. These authors have achieved high performance in stress classification in different contexts employing AI and biometrics (Albaladejo-González & Ruipérez-Valiente, 2022; Albaladejo-González et al., 2022; Panicker & Gayathri, 2019; Pourmohammadi & Maleki, 2020; Siirtola & Röning, 2020). Furthermore, we found one example of stress regression; however, authors of Siirtola and Röning (2020) finally chose a threshold and transformed the regression into a binary classification. It is also worth noting that authors of Siirtola and Röning (2020) and us predicted user-reported stress rather than whether subjects are in a stress phase or not (Albaladejo-González & Ruipérez-Valiente, 2022; Albaladejo-González et al., 2022; Panicker & Gayathri, 2019; Pourmohammadi & Maleki, 2020). Another difference is that after extracting the features of each window from the biometrics, other authors did not apply an aggregation of the windows for each user as we did due to the calculation of the executive function metrics for each subject and not each time window. This allowed them to have multiple inputs per user; thus, they had more data to train their models.

In our results, we also observed the highest correlations in the variables obtained from heart rate and skin temperature, and we can

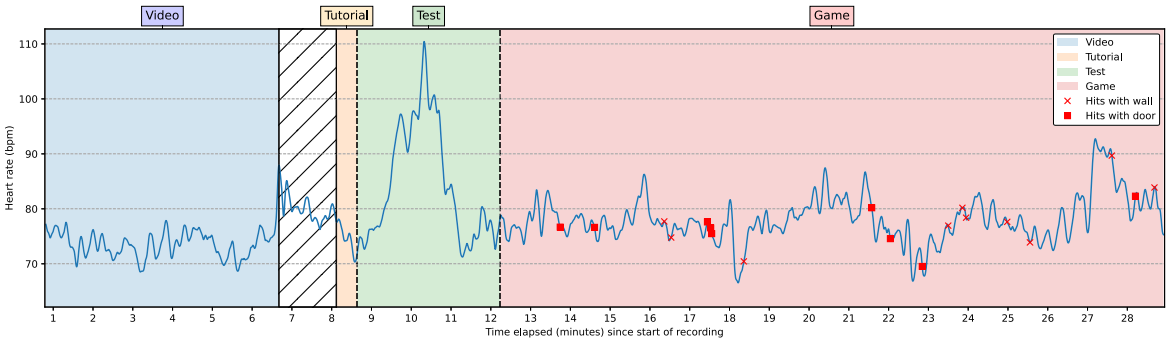


Fig. 8. Example of one subject's heart rate during the case study.

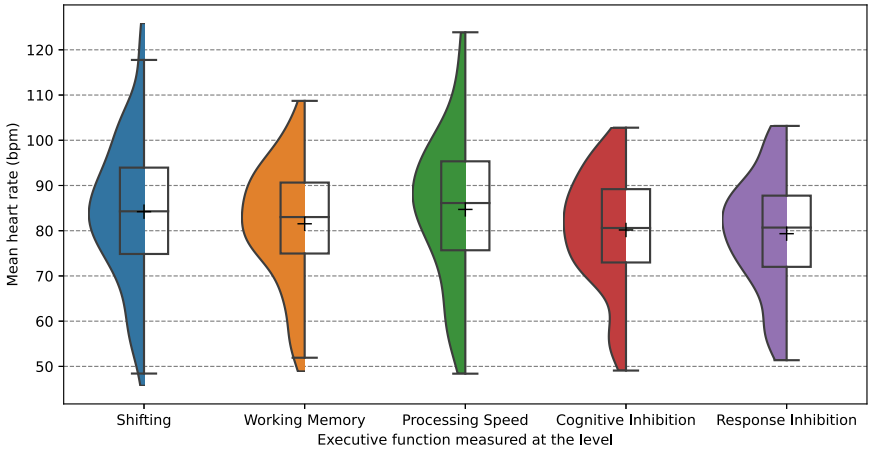


Fig. 9. Boxplot of the average heart rate within each level type.

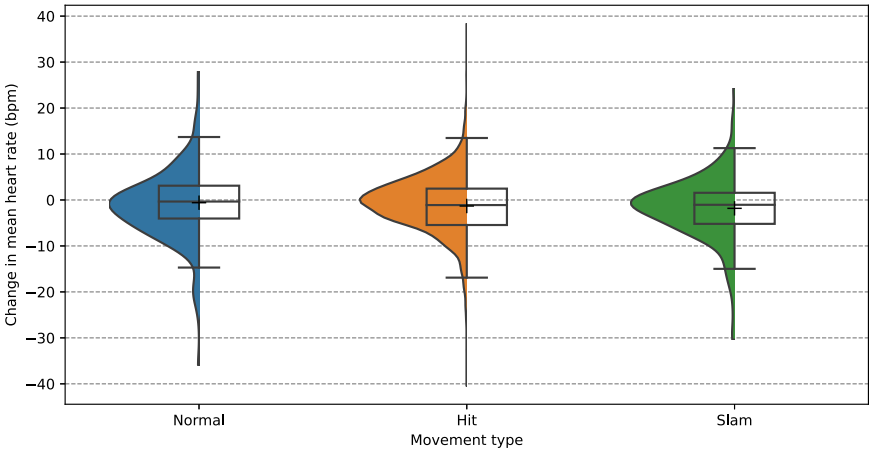


Fig. 10. Change in average heart rate between 30 seconds before a movement and 3 seconds after.

use these variables to develop stress detectors. Other authors have also previously noticed the importance of heart rate in stress prediction (Panicker & Gayathri, 2019; Zontone et al., 2019) and skin temperature (Siirtola & Röning, 2020). This conclusion is helpful because low-cost sensors in commercial wristbands can record these signals.

During RQ2, we found interesting correlations between heart rate and some executive function variables. One reason might be that heart rate and some executive function variables are both influenced by stress. This finding is relevant for the affective computing field because some AI models, such as linear regressors, do not work properly with correlated input variables. These correlations may also be one of the reasons why the best RQ1.2 configurations only use two variables instead of all variables.

In RQ3, we detected that during the test phase, which includes the mental arithmetic task and the inclusion of the fictitious professor, the subjects' heart rate increased compared to the video phase. On the other hand, the levels focused on measuring shifting and processing speed showed a higher heart rate distribution, mainly compared to the response inhibition levels. The levels focused on processing speed stand out for the presence of doors that open and close automatically, and levels of shifting are characterized by arrows indicating the opposite path to the one the subject should follow. In contrast, there is no clear cause for the null results obtained in the analysis of collisions and heart

rate; maybe the stakes of playing the AquaPressure video game are that high for participants to get stressed over small errors.

6.2. Implications

The results of RQ1 combined with results obtained by other authors demonstrate that stress prediction is feasible using biometrics, especially heart rate and skin temperature, which are available in many non-invasive commercial wristbands. This idea reinforces the argument that non-invasive stress detectors can be developed and used in many applications, such as work or educational environments, avoiding excessive exposure to high and dangerous stress levels. Furthermore, we found that executive function metrics do not provide essential information for stress prediction and that biometrics have a higher prediction power. In this paper, we are pioneers in combining biometrics and GBA trace data for stress prediction using AI. Although our experiments have shown that executive function metrics are not essential for predicting stress, we suggest their test in other affective computing applications. In RQ2 and RQ3, we also discovered interesting findings (mentioned in Section 6) that other authors could continue investigating to achieve more complete conclusions with high impact.

Finally, the case study combined with the stress predictors developed in RQ1 can be utilized to find subjects who do not manage stress correctly. The stress detector would indicate the subject's perceived stress, and through AquaPressure, we could obtain his executive function performance. High stress and low performance in executive function indicate poor stress management. It is essential to detect professionals who make critical decisions under stressful conditions that do not handle these situations appropriately. In addition, detecting poor stress management in subjects is also relevant because it affects their learning and performance. Furthermore, the case study can be utilized to train stress management, but we recommend further investigations into the long-term use of the case study.

6.3. Limitations

The main limitation of the paper is the number of subjects in our dataset, mainly due to the difficulty in finding volunteers to record biometrics. However, as we mentioned before, the size of our dataset is normal in the context of stress with biometrics. The other limitation to mention is that if we use our case study to detect individuals who do not manage stressful situations properly, some stressors may lose efficacy in sustained use, which is particularly relevant to be utilized to improve stress management. Therefore, we recommend new experiments about the long-term application of this case study.

7. Conclusions and future work

In the paper at hand, we developed a case study to record four biometric signals, twenty-two executive function metrics, and two stress metrics to research them together. The experiments showed that biometric variables, especially heart rate and skin temperature, had a high prediction power of stress metrics. In contrast, including executive function metrics in stress prediction was not essential, confirming with the results of other authors that stress prediction is feasible through biometrics. Furthermore, we observed correlations between executive function metrics and biometrics, especially heart rate data. This finding can be explored in other research to know the cause of these correlations. In the analysis of the heart rate during the case study, we found that the subject's heart rate increased during the test phase (one stress phase) compared to the video phase (non-stress phase). Besides, the GBA levels focused on measuring shifting and processing speed showed a higher heart rate distribution, mainly compared to other levels, especially the response inhibition levels. Finally, the developed case study enables the measurement of executive functions and biometrics under stressful conditions; together with the developed AI stress predictors, it can be utilized to find subjects who do not manage stress correctly.

In the future, these results could be replicated in new case studies to make additional and more robust conclusions, for example, by analyzing the correlations detected between biometric variables and executive function variables. Besides, other researchers may focus on validating some of our conclusions through variants of our case study. On the other hand, a front end could be implemented to facilitate the application of the case study to detect individuals with poor stress management. The long-term effect of the case study could also be analyzed, including different variants of the test phase and randomized levels, mainly to use the case study and the stress predictors for stress management training.

CRediT authorship contribution statement

Mariano Albaladejo-González: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Visualization. **Rubén Gaspar-Marco:** Conceptualization, Software, Validation, Formal analysis, Data curation, Writing – review & editing, Visualization. **Nancy Tsai:** Conceptualization, Resources, Writing – review & editing. **Félix Gómez Mármol:** Conceptualization, Writing – review & editing. **José A. Ruipérez-Valiente:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank Manuel Gómez Moratilla for processing the executive function metrics and Philip Tan, Louisa Rosenheck, William Freudheim, and Leslie Coles for their work on developing AquaPressure. Finally, this work has been partially funded under grant PID2021-122466OB-I00 (Spain), by MCIN/AEI/10.13039/501100011033/FEDER (Spain), (both founded by Spanish Ministry of Science, Innovation and Universities), by the strategic project CDL-TALENTUM (Spain) from the Spanish National Institute of Cybersecurity (INCIBE), and by REASSESS project (grant 21948/JLI/22 (Spain)), funded the Seneca Foundation, Science and Technology Agency of the Region of Murcia.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2023.122864.

References

Albaladejo-González, M., & Ruipérez-Valiente, J. A. (2022). Supporting stress detection via AI and non-invasive wearables in the context of work. In *Advances in analytics for learning and teaching* (pp. 77–97). Springer International Publishing.

Albaladejo-González, M., Ruipérez-Valiente, J. A., & Mármol, F. G. (2022). Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. *Journal of Ambient Intelligence and Humanized Computing*.

de Assis Faria, C., Alves, H. V. D., & Charchat-Fichman, H. (2015). The most frequently used tests for assessing executive functions in aging. *Dementia and Neuropsychologia*, *9*(2), 149–155.

Dhabhar, F. S. (2018). The short-term stress response – Mother nature's mechanism for enhancing protection and performance under conditions of threat, challenge, and opportunity. *Frontiers in Neuroendocrinology*, *49*, 175–192.

Durbin, J., & Watson, G. S. (1950). Testing for seriall correlation in least squares regression. I. *Biometrika*, *37*(3–4), 409–428.

Empatica (2023). E4 wristband: Real-time physiological signals: Wearable ppg, Eda, temperature, motion sensors.

Friedl, K. E. (2018). Military applications of soldier physiological monitoring. *Journal of Science and Medicine in Sport*, *21*(11), 1147–1153.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701.

Gomez, M. J., Ruiperez-Valiente, J. A., & Clemente, F. J. G. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies*, 1–16.

Greene, S., Thapliyal, H., & Caban-Holt, A. (2016). A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine*, *5*(4), 44–56.

Haouij, N. E., Poggi, J.-M., Sevestre-Ghalila, S., Ghozi, R., & Jaïdane, M. (2018). AffectiveROAD system and database to assess driver's attention. In *Proceedings of the 33rd annual ACM symposium on applied computing* (pp. 800–803). ACM.

Hu, X., Chen, J., Wang, F., & Zhang, D. (2019). Ten challenges for EEG-based affective computing. *Brain Science Advances*, *5*(1), 1–20.

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1–2), 76–81.

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). The SWELL knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 291–298). ACM.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, *53*(4), 1689–1696.

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the spielberger state—Trait anxiety inventory (STAI). *British Journal of Clinical Psychology*, *31*(3), 301–306.

McCarthy, C., Pradhan, N., Redpath, C., & Adler, A. (2016). Validation of the empatica E4 wristband. In *2016 IEEE EMBS international student conference (ISC)* (pp. 1–4). IEEE.

Mohammadi, A., Fakharzadeh, M., & Baraeinejad, B. (2022). An integrated human stress detection sensor using supervised algorithms. *IEEE Sensors Journal*, *22*(8), 8216–8223.

Motogna, V., Lupu-Florian, G., & Lupu, E. (2021). Strategy for affective computing based on HRV and EDA. In *2021 international conference on E-health and bioengineering (EHB)* (pp. 1–4). IEEE.

Nemenyi, P. B. (1963). *Distribution-Free Multiple Comparisons* (Ph.D. thesis), Princeton University.

O'Connor, D. B., Thayer, J. F., & Vedhara, K. (2021). Stress and health: A review of psychobiological processes. *Annual Review of Psychology*, *72*(1), 663–688.

Panicker, S. S., & Gayathri, P. (2019). A survey of machine learning techniques in physiology based mental stress detection systems. *Biocybernetics and Biomedical Engineering*, *39*(2), 444–469.

Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2019). The impact of stress on students in secondary school and higher education. *International Journal of Adolescence and Youth*, *25*(1), 104–112.

Plieger, T., & Reuter, M. (2020). Stress & executive functioning: A review considering moderating factors. *Neurobiology of Learning and Memory*, *173*, Article 107254.

Pluntke, U., Gerke, S., Sridhar, A., Weiss, J., & Michel, B. (2019). Evaluation and classification of physical and psychological stress in firefighters using heart rate variability. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2207–2212).

Pourmohammadi, S., & Maleki, A. (2020). Stress detection using ECG and EMG signals: A comprehensive study. *Computer Methods and Programs in Biomedicine*, *193*, Article 105482.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Laerhoven, K. V. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction* (pp. 400–408). ACM.

Shields, G. S., Sazma, M. A., & Yonelinas, A. P. (2016). The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol. *Neuroscience and Biobehavioral Reviews*, *68*, 651–668.

Siirtola, P., & Röning, J. (2020). Comparison of regression and classification models for user-independent and personal stress detection. *Sensors*, *20*(16), 4402.

Szakonyi, B., Vassányi, I., Schumacher, E., & Kósa, I. (2021). Efficient methods for acute stress detection using heart rate variability data from Ambient Assisted Living sensors. *BioMedical Engineering OnLine*, *20*(1).

Tsai, N., Eccles, J. S., & Jaeggi, S. M. (2019). Stress and executive control: Mechanisms, moderators, and malleability. *Brain and Cognition*, *133*, 54–59.

Vasanthakumari, S. (2019). Soft skills and its application in work place. *World Journal of Advanced Research and Reviews*, *3*(2), 066–072.

Wolff, M., Enge, S., Kräplin, A., Krönke, K.-M., Bühringer, G., Smolka, M. N., & Goschke, T. (2020). Chronic stress, executive functioning, and real-life self-control: An experience sampling study. *Journal of Personality*.

Zamkah, A., Hui, T., Andrews, S., Dey, N., Shi, F., & Sherratt, R. S. (2020). Identification of suitable biomarkers for stress and emotion detection for future personal affective wearable sensors. *Biosensors*, *10*(4), 40.

Zontone, P., Affanni, A., Bernardini, R., Piras, A., & Rinaldo, R. (2019). Stress detection through electrodermal activity (EDA) and electrocardiogram (ECG) analysis in car drivers. In *2019 27th European signal processing conference (EUSIPCO)* (pp. 1–5).