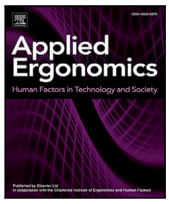
Applied Ergonomics 123 (2025) 104418

Contents lists available at ScienceDirect

Applied Ergonomics

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



Assessing operator stress in collaborative robotics: A multimodal approach

Simone Borghi [a,*], Andrea Ruo [b], Lorenzo Sabattini [b], Margherita Peruzzini [c], Valeria Villani [b]

[a] Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy
[b] Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Reggio Emilia, Italy
[c] Department of Industrial Engineering, University of Bologna, Bologna, Italy

ARTICLE INFO

Keywords:
Human–Robot Collaboration
Cobot programming
Stress evaluation
Wearable sensors
Psycho-physiological signals
Human monitoring

ABSTRACT

In the era of Industry 4.0, the study of Human–Robot Collaboration (HRC) in advancing modern manufacturing and automation is paramount. An operator approaching a collaborative robot (cobot) may have feelings of distrust, and experience discomfort and stress, especially during the early stages of training. Human factors cannot be neglected: for efficient implementation, the complex psycho-physiological state and responses of the operator must be taken into consideration. In this study, volunteers were asked to carry out a set of cobot programming tasks, while several physiological signals, such as electroencephalogram (EEG), electrocardiogram (ECG), Galvanic skin response (GSR), and facial expressions were recorded. In addition, a subjective questionnaire (NASA-TLX) was administered at the end, to assess if the derived physiological parameters are related to the subjective perception of stress. Parameters exhibiting a higher degree of alignment with subjective perception are mean Theta (76.67%), Alpha (70.53%) and Beta (67.65%) power extracted from EEG, recovery time (72.86%) and rise time (71.43%) extracted from GSR and heart rate variability (HRV) metrics PNN25 (71.58%), SDNN (70.53%), PNN50 (68.95%) and RMSSD (66.84%). Parameters extracted from raw RR Intervals appear to be more variable and less accurate (42.11%) so as recorded emotions (51.43%).

1. Introduction

A cobot is a robot type designed to work alongside humans in a shared workspace (Franklin et al., 2020). Unlike traditional industrial robots, which are typically caged off from human workers due to safety concerns, cobots are designed to be safe to work with Association et al. (1999).

There is a growing prevalence of cobots being employed in various industries: the U.S. market for collaborative robots is constantly growing, valued at $1.23 billion in 2022 and projected to grow at a Compound Annual Growth Rate (CAGR) of 32.0% from 2023 to 2030.[1] This is highly explanatory of the role that cobots have and will increasingly have.

Cobots possess several notable characteristics that make them valuable in several industrial contexts, such as safety towards humans (Javaid et al., 2022), user-friendliness (Taesi et al., 2023), flexibility and adaptability (Javaid et al., 2022). These features enable cobots to collaborate with humans in several industries, such as manufacturing, healthcare, and logistics. Main applications relate to assembly, packaging, and material handling (Javaid et al., 2022; Aaltonen and Salmi, 2019). In these applications, they can tackle tasks that are particularly repetitive, hazardous, or high-precision demanding, reducing, at the same time, the risk of accidents.[2]

Despite that, the best synergy between humans and cobots is achieved when several factors are satisfied: a highly efficient and technological cobot that lacks the characteristics necessary for a correct synergy with the operator would be useless and indeed detrimental to productivity. Typically, collaborative robots are considered more user-friendly when they are perceived as easy to use, adaptable, and possess human-like qualities (Moser et al., 2022; Valente and Avram, 2023).

Cobots are capable of carrying out tasks autonomously and with high precision but need to be programmed by a human operator (George et al., 2023). Programming cobots can be intricate and challenging for beginners, often resulting in stressful scenarios. The human dimension, and in particular the emotional and psycho-physiological one related to stress, must therefore be taken into consideration.

According to the European Foundation for the Improvement of Living and Working Conditions, work-related stress can lead to various disorders (Macias et al., 2007) of a physical and mental nature and it is a common issue reported by around 51% of European workers and significant concern for almost 80% of EU companies (Irastorza et al., 2016).

* Corresponding author.
E-mail address: simone.borghi@unimore.it (S. Borghi).

[1] https://www.grandviewresearch.com/industry-analysis/collaborative-robots-market
[2] https://www.universal-robots.com/blog/five-ways-cobots-can-improve-manufacturer-productivity/

https://doi.org/10.1016/j.apergo.2024.104418
Received 27 March 2024; Received in revised form 7 October 2024; Accepted 4 November 2024
Available online 16 November 2024

The significance of workplace stress is highlighted in a recent study[3] that conducted a survey involving at least 1000 employees. The survey reveals the percentage of employees in the United States who indicated their stress levels as consistently high, moderate, or low from 2016 to 2020: more than 80% declare a level of stress ranging from moderate to high.

Stress can be defined as a state of mental or emotional strain or tension resulting from adverse or demanding circumstances (Bryant et al., 2011). It is a natural response to challenging situations and can be beneficial in small doses, but chronic stress can have negative effects on physical and mental health leading to a range of physical and mental health problems, including high blood pressure, heart disease, depression, anxiety, weakening of the immune system, and insomnia (Mariotti, 2015). Stress is a complex concept with different proposed theories: Levine (2005) contended that it is simply a form of activation, which can become detrimental when prolonged over extended periods; Ursin and Eriksen (2010) offers a theory with a cognitive focus, while Koolhaas et al. (2011) provides a more biologically oriented approach. Herman et al. (2003) delineates a conceptually valuable division between "reactive responses" triggered primarily by internal stimuli (such as pain, homeostatic adjustments, and inflammatory signals) and anticipatory responses to external stimuli (including encounters with predators, social challenges, and more), as well as responses linked to memory. In light of numerous investigations, the main focus revolves around two fundamental stress theories, namely the General Adaptation theory proposed by Selye (1946) and Cannon's Fight-Flight response (Cannon, 1932). For this work, Cannon's approach was used: events that may appear difficult and threatening could induce different physiological reactions in the subject, preparing it to confront the challenge. These psycho-physiological responses can be observed and analyzed.

1.1. Paper contribution and organization

Building upon this, the objective of this study was to evaluate how learning cobot programming and engaging in the act of programming a cobot for the first time influence an individual's stress levels. For this purpose, this study used a learning platform defined in previous work (Hansen et al., 2023), which provides a safe and interactive environment to learn about the capabilities and limitations of collaborative robots and helps users gain the skills and knowledge needed. To assess the users' stress while learning to program a robot, we recorded a wide set of psycho-physiological signals to monitor their involuntary responses induced by stress. The main contributions of this research are:

1. Proposing a multimodal methodology to assess the operator's stress in HRC programming tasks, and
2. Demonstrating how EEG, ECG, GSR, and facial expression parameters can be interpreted to detect human stress conditions in HRC programming tasks.

As evidenced by the important number of recent works described in Section 2.2, stress evaluation of operators involved in HRC contexts is a hot topic. The objective of this work was to use different types of sensors in combination with subjective questionnaires, trusting that the multimodal approach can provide more precise information about the complex psycho-physiological responses at play. This work could help in bridging and integrating points that are still unknown in stress evaluation in general and HRC in particular.

In the following sections, we recall the state-of-the-art on this topic in Section 2. Then, the experimental setup, the organization of the tasks, and how the participants were selected and trained is described in Section 3. A detailed description of the signals obtained and how these are analyzed is provided in Section 4, to get from a raw signal to informative metrics. A thorough statistical analysis was performed to gather precise information concerning our case study: it is reported in Section 6 and elaborated upon in Section 7.

1.2. Paper organization

In the following sections, the experimental setup, the organization of the tasks, and how the participants were selected and trained will be described: as already mentioned above, the experiment aims to simulate what could be a real context of education in the use of cobots. A detailed description of the signals obtained and how these are analyzed will follow, to get from a raw signal to informative metrics. A thorough statistical analysis was performed to gather precise information concerning our case study, elaborated upon in Section 7.

2. Related work

2.1. Stress detection

As described in Section 1, stress is a complex physiological and psychological response that involves intricate interactions between various hormonal, neurological, and behavioral factors. So, to detect stressing conditions with the utmost reliability and comprehensive detail, employing a multimodal approach that incorporates metrics derived from various biological signals is essential. According to Giannakakis et al. (2019) there are several involuntary responses induced by stress, mediated by the autonomic nervous system that can create characteristic patterns in some psycho-physiological signals, such as ECG, GSR, EEG, and facial expression.

As evidenced by the literature, ECG is a physiological measure that can be used to evaluate stress (Andersson, 2017). The cardiac signal is characterized by peaks denoted with the letters P, Q, R, S, and T (De Luna, 2008). R-peaks indicate ventricular depolarization of the heart and are the most prominent: several analyses exploit the distribution of these peaks and the distances between successive R peaks, also known as RR intervals (RRi). This parameter, which shows an inverse correlation with stress, and its inverse, Heart Rate (HR), are the most widely adopted and straightforward measure to estimate stress level (Giannakakis et al., 2019). HRV, another ECG derived parameter, is the distribution of RR intervals [RRi, RRi+1,...] over a given time interval, and reflects the activity of the sympathetic and vagal components of the autonomic nervous system and is related to stress (T. F. o. t. E. S. o. C. t. N. A. S. o. P. Electrophysiology, 1996).

GSR, also known as ElectroDermal Activity (EDA) of the skin is another signal that can be monitored in a simple and non-invasive way to evaluate stress (Boucsein, 2012). It measures the electrical conductance of the skin, which is affected by the activity of sweat glands, activated during states of emotional arousal and anxiety, and found to be positively correlated with stress. GSR signal (see also Section 4) can be decomposed into two main components: a slower and non-specific tonic component due to the basic activity of the sweat glands and a faster stress-specific phasic phase (Zou and Ergan, 2021). A stress-induced stimulus creates a peak in the phasic component: significant features can be derived from this, such as rise time (Hoehn-Saric et al., 1989) which is inversely proportional to stress, the amplitude of the peak (Acerbi et al., 2017) and recovery time (Hoehn-Saric et al., 1989) (see Fig. 1).

Also, signals obtained from the electrical activity of the brain are useful for stress evaluation and stress monitoring. Brain signals obtained from EEG can be divided into different frequency bands: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–12 Hz), Beta (12–30 Hz), and Gamma (> 30 Hz). Each of these corresponds to different activity of the brain (Zulkurnaini et al., 2012; Buzsaki, 2006). Stress can disrupt the balance of these brainwaves, often leading to an overproduction

[3] https://www.statista.com/statistics/728584/respondents-affected-by-high-level-stress-in-us/

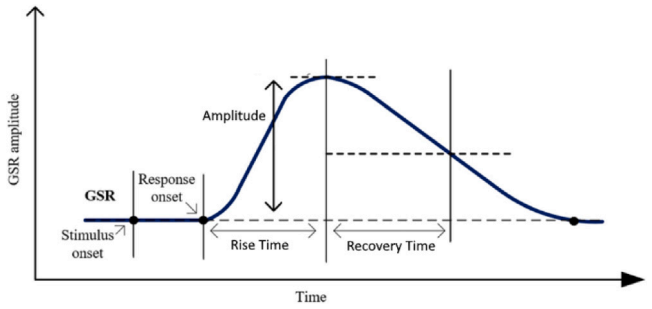


Fig. 1. GSR Phasic stress peaks with some of the main metrics reported: the `Rise Time`, that is the time interval between the beginning of the peak and its maximum point, the `Recovery Time`, namely the time of descent from the peak and the height of the peak or `Amplitude`.

of Beta (Chandra et al., 2017) and Theta wave activity (Ruo et al., 2022; Xavier et al., 2020) and a decrease in Alpha waves (Teo et al., 2020). The literature Hayashi et al. (2009) suggests that the Alpha activity in EEG signals may be reduced during stress conditions, while Beta activity may be increased: it follows that the $\frac{\beta}{\alpha}$ power ratio ($\frac{\beta}{\alpha}$ ratio) (Gatzke-Kopp et al., 2014) is considered as a measure of cognitive load associated with the arousal dimension and stress. Furthermore, EEG Alpha Asymmetry, which refers to the difference in power or amplitude of EEG signals in the Alpha frequency band between the two frontal lobe hemispheres, is a stress-related metric. In particular, an increase in Alpha activity of the left frontal hemisphere respect to the right is positively associated with stress (Berretz et al., 2022).

Furthermore, studies (Giannakakis et al., 2017) have shown that facial expressions can be used to detect stress and anxiety levels: for example, the corrugator supercilii muscle, which is responsible for frowning, is activated during stress conditions, while the activity of the zygomaticus major muscle, responsible of smiling, is decreased during stress conditions.

While objective parameters obtained from the analysis of physiological signals can provide analytical information, they may not capture the individual's subjective experience accurately: the two dimensions are complementary and cannot be separated from each other. Individual assessment questionnaires are essential to capture this aspect. The NASA Task Load Index (NASA-TLX) (Hart and , tlx) is a popular tool used to evaluate mental workload and stress: it is a subjective, multidimensional assessment tool that rates perceived stress to assess the effectiveness of a task, system, or team, as well as other aspects of performance. Other questionnaires widely used for the evaluation of subjective stress and mental workload are: the Workload Profile (WP) (Tsang and Velazquez, 1996), the Subjective Workload Assessment Technique (SWAT) (Vidulich and Tsang, 2012), and others (J.R. Crawford, 2004; Cohen et al., 1994; Holmes and Rahe, 1967).

2.2. Stress assessment in HRC

Recently, many works have been carried out, demonstrating the importance of stress assessment in the field of collaborative robotics. The study of Mariscal et al. (2023), mainly focusing on pupil diameter, aimed to assess occupational risks in terms of mental stress to determine whether a worker experiences greater stress when working in collaboration with a cobot rather than with another person while performing the same production-line process. The work of Zakeri et al. (2023) discusses the assessment of the mental workload of factory workers in an HRC environment, using EEG signals and subjective questionnaires. In the study by Carissoli et al. (2023), the researchers evaluated stress experienced by cobot programmers, analyzing EDA. Furthermore, the study of Pollak et al. (2020) focused on the levels of psychological (primary and secondary stress appraisal) and physiological (heart rate) stress in human operators working in two different cobot modes (i.e., manual and autonomous). In a recent study by Bussolan et al. (2023), a method similar to the one explored in this work was used to assess stress levels and cognitive load in operators engaged in collaborative robotics tasks, by analyzing cardiac, muscular and skin activity. The proposed study extends human modeling in by Bussolan et al. (2023) including also the analysis of brain activity and facial response, while focusing on collaborative tasks related to robot programming.

3. Material and methods

The experiment consisted in training subjects novice to cobot programming to perform a pick-and-place task. The study comprises three distinct stages, outlined extensively as follows: introduction to the learning materials, baseline measurement task, and hands-on practice. The educational framework has been meticulously designed to progressively furnish information essential for each phase of the tasks, facilitating an accessible learning experience for beginners in cobot programming. This incorporates informative slides, videos, and audio components. While they were exposed to the training program, their physiological activity was recorded and their subjective feedback was collected to assess the stress induced by HRC and robot learning. The recorded data are available in the publicly available dataset SenseCobot,[4] described in Borghi et al. (2024).

3.1. Participants

The experiment was conducted with $N = 21$ participants (17 male and 4 female), between the ages of 22 and 29 (M = 22.43 years, SD = 2.09). To simulate a real-world situation of a novice operator approaching a cobot, the only inclusion criterion was the absence of any prior experience in cobot programming (Borghi et al., 2024). Participants, all management and mechatronic engineering students, were recruited voluntarily with online scheduling and were instructed to avoid substances like coffee, nicotine, and alcohol on the day of the experiment, as these could potentially affect brain activity and stress-related physiological signals. Each participant filled in a consent form in which they authorized the use of their data for scientific research purposes.

For this particular experiment, a cobot UR10e equipped with a robotic gripper was used, but the procedure applies to other models and usage scenarios.

A learning platform was organized according to the previous work presented in Hansen et al. (2023), which aims at making cobot programming as easy as possible for novices. This previous work hypothesized that if unskilled workers were better informed about collaborative robots' functionalities and received systematic training, they could gain the ability to independently program cobots in an effective way. The platform provides information such as video, audio, and explanatory slides and guides learners through the different skills required for robot programming.

The workstation was arranged as depicted in Fig. 2. The attendee stood in front of a table, with a touchscreen monitor that had a webcam on the left side and the cobot on the right side. Three reference points (A, B, and C) and a box layout, necessary for the task, were marked on a cardboard on the table. The accessories necessary for task execution, together with a manual switch necessary for activating the free-drive mode, were placed within easy reach, to minimize operator's movements and distractions. The experiment was logically organized following a gradual progression of the complexity of the tasks, as described in Section 3.4.

[4] https://doi.org/10.5281/zenodo.8363762

3.2. Sensors and signals used

Different wearable devices were used to collect several physiological signals. The iMotions software platform[5] was used, since it facilitates data synchronization and organization, thanks to the integration of different communication protocols.

3.3. Experimental set-up

The sensors used in teh experimental set-up are shown in Fig. 3.

A Shimmer3 ECG[6] with a sampling rate of 512 Hz and equipped with five Ag/AgCl electrodes was used to measure the ECG. Moreover, a Shimmer3 GSR[7] with a sampling rate of 512 Hz was used to detect real-time GSR. For brain activity monitoring, an Enobio 20[8] EEG Helmet with 20 channel electrodes positioned according to the 10:20 standard (Ruo et al., 2022), organized in 17 EEG channels, 2 electrooculography (EOG) ocular electrodes, and a reference EXT channel, was used. The sampling rate of this device is 500 Hz. For recording facial expressions, the iMotions AFFDEX integrated toolkit was used (Bishay et al., 2022). This module is a convolutional and recurrent neural network trained for recording the 3D position of the user's head, detects the action units of the face, and recognizes the fundamental emotions (joy, surprise, fear, disgust, anger), plus neutral, sentimentality, attention, and confusion. AFFDEX measures accurately unfiltered and unbiased facial expressions of emotion, using just a standard webcam. In addition to physiological signals, other assessments were taken into consideration, such as time of task execution, number of errors, and programming mode used. More details can be found in Borghi et al. (2024).

3.4. Study design

A detailed description of the experiment can be found in the previous work (Borghi et al., 2024). The experimental setup was structured to include the 4 macro-phases described hereafter, and summarized in Fig. 4. The tasks are performed in succession, with a short break between them necessary for the participant to complete the questions of the subjective questionnaire (see Fig. 5).

1. **Introduction to learning materials** - This phase aimed to introduce the users to cobot programming using dedicated learning material and assigning four basic tasks to complete (Borghi et al., 2024). Participants were not bound by time constraints and were allowed to refer to the instructions as needed, with no monitoring of errors. The four tasks performed are the following:
    - **Task 1** = moving the robotic gripper of the cobot arm from point A to point B (see Section 2 in Fig. 2).
    - **Task 2** = directing the cobot's arm to reach three specified points (A, B, and C) (see Fig. 2).
    - **Task 3** = programming the cobot to reach an arbitrary height from point A(A') and then approach point A while maintaining its orientation perpendicular to the working surface. After a short pause, the cobot was to ascend back to point A'.
    - **Task 4** = programming the cobot to perform a pick-and-place operation in which the user had to use the teach pendant to pick up a screw placed at point A and move it to point B by activating the robotic gripper (see rectangular shape on Section 2 in Fig. 2).

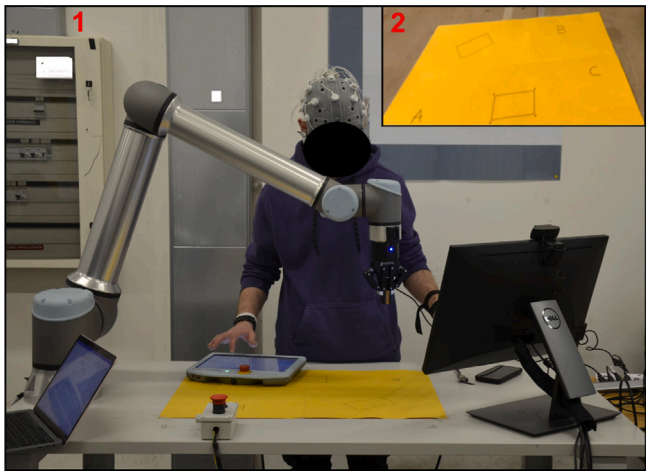


Fig. 2. Experimental set-up. The participant (1), dressed with sensors in front of the workbench, has the cobot on their right and the monitor with learning materials on their left. On Top right (2), a detail of the workbench with the three points A, B, C and the rectangular shape where to place the box.

2. **Sensors setup and baseline** - The primary purpose of the baseline was to objectify and standardize the psycho-physical conditions for each participant, establishing a reference point for comparing signals obtained during other tasks. This phase involved recording the participant's signals while they were positioned in front of a stationary blue screen for three minutes, in the absence of sounds or other visual stimuli.
3. **Hands-on practice** - Participants were assigned to perform five tasks, with the first four being the same as those of the "Introduction to learning materials" phase, plus one additional new task: Task 5. For Task 5, participants were tasked with programming a cobot to manipulate a box initially positioned at a 45° angle relative to the surface, ensuring precise placement onto a predetermined location on the plane. Following this, participants were required to guide the gripper robot along a predefined path at a consistent speed, simulating a gluing operation. This task encompassed all the previously acquired skills and presented a more intricate challenge. These tasks were structured in a sequence of increasing complexity, as shown in Fig. 4 to simulate the challenges one might encounter in a real work environment. In all cases, users had the option of moving the cobot arm, via the teach pendant, using either jogging mode or free-drive mode. Furthermore, no constraints have been defined for the user, to observe the possible number of reactions, errors, and strategies without imposing bias. The effect of the user's choice was documented for the potential assessment of stress variation, errors, or variations in psycho-physiological signals.
4. **Questionnaire** - After each task, participants were asked to fill a reduced version of the NASA-TLX questionnaire, expressing their perceived level of stress concerning only two categories: physical exertion and the overall effort. It is worthwhile noting that the NASA-TLX questionnaire was formerly introduced to measure the overall task load. However, in the proposed analysis it was used for stress evaluation, as in previous works such as Favre-Félix et al. (2022), Ghanavati et al. (2019), Said et al. (2020) and Bussolan et al. (2023). The evaluation of the subjective stress and mental workload was expressed using a 7-point Likert scale (1: "very low"; 7: "very high"). If the sum of the scores of the two questions used was lower than 7, the task was labeled as NO STRESS; otherwise, it was labeled as STRESS (Borghi et al., 2024).

[5] https://imotions.com/
[6] https://shimmersensing.com/product/shimmer3-ecg-unit-2/
[7] https://shimmersensing.com/product/shimmer3-gsr-unit/
[8] https://www.neuroelectrics.com/solution/software-integrations/nic2

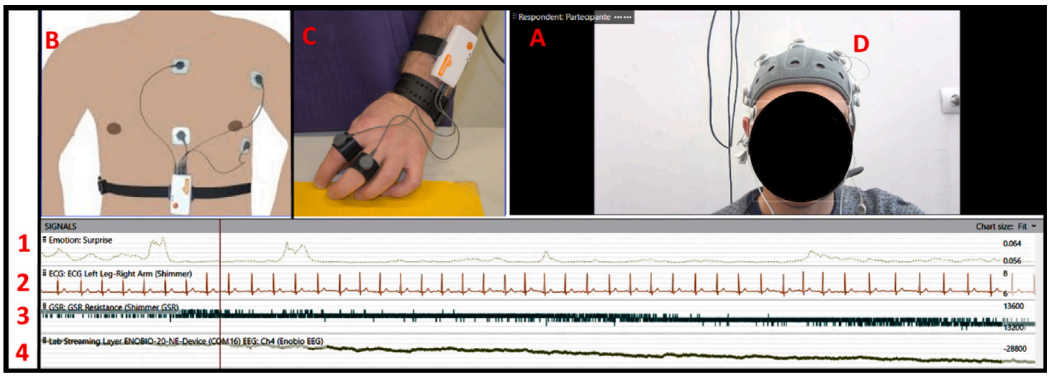


Fig. 3. The figure depicts the sensors utilized (denoted by the red letter) along with examples of their corresponding signals (indicated by the textcolorredred number). Specifically, A is associated with the iMotions AFFDEX module, and its corresponding signal is labeled as 1. B represents the Shimmer3 ECG sensor, and its ECG signal is identified as 2. C corresponds to the Shimmer3 GSR sensor, and its respective signal is marked as 3. Lastly, D showcases the Enobio20 EEG helmet, and the EEG signal it generates is denoted by the number 4.

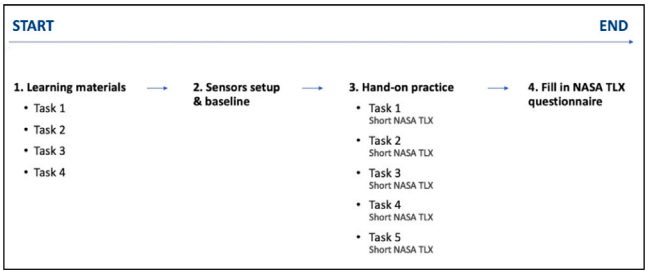


Fig. 4. Experimental protocol. Temporal succession of the phases of the experiment. As can be seen in the figure, at the initial stage (learning materials), the participant performs the 4 tasks without constraints. Then, the participant is dressed with the sensors and performs the Baseline (sensors set-up and baseline). Subsequently, the 4 previously performed tasks plus a new one (Task_5) are repeated for the participant dressed with sensors, each interspersed with a questionnaire. The experiment ends with a final complete NASA-TLX.

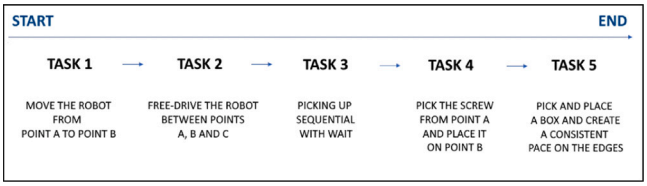


Fig. 5. A visual representation illustrating the chronological order of each task, accompanied by brief descriptions for clarity (refer to the video in the Zenodo repository for further details, available at https://doi.org/10.5281/zenodo.8363762, in the "Video_Tasks" folder).

4. Processing and analysis of recorded data

A signal processing pipeline was implemented in Python to extract the features and metrics that characterize stress, as indicated by the current body of literature (see Section 2.1). In particular, the Scipy library[9] was used for ECG signal processing, Neurokit 2 library[10] for GSR analysis, and MNE library[11] for processing EEG. Before the analysis of signals obtained and synchronized by iMotions, the timestamp was converted into a suitable format, and null values, gross recording errors, or files unusable for the study were eliminated. Due to recording errors, Participant 1's Baseline signal and Participant 10's entire ECG recording were discarded.

[9] https://docs.scipy.org
[10] https://neurokit2.readthedocs.io/en/legacy_docs/
[11] https://mne.tools/stable/index.html

The time windows chosen for metric extraction were set to a 10 s, an interval deemed sufficient to capture the slowest events (identified in this work as those from the GSR signal, according also with the literature Memar and Mokaribolhassan, 2021) and obtain meaningful HRV analysis (Melo et al., 2018).

4.1. ECG signal processing

As discussed in Section 2.1, the time intervals between consecutive R peaks of an ECG signal, known as RR intervals, and their derived measures are important parameters for stress detection. The steps performed to obtain this metric are described below:

1. **ECG data normalization:** The ECG signal was normalized by considering the 10th and 90th percentiles to mitigate subjective variations (Hong et al., 2020).
2. **ECG filtering:** To remove noise and artifacts corrupting the recorded signal,
   a band-pass filter with bandwidth from 10 Hz to 75 Hz was used (Tompkins, 1993).
3. **R peaks finding:** R peaks were detected using the "find peaks" function in the SciPy library. To discard erroneously detected peaks, the procedure in Kassinopoulos et al. (2021) was considered. Specifically, the fluctuations of the filtered ECG signal were smoothed by applying a Moving Average technique, with a window set to 1000 samples, and subsequently, a threshold, corresponding to one standard deviation above the moving average, was determined (Kassinopoulos et al., 2021).
4. **RR intervals and HR calculation:** RR intervals between successive heartbeats were calculated by determining the differences between consecutive R-peaks. The instantaneous HR could then be derived, resulting in the number of beats per minute (Mayapur, 2018).
5. **HRV metrics extraction:** Some of the main metrics related to HRV, such as PNN25, PNN50, root mean square of successive differences (RMSSD), and RR intervals standard deviation (SDNN) were derived (Kleiger et al., 2005; Shaffer and Ginsberg, 2017; Villani et al., 2020)). In particular, SDNN was calculated as standard deviation of RR intervals, RMSSD as mean of the squared differences between successive RR intervals in squared root. PNN25 and PNN50 were obtained as the percentage of consecutive RR intervals differences that are greater than 25 or 50 ms respectively.

As a result, the ECG metrics used for subsequent analysis are RR Intervals expressed in ms and HRV metrics PNN25, PNN50, RMSSD, and SDNN.

4.2. GSR signal processing

For the extraction of significant metrics from GSR, the following steps were implemented, using the Neurokit 2 library.

1. **GSR filtering:** To enhance the signal-to-noise ratio of the recorded GSR signal, the `nk.eda_process` module was used. It consists in a two-step filtering process. Firstly, the high-frequency noise is removed. Secondly, since the sensor is mou nted on the wrist during the execution of a dynamic task, motion artifacts are removed.
2. **Tonic and phasic components extraction:** tonic and phasic components of the GSR signal were extracted by a dedicated data processing, designed according to the model proposed by Aqajari et al. (2021). This was done by the Neurokit 2 `nk.eda_process` module.
3. **GSR features extraction:** Finally, parameter analysis and peak identification were executed with the `nk.eda_peaks` module.

The GSR metrics used for subsequent analysis are `Amplitude` (amplitude of the peaks in μS), `Rise Time` and `Recovery Time` (each of these in ms).

4.3. EEG signal processing

MNE Python library offers a wide range of modules for detailed analysis of the EEG signal. Based on the consideration of Section 2.1, the frequency bands that are commonly used for stress analysis are Theta, Alpha, and Beta. EEG processing was organized in the following manner:

1. **EEG signal pre-processing:** Recorded data were pre-processed by discarding the beginning and end of the recording signal (5 s) and downsampled from 500 Hz to 250 Hz. Channels related to EOG, necessary for detecting eye movement artifacts but useless for analysis, were then discarded.
2. **EEG filtering:** A band-pass filter with low cut-off frequency of 1 Hz and high cut-off frequency of 30 Hz was applied to remove noise and artifacts, according to previous works (Karpiel et al., 2021).
3. **Bad channels and artifacts removal:** Channels that exhibit statistically deviant behavior compared to the other channels were identified as bad channels and removed (Alotaiby et al., 2015). These channels exhibit very different behavior due to loss of electrode contact or other artifacts. In this work, the selection of bad channels was carried out with the kurtosis method of the MNE library. To carry out statistical comparisons, the following approach was considered: if the same channels were identified as bad for at least four of the five subsequent tasks, including the baseline task, they were removed. Subsequently, any movement-related artifacts, including ocular artifacts, were removed using the Independent Component Analysis (ICA) method (Hyvärinen, 2013) from the remaining channels.
4. **Band power analysis:** The script used calculates band power by applying the Welch method to compute Power Spectral Density (PSD) of the frequency bands of interest (Theta, Alpha, and Beta) using the trapezoidal numerical integration method for each 5-second interval in the selected channels (Huberty et al., 2021).
5. **EEG features extraction:** The ratio between Beta and Alpha components within the same channel was computed, together with the Asymmetry of Alpha between electrodes placed in symmetrical areas of the head. As for the findings reported in previous studies (Raufi and Longo, 2022; Fuentes-García et al., 2020; Holm et al., 2009) regarding the Theta channel selection, the Power Spectrum for the specified frontal and temporal channels, namely F7, F3, F4, F8, Fz, T7, and T8, was averaged. Similarly, for the Alpha and Beta power spectra, the channels P7, P3, P4, P8, Pz, O1, O2, and F7, F3, F4, F8, Fz were chosen, respectively (see Table 1). The channels used in the evaluation of Alpha Asymmetry are two frontal channels F3 and F4. Alpha Asymmetry was calculated according to Giannakakis et al. (2017) as the subtraction of the alpha's power (α) natural logarithm of the left hemisphere channel (P3) from that of the right hemisphere (P4), as provided by the following equation

$$\text{Alpha Asymmetry} = \ln(\alpha_{P3}) - \ln(\alpha_{P4}) \quad (1)$$

Table 1
Frequency Bands and Corresponding Cortical Areas and Selected Channels.

Frequency Bands	Cortical Area	Channels
Theta	Frontal and Temporal	F7 - F3 - F4 - F8 - Fz - T7 - T8
Alpha	Parietal and Occipital	P7 - P3 - P4 - P8 - Pz - O1 - O2
Beta	Frontal	F7 - F3 - F4 - F8 - Fz

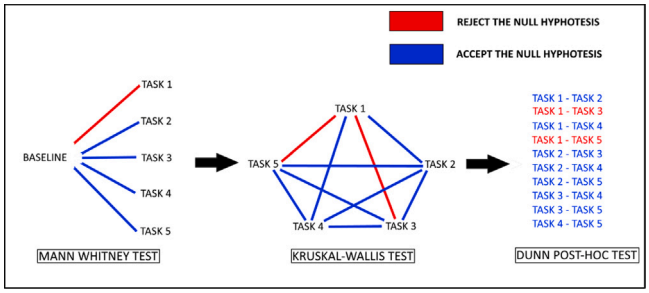


Fig. 6. Succession of statistical tests performed for each metric extracted from each signal. In blue, the statistical test accepts the null hypothesis: samples confronted are similar. In red, the statistical test rejects the null hypothesis: samples confronted are different.

Most research findings indicate that during times of stress, there tends to be increased alpha activity in the right frontal region compared to the left (Acharya et al., 2012), resulting in a negative value of alpha asymmetry.

As a result, the EEG metrics used for subsequent analysis are `Mean Alpha Power`, `Mean Beta Power`, `Mean Theta Power`, `Mean Beta/Alpha Power Ratio` and `Alpha Asymmetry`.

4.4. Emotions processing

The raw AFFDEX module output of the iMotions software is a time series of the selected emotions: anger, disgust, fear, joy, sadness, and surprise. The algorithm implemented for analyzing emotions was structured in the following way:

1. **Data normalization:** Each column of the raw output was normalized by dividing the absolute value of each data point by the corresponding maximum value so that the signals vary between 0 and 1.
2. **Emotions features extraction:** For each frame of registration the dominant emotion recorded was selected, to have a time-dependent evolution of emotions.

5. Methodology for statistical analysis

The main objective of the statistical analysis was to evaluate whether a significant difference among the psycho-physiological parameters collected during task execution can be observed if compared with the baseline.

To this end, the following research questions were considered:

Q1: Is the physiological response coherent with task complexity?

Q2: Is the physiological response strongly individual?
Q3: Are variations in the physiological response coherent with variations in the subjective perceptions?

To answer the research questions, we considered the parametric t-test or non-parametric Mann Whitney test, parametric ANOVA or non-parametric Kruskal Wallis test, and parametric Tukey or non-parametric Dunn Post-hoc test (See Fig. 6). The tests used were all non-parametric, since the samples have different numbers of elements, and the condition of normality was not verified for all the metrics of interest. Moreover, the various tasks in the experimental protocol, as described in Section 3.3, showed a progressively increasing level of complexity and the same sequence was maintained for all participants. As a result, statistical independence of the samples was not presumed for this experiment. For all statistical tests carried out for analysis, two significance values (α) of 0.05 and 0.01 were used.

The statistical analysis is reported in detail below:

- **Normality test:** The data in each sample were tested using the D'Agostino and Pearson normality test with $\alpha = 0.05$, to determine if they follow a normal distribution (Ghasemi and Zahediasl, 2012). This test was selected because of its robustness and precision even on modest-sized samples. This step was crucial and a requirement for subsequent analyses.
- **Comparison of physiological parameters between tasks and baseline:** During the execution of the five tasks, the signals of individual participants were statistically compared to their respective baseline (See Fig. 6). If the normality of the data was confirmed, a parametric t-test was conducted, otherwise, a non-parametric Mann–Whitney test with the same value was conducted instead (Fay and Proschan, 2010). The objective was to verify if the psycho-physiological parameters, as expected, were different compared to the resting condition.
- **Differences between tasks:** In order to reply to research question **Q1**, the study has to verify if the metrics extracted were different between tasks. If data were normally distributed, a parametric ANOVA test was used, otherwise, a non-parametric Kruskal–Wallis test (See Fig. 6); each of these tests can be applied to samples of a number greater than two (Corder and Foreman, 2011).
- **Dunn Post Hoc test with Bonferroni Correction:** In order to reply to question **Q2**, the stress response among individuals has to be investigated, for each task. If the Kruskal–Wallis test yielded a rejection for $\alpha = 0.05$, that is, the compared samples are different, a subsequent parametric Tukey test or non-parametric Dunn test was conducted to assess variations among individual pairs of tasks, distinguishing those that exhibited differences from those that did not (Dinno, 2015) (See Fig. 6).
- **Comparison with Subjective Evaluations:** In order to reply to research question **Q3**, the outcomes of the statistical analysis with the subjective assessments derived from the NASA-TLX questionnaire: this facilitated the harmonization of quantifiable physiological data with the participant's perspective. The objective was to evaluate how many times a variation detected in statistical analysis corresponds to a variation in subjective perception. This gives us information on the percentage of agreement between the observed stress-related physiological parameter and the subjective perception.

Statistical analysis was carried out using ad-hoc Python programs and Python standard libraries such as Scipy,[12] Scikit-Learn.[13] and Statsmodels[14]

[12] https://docs.scipy.org
[13] https://scikit-learn.org/stable/
[14] https://www.statsmodels.org/stable/index.html

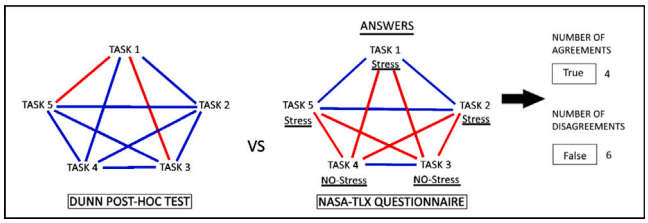


Fig. 7. Schematic comparison between the statistical result obtained during the execution of the Dunn's Post Hoc test and the subjective responses from the NASA-TLX questionnaire. In blue, samples confronted with the Dunn Statistical test are similar, and answers from NASA-TLX are the same. In red, samples confronted with the Dunn Statistical test are different, as are the answers from NASA-TLX. On the right, in **NUMBER OF AGREEMENTS**, **TRUE**, represents the number of times in which a difference recorded by the statistical test corresponds to a difference observed in the questionnaire answers, **FALSE** otherwise.

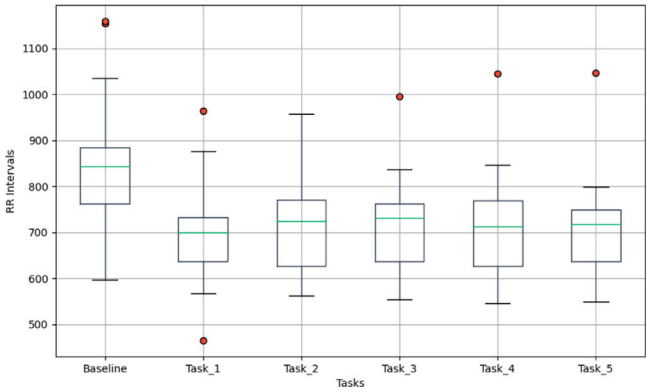


Fig. 8. Box plot of ECG RR `Intervals` for baseline and the different robot programming tasks.

6. Results

Table 2 reports mean and standard deviation values for RR interval, HRV, GSR, EEG, and emotions metrics. An overall discussion is provided in Section 7. To understand how such values are representative of stress conditions, we carried out the statistical tests described in Section 5. In the remaining of this section, we report the main results achieved for the different physiological signals under analysis.

6.1. ECG analysis

The statistical analyses described in Section 5 were carried out on the RR intervals and HRV metrics obtained from ECG signals recorded during baseline and robot programming tasks, as described in Section 4.1. For ECG signals, data from two subjects were discarded due to poor quality of recorded data (Borghi et al., 2024), thus resulting in a total valid sample size of 19 subjects.

Fig. 8 reports the box plot of RR `Intervals` recorded for all 19 valid samples. Moreover, Table 3 summarizes the number of subjects for which a statistically significant difference was observed for RR `Intervals` among baseline and the robot programming tasks, while Table 4 reports the same type of results for `PNN25`, `PNN50`, `RMSSD`, and `SDNN`. Specifically, the first set of rows reports the results of the Mann–Whitney test, which was used to investigate differences in the pairs consisting of baseline and each robot programming task. As shown in Table 3, a statistically significant difference for RR `Intervals` is observed for all the considered 19 test participants, such as to reject the null hypothesis of similarity of sample distribution, while this number decreases in Table 4 for HRV metrics. In summary, a significant difference compared to baseline situation is observed for RR `Intervals`

Table 2
Mean (**Mean**) and standard deviation (**Std**) of all the considered metrics for baseline and programming tasks. RR Intervals, SDNN, RMSSD, Rise Time and Recovery Time are in ms, Alpha Power, Beta Power and Theta Power are in $\frac{V^2}{Hz}$, Amplitude is in μS.

Tasks	ECG Metrics	Mean	Std	EEG Metrics	Mean	Std	GSR Metrics	Mean	Std
Baseline	RR intervals	947.50	240.33	Alpha Power	11.90	7.72	Rise Time	271.55	17.40
Task 1		799.62	222.71		14.70	5.65		243.99	23.50
Task 2		831.03	253.39		16.49	7.33		246.50	37.37
Task 3		816.35	248.16		17.07	8.02		240.92	27.72
Task 4		826.06	269.43		15.75	8.24		268.64	27.70
Task 5		812.72	257.86		16.40	7.03		282.10	19.64
Baseline	PNN25	83.56	10.13	Beta Power	21.89	11.34	Recovery Time	323.34	38.92
Task 1		73.44	18.62		26.70	12.64		320.97	45.59
Task 2		74.95	17.24		27.97	10.42		312.92	46.27
Task 3		75.24	17.29		26.43	10.17		309.73	47.50
Task 4		73.61	17.86		31.14	14.42		336.06	47.76
Task 5		72.91	18.64		33.18	11.98		378.75	38.20
Baseline	PNN50	77.53	13.21	Theta Power	35.92	30.61	Amplitude	7.19	20.4
Task 1		65.63	20.26		50.20	31.77		40.3	48.8
Task 2		67.22	18.67		55.70	35.68		85.1	159
Task 3		68.12	19.75		56.10	35.48		94.4	157
Task 4		66.16	19.85		47.58	31.44		24.4	26.6
Task 5		65.38	20.80		45.49	24.92		15.2	20.9
Baseline	SDNN	37.35	17.14	Beta/Alpha Ratio	2.58	1.78			
Task 1		53.83	41.66		1.96	0.65			
Task 2		55.14	43.36		1.98	0.82			
Task 3		52.52	38.07		1.84	0.83			
Task 4		43.65	31.77		2.19	0.79			
Task 5		44.91	28.41		2.31	0.76			
Baseline	RMSSD	38.80	23.39	Alpha Asymmetry	−0.11	0.37			
Task 1		54.16	49.20		−0.15	0.38			
Task 2		54.15	49.26		−0.16	0.40			
Task 3		52.35	44.28		−0.16	0.37			
Task 4		42.23	37.22		−0.24	0.47			
Task 5		43.41	34.22		−0.04	0.34			

Table 3
Statistical tests for RR intervals. The columns on the right report the number of test participants for which a statistically significant difference was observed among the different conditions for $\alpha = 0.05$ and $\alpha = 0.01$. Rows in blue denote consecutive tests. Total sample size = 19.

Test	Conditions	$\alpha = 0.05$	$\alpha = 0.01$
Mann–Whitney	Task 1 - Baseline	19	19
	Task 2 - Baseline	19	19
	Task 3 - Baseline	19	19
	Task 4 - Baseline	19	19
	Task 5 - Baseline	19	19
Kruskal–Wallis	All Tasks	19	19
Dunn Post-Hoc	Task 1 - Task 2	9	9
	Task 1 - Task 3	13	11
	Task 1 - Task 4	16	15
	Task 1 - Task 5	16	16
	Task 2 - Task 3	11	10
	Task 2 - Task 4	14	14
	Task 2 - Task 5	13	11
	Task 3 - Task 4	9	9
	Task 3 - Task 5	13	11
	Task 4 - Task 5	12	11

for all the considered subjects, while less average variability between tasks and baseline is found for HRV metrics. This is confirmed by Fig. 8 and Table 2. As shown in Fig. 8, net of some outliers, the baseline has the greatest median value of RR Intervals, while RR Intervals during the programming tasks are, on average, lower. This is indicative of an increase in the heart rate and greater activation of the subject. Furthermore, it is observed that the mean values of RR Intervals are lower in Task 1, Task 4, and Task 5; in particular, Task 5 seems to be the one with a smaller range of values. Moreover, when excluding the baseline, the most extensive range of values is observed in Task 1 and Task 2, suggesting a higher variability in the participant's responses, and, regarding the difference between the first and third quartile, the greater distribution of RR interval values is observed in Task 2 and Task 4.

To investigate further in detail the differences among the different tasks, we carried out the non-parametric Kruskal–Wallis test (second set of rows in Tables 3 and 4). Indeed, this test is performed on more than two sample populations to determine if a significant difference in data distribution exists. As demonstrated in Table 3, for all 19 subjects in our study, a statistically significant difference is observed among the distribution of RR Intervals collected during the five robot programming tasks, regardless of baseline values. This number decreases for HRV related metrics, as shown in Table 4. It appears that the distribution of raw RR Intervals shows a greater difference between tasks, while HRV metrics show less variability. However, this test does not provide insights about the differences within the tasks. To this end, we performed a Dunn Post-Hoc test, to assess which pair of conditions exhibited a statistically significant difference. The last set of rows (Dunn Post-Hoc) in Tables 3, 4 reports, for all the possible pairs of robot programming tasks, the number of subjects for which statistically significant differences were found with this test. In the tables, we highlight in blue the pairs referred to conditions administered in sequence. The results show that RR Intervals in such pairs of consecutive tasks were more similar than when comparing tasks administered not in succession: indeed, the tasks were designed with increasing complexity. This observation is not always verified for HRV metrics (see Table 4).

Finally, we compared the distribution of RR intervals and HRV metrics with the results of NASA-TLX subjective questionnaires, to assess whether they agree or not. In particular, following Fig. 7, we computed how often the Dunn's Post Hoc test for each metrics and the replies to the NASA-TLX questionnaire detected a change in the stress level (from "STRESS" to "NO-STRESS" and vice versa) or the same stress level between any pair of tasks. Results are shown in Table 5, respectively, where we report the average level of agreement between statistical tests and subjective questionnaire. These results show how

Table 4
Statistical tests for PNN25, PNN50, RMSSD, and SDNN. The columns on the right report the number of test participants for which a statistically significant difference was observed among the different conditions for $\alpha = 0.05$ and $\alpha = 0.01$. Rows in blue denote consecutive tests. Total sample size = 19.

		PNN25		PNN50		RMSSD		SDNN	
Test	Conditions	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Mann–Whitney	Task 1 - Baseline	6	6	10	7	5	2	3	2
	Task 2 - Baseline	8	5	7	5	6	5	3	1
	Task 3 - Baseline	8	4	7	5	7	7	8	4
	Task 4 - Baseline	11	7	10	8	7	7	7	5
	Task 5 - Baseline	10	7	11	9	7	6	6	5
Kruskal–Wallis	All Tasks	5	5	5	5	4	4	6	6
Dunn Post-Hoc	Task 1 - Task 2	2	2	1	1	0	0	2	2
	Task 1 - Task 3	2	2	3	3	0	0	2	2
	Task 1 - Task 4	2	2	1	1	0	0	2	2
	Task 1 - Task 5	3	3	4	4	1	1	3	3
	Task 2 - Task 3	1	1	3	3	2	2	3	3
	Task 2 - Task 4	1	1	1	1	0	0	1	1
	Task 2 - Task 5	1	1	0	0	0	0	2	2
	Task 3 - Task 4	1	1	1	1	0	0	1	1
	Task 3 - Task 5	1	1	2	2	0	0	3	3
	Task 4 - Task 5	0	0	1	1	0	0	3	3

Table 5
Agreement between RR Intervals, PNN25, PNN50, RMSSD, and SDNN with subjective responses from NASA-TLX, when comparing tasks. Values are in percentage over 19 subjects.

Conditions	Agreement %				
	RR Intervals	PNN25	PNN50	RMSSD	SDNN
Task 1 - Task 2	47.37	78.95	84.21	84.21	89.47
Task 1 - Task 3	36.84	78.95	73.68	73.68	78.95
Task 1 - Task 4	31.58	63.16	57.89	57.89	63.16
Task 1 - Task 5	36.84	63.16	57.89	52.63	52.63
Task 2 - Task 3	52.63	84.21	73.68	68.42	73.68
Task 2 - Task 4	36.84	68.42	68.42	63.16	68.42
Task 2 - Task 5	47.37	52.63	57.89	52.63	68.42
Task 3 - Task 4	42.11	78.95	78.95	73.68	78.95
Task 3 - Task 5	57.89	63.16	57.89	63.16	63.16
Task 4 - Task 5	31.58	84.21	78.95	78.95	68.42
Average	**42.11**	**71.58**	**68.95**	**66.84**	**70.53**

changes in the considered metrics were consistent with self-reported stress levels. In particular, the table shows that HRV metrics exhibit greater agreement with self-reported stress levels than RR Intervals.

6.2. GSR analysis

As described in depth in Section 4.2, the metrics used for statistical analysis of the GSR signal were Rise Time, Recovery Time, and Amplitude. The same analysis procedure as that reported in Section 6.1 was carried out on the GSR data of all the 21 test participants: Mann–Whitney test for pairwise comparison between baseline and each robot programming task, Kruskal–Wallis test among all the robot programming tasks and the Dunn Post-Hoc test for multiple pairwise comparisons between all pairs of robot programming tasks. The results are reported in Table 6, which shows that Amplitude is the metrics with greater variability compared to the rest situation.

Finally, the percentage of agreement between the Dunn statistics tests performed on Rise Time, Recovery Time, and Amplitude and the respective subjective questionnaires are reported in Table 7. The Rise Time and Recovery Time metrics show the highest percentage of agreement with subjective perception (NASA-TLX), on average above 70%. Conversely, poor agreement is found for Amplitude, although this is the metrics exhibiting the greatest differences among the different experimental tasks and baseline.

6.3. EEG analysis

As cited in Section 1, the metrics obtained from EEG signals were Mean Alpha Power, Mean Beta Power, Mean Theta Power, Mean Beta/Alpha Power Ratio and Alpha Asymmetry between channels F3-F4.

summarizes the results of the statistical analyses reported in Section 5, namely the Mann–Whitney, Kruskal–Wallis and Dunn Post-Hoc tests. The table shows that most differences were found between all the programming tasks and the baseline condition for all the metrics, while Alpha Asymmetry was the metrics that highlighted differences among the programming tasks for some subjects.

The comparisons of the brain parameters mean Alpha, Beta, and Theta Power spectrum, Alpha Asymmetry, and mean Beta to Alpha ratio of the power spectrum and the responses from the NASA-TLX questionnaires are reported in Table 9. The metrics derived from the EEG are among the most precise in describing and predicting the individual stress state, with accuracy percentages always above 50% and with Theta Power Mean close to 80%.

6.4. Analysis of emotional response

Statistical tests were performed to evaluate emotional responses during the execution of tasks. AFFDEX module was used to extract the main emotions: joy, anger, sadness, surprise, fear, and disgust. Specifically, the software provides the likelihood of the different emotions for each video frame, as described in Borghi et al. (2024).

The results of the statistical analysis are reported in Table 10. A multivariate analysis of variances (MANOVA) (Warne, 2019) was performed on the raw time series on the emotions recorded during the tasks. Before using this test, the requirements of normality of the data distribution and homomorphy of the covariance matrices were verified. The MANOVA returned that there is a noticeable distinction between the tasks for all 21 participants. Following the use of the MANOVA test, for a more detailed statistical analysis of emotion signals, a Chi-square test for consistency (Aslam and Smarandache, 2023) was conducted to compare the distribution of emotions for each participant in each task with their corresponding baseline. A Chi-square test for consistency (Aslam and Smarandache, 2023) was conducted on the emotion distributions across the five tasks for each participant to determine whether there was a significant difference among them. Finally, a Bonferroni Post-Hoc test (Aslam and Albassam, 2020) was performed to evaluate those pairs of tasks that showed more differences. As a last result, the results of the comparisons with the NASA-TLX questionnaires are shown in Table 11.

7. Discussion

Fig. 9 presents an overview of the metrics obtained from the analysis of psycho-physiological signals and summarizes the agreement (in percentage) between subjective response and quantitative metrics derived

Table 6
Statistical tests for GSR data. The columns on the right report the number of test participants for which a statistically significant difference was observed among the different conditions for $\alpha = 0.05$ and $\alpha = 0.01$. Rows in blue denote consecutive tests. Total sample size = 21.

		Rise Time		Recovery Time		Amplitude	
Test	Conditions	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Mann–Whitney	Task 1 - Baseline	8	6	5	2	20	19
	Task 2 - Baseline	11	10	5	4	19	19
	Task 3 - Baseline	11	7	7	4	19	19
	Task 4 - Baseline	8	4	2	2	20	18
	Task 5 - Baseline	10	6	5	2	19	19
Kruskal–Wallis	All tasks	7	4	6	5	21	21
Dunn Post-Hoc	Task 1 - Task 2	7	3	1	1	12	12
	Task 1 - Task 3	3	2	1	1	17	17
	Task 1 - Task 4	5	4	1	1	16	14
	Task 1 - Task 5	5	1	1	1	17	13
	Task 2 - Task 3	1	1	2	1	18	17
	Task 2 - Task 4	5	4	2	2	16	16
	Task 2 - Task 5	5	4	3	2	18	17
	Task 3 - Task 4	4	4	1	0	16	14
	Task 3 - Task 5	3	3	1	0	16	16
	Task 4 - Task 5	1	1	1	1	12	10

Table 7
Agreement between GSR metrics and subjective responses from NASA-TLX, when comparing tasks. Values are in percentage over 21 subjects.

Conditions	Agreement %		
	Rise Time	Recovery Time	Amplitude
Task 1 - Task 2	71.43	76.19	42.86
Task 1 - Task 3	71.43	66.67	42.86
Task 1 - Task 4	57.14	57.14	61.90
Task 1 - Task 5	57.14	66.67	47.62
Task 2 - Task 3	80.95	80.95	28.57
Task 2 - Task 4	76.19	80.95	33.33
Task 2 - Task 5	80.95	71.43	38.10
Task 3 - Task 4	71.43	76.19	33.33
Task 3 - Task 5	71.43	71.43	42.86
Task 4 - Task 5	76.19	80.95	66.67
Average	**71.43**	**72.86**	**43.81**

from physiological signals and estimated emotions. The metrics were sorted from left to right following an increasing order of percentage of agreement between the metric and the subjective evaluation. So in the rightmost part, there are the metrics that demonstrate a higher correlation with subjective evaluations of stress from the NASA-TLX questionnaire.

7.1. ECG discussion

The statistical examination of cardiac activity in terms of RR intervals and HRV metrics suggests some interesting findings. From a statistical perspective, the raw RR Intervals values sampled during task execution are notably and consistently lower than those recorded during rest conditions or baseline situations for all participants (see Table 2 and the results of the Mann–Whitney test in Table 3). This stress-related parameter decreases during task execution, suggesting a greater state of emotional arousal. Note that, as defined in the research question **Q2**, cardiac parameters, as well as individual perceptions, are highly variable between subjects (see Fig. 8). Furthermore, for all the samples analyzed, a statistically significant variation in the RR distribution is observed during different tasks (see Table 3), with the lowest values in Task 1, Task 4, and Task 5 (see Fig. 8). The last two are the most complex tasks, while the first could represent a certain anxiety in approaching the cobot (Fig. 8). These results are in agreement with the research question **Q1**. In Table 3 the Dunn Post-Hoc test indicates lower values in the comparison between Task 1 vs Task 2, Task 2 vs Task 3, Task 3 vs Task 4, and Task 4 vs Task 5, all of which are consecutive tasks. This implies that the instances where a statistically significant difference is observed in the RR Intervals parameter are fewer than those observed in non-consecutive tasks. The parameters appear to exhibit similarity, demonstrating temporal inertia, where preceding events seem to influence subsequent ones, irrespective of task complexity (Table 5).

In research, it is essential to assess the alignment between objective evaluations of stress-related parameters and subjective perceptions. The comparison of results from the Dunn test and NASA-TLX responses, specifically regarding RR Intervals distribution in the given experimental setup, reveals an agreement of 47.5%. The limited correlation observed between cardiac metrics and subjective perceptions appears to validate the constraints associated with relying solely on subjective assessments to describe stress (Föhr et al., 2015). Such assessments are heavily influenced by cultural and personal biases, as well as various objective and subjective variables.

HRV metrics, which are widely used in the literature to assess stress states, show lower differences than raw RR intervals between tasks and the baseline, as well as between different tasks (see Table 4). The HRV metrics decrease also during tasks compared with baseline, suggesting an higher stress level (see Table 2) . However, as illustrated in Fig. 9, PNN25, PNN50, RMSSD, and SDNN demonstrate strong precision, consistently aligning with subjective stress evaluations at a rate of 65% to 75%.

7.2. GSR discussion

From the statistical analysis of GSR derived metrics in Table 6, it is clear that Amplitude is the parameter that shows more differences between the robot programming tasks and the rest condition, unlike Rise Time and Recovery Time, which appear to be less prone to variations compared when compared to a baseline scenario. Indeed, considering the Kruskal–Wallis outcomes with $\alpha = 0.01$, statistically significant differences were observed in 4 out of 21 samples for RiseRise Time, 5 out of 21 for Recovery Time and 21 out of 21 for Amplitude. Again with reference to Table 6, the outcomes of the Dunn's tests performed on samples with significant Kruskal–Wallis test results appear to validate the observations made in the ECG analysis: the metrics exhibit reduced variability throughout the execution of consecutive tasks, as indicated by the lower incidence of significant results in subsequent tests. Table 2 highlights that Recovery Time, a GSR metric directly related to stress, shows a positive increase during the evolution of tasks, with greater values on Task 4 and Task 5. Amplitude, shows greater values during task execution, while Rise Time, a metric inversely related to stress, has values that tend to be lower during activities compared to the baseline level.

Comparing the results obtained from the variation in parameters detected by the Dunn Post Hoc test and the NASA-TLX responses, we

Table 8
Statistical tests for EEG data. For each metrics, the columns on the right report the number of test participants for which a statistically significant difference was observed among the different conditions for $\alpha = 0.05$ and $\alpha = 0.01$. Rows in blue denote consecutive tests. Total sample size = 21.

		Mean Alpha Power		Mean Beta Power		Mean Theta Power	
Test	Conditions	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Mann–Whitney	Task 1 - Baseline	10	10	7	6	6	5
	Task 2 - Baseline	8	7	10	8	6	3
	Task 3 - Baseline	8	7	11	9	10	8
	Task 4 - Baseline	13	12	16	15	11	11
	Task 5 - Baseline	14	13	17	17	14	12
Kruskal–Wallis	All tasks	12	11	14	13	7	2
Dunn Post-Hoc	Task 1 - Task 2	1	0	2	1	0	0
	Task 1 - Task 3	0	0	3	1	0	0
	Task 1 - Task 4	4	2	3	2	1	1
	Task 1 - Task 5	4	2	4	2	1	1
	Task 2 - Task 3	1	0	2	2	0	0
	Task 2 - Task 4	2	1	3	2	0	0
	Task 2 - Task 5	5	3	4	2	1	1
	Task 3 - Task 4	0	0	3	2	0	0
	Task 3 - Task 5	5	3	7	6	1	1
	Task 4 - Task 5	4	2	4	4	1	0

		Mean Beta/Alpha Power Ratio		Alpha Asymmetry	
Test	Conditions	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Mann–Whitney	Task 1 - Baseline	8	6	8	7
	Task 2 - Baseline	8	8	6	6
	Task 3 - Baseline	8	8	6	5
	Task 4 - Baseline	15	12	11	10
	Task 5 - Baseline	13	12	8	8
Kruskal–Wallis	All tasks	10	10	12	12
Dunn Post-Hoc	Task 1 - Task 2	2	1	3	3
	Task 1 - Task 3	3	2	2	2
	Task 1 - Task 4	5	4	8	7
	Task 1 - Task 5	4	3	4	3
	Task 2 - Task 3	3	3	4	4
	Task 2 - Task 4	5	4	9	8
	Task 2 - Task 5	3	3	8	7
	Task 3 - Task 4	1	0	7	7
	Task 3 - Task 5	5	5	7	5
	Task 4 - Task 5	3	2	9	8

Table 9
Agreement between EEG metrics and subjective responses from NASA-TLX, when comparing tasks. Values are in percentage over 21 subjects.

Conditions	Agreement %				
	Mean Alpha Power	Mean Beta Power	Mean Theta Power	Mean Beta/Alpha Power Ratio	Alpha Asymmetry
Task 1 - Task 2	89.47	82.35	80.00	78.57	75.00
Task 1 - Task 3	73.68	58.82	73.30	64.29	58.33
Task 1 - Task 4	63.16	70.59	66.70	64.29	58.33
Task 1 - Task 5	57.89	58.82	60.00	64.29	83.33
Task 2 - Task 3	73.68	70.59	80.00	57.14	58.33
Task 2 - Task 4	78.95	76.47	80.00	57.14	50.00
Task 2 - Task 5	57.89	52.94	80.00	50.00	50.00
Task 3 - Task 4	78.95	82.35	86.70	78.57	41.67
Task 3 - Task 5	57.89	52.94	73.30	57.14	66.67
Task 4 - Task 5	73.68	70.59	86.70	71.43	33.33
Average	**70.53**	**67.65**	**76.67**	**64.29**	**57.50**

observe agreement percentages of 71.43% for `Rise Time`, 72.86% for `Recovery Time`, and 43.81% for `Amplitude`. As Fig. 9 shows, `Rise Time` and `Recovery Time` metrics stand out as excellent indicators for predicting and characterizing an individual's perception of stress.

7.3. EEG discussion

The comprehensive statistical analysis conducted using metrics derived from brain signals proves to be some of the most accurate methods for evaluating perceived subjective stress within the scope of this cobot programming study, with an agreement percentage for `Theta Power Mean`, `Mean Alpha Power`, and `Mean Beta Power` respectively of 76.67%, 70.53%, and 67.65% (see Fig. 9). The trend and evolution of these metrics during the evolution of the tasks are reported in Table 2. The results of the Mann–Whitney test in indicate that the derived EEG parameters show lower variability compared with the baseline, in contrast to the cardiac parameter. Furthermore, as discussed in research question **Q2**, there is variability in cerebral parameters across tasks, as evidenced by Kruskal–Wallis tests. This suggests distinct mental demands associated with tasks of varying complexity. The application of Dunn's Post Hoc test to brain parameters provides additional support for the hypothesis that there is a temporal dependence in the biological signal between tasks. This is evidenced by the lower significance values observed when comparing consecutive task.

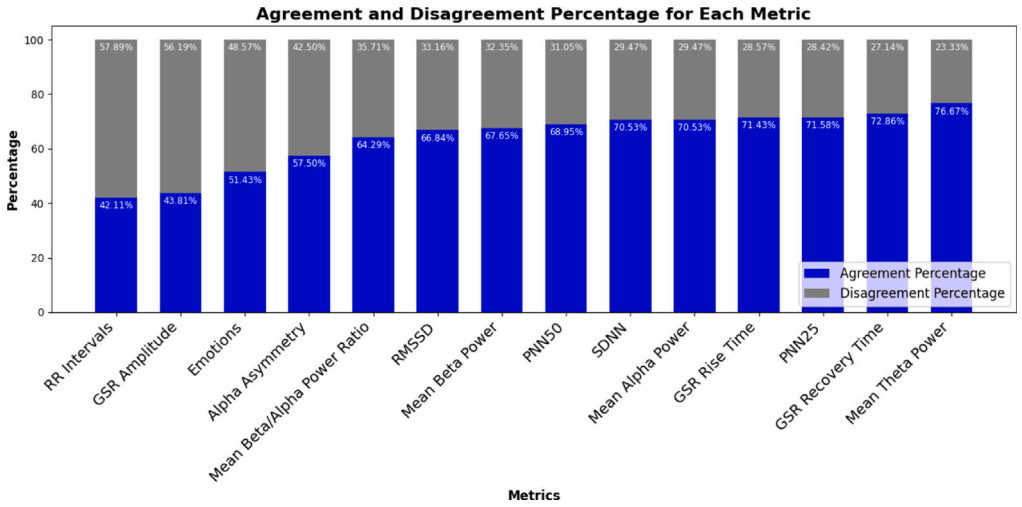


Fig. 9. Percentage of agreement between psycho-physiological metrics and subjective questionnaire NASA-TLX.

Table 10
Statistical tests for the analysis of emotional response. The columns on the right report the number of test participants for which a statistically significant difference was observed among the different conditions for $\alpha = 0.05$ and $\alpha = 0.01$. Rows in blue denote consecutive tests. Total sample size = 21.

Test	Conditions	$\alpha = 0.05$	$\alpha = 0.01$
MANOVA	All tasks	21	21
Chi square	Task 1 - Baseline	16	11
	Task 2 - Baseline	12	11
	Task 3 - Baseline	11	9
	Task 4 - Baseline	12	9
	Task 5 - Baseline	14	13
Chi square with consistency	All tasks	18	16
Bonferroni	Task 1 - Task 2	17	17
	Task 1 - Task 3	20	20
	Task 1 - Task 4	21	21
	Task 1 - Task 5	21	21
	Task 2 - Task 3	7	7
	Task 2 - Task 4	11	11
	Task 2 - Task 5	14	11
	Task 3 - Task 4	3	3
	Task 3 - Task 5	5	5
	Task 4 - Task 5	1	1

Table 11
Agreement between emotional response and subjective responses from NASA-TLX, when comparing tasks. Values are in percentage over 21 subjects.

Conditions	Agreement %
Task 1 - Task 2	38.10
Task 1 - Task 3	33.33
Task 1 - Task 4	38.10
Task 1 - Task 5	42.86
Task 2 - Task 3	57.14
Task 2 - Task 4	47.62
Task 2 - Task 5	57.14
Task 3 - Task 4	66.67
Task 3 - Task 5	52.38
Task 4 - Task 5	80.95
Average	**51.43**

7.4. Emotions discussion

The capture of emotions in the dynamic setting of this project faced challenges due to occlusions and registration loss. Precision is likely higher in static scenarios, where the subject remains stationary in front of the camera, allowing for continuous recording. The level of concordance between recorded emotions and subjective assessments from NASA-TLX stands at 51.43% confirming research question **Q3**.

8. Conclusion

The purpose of this study was to investigate the use of psycho-physiological data to assess the stress conditions of operators involved in cobot programming tasks. For this purpose, the cobot programming platform used in this work mirrors an authentic scenario, where incremental learning is paired with the step-by-step execution of tasks based on their increasing difficulty, together with the collection of different human data through wearable devices. All participants followed the same sequence of tasks and utilized a platform designed to mimic a real learning environment suitable for industrial use. Despite this standard approach, the physiological parameters between different participants present high variability and are very subjective (research question **Q1**). The test demonstrated that the participant's psycho-physiological response is influenced not only by the task's difficulty (research question **Q2**), but also by the temporal succession of the same: physiological responses observed in temporally successive tasks generally are similar to each other. In addition to the tasks that are more complex and require greater skills, in this experiment, a certain distrust seems to be observed in approaching the cobot, highlighted by an increase in the stress parameters in the early stages.

This study showed that the metrics derived from GSR, EEG and HRV signals are among the most precise in describing the subjectively perceived stress state. This means that, in the specific context of this experiment and with the analysis methods used, the brain parameters, linked to sweating and those of HRV derived from raw RR intervals distribution show the greatest coherence with the subjective evaluations, responding positively to the research question **Q3**, as opposed to facial expressions and raw cardiac parameters. It also suggests that the level of accuracy for stress prediction could further increase by using the combination of some of these metrics.

The achieved results could be used in the future to design a suitable set-up to be adopted in industrial contexts, where the use of a large number and bulky sensors is not possible. Indeed, this work supports the selection of the most suitable signals and metrics and defines a potential set-up as well as a protocol analysis to detect human stress conditions in HRC contexts.

Future research should strive to elucidate stress-related psycho-physiological responses in other real contexts beyond those of human cobot programming to observe whether the responses are context-specific or independent. Furthermore, further analysis of the data obtained will be carried out to observe which metrics obtained from signals may be the best in the classification and prediction of stress statistics.

CRediT authorship contribution statement

Simone Borghi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrea Ruo:** Writing – review & editing, Methodology, Conceptualization. **Lorenzo Sabattini:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Margherita Peruzzini:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Valeria Villani:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgments

This research was supported by the project "FAR 2021 Attrezzature Misura 2" funded by the University of Modena and Reggio Emilia, Italy, PI Prof. Margherita Peruzzini and the PRIN PNRR 2022 "ART.I.DE" project funded by the Italian Ministry of Education and Merit.

References

Aaltonen, I., Salmi, T., 2019. Experiences and expectations of collaborative robots in industry and academia: Barriers and development needs. Procedia Manuf. 38, 1151–1158.

Acerbi, G., Rovini, E., Betti, S., Tirri, A., Rónai, J.F., Sirianni, A., Agrimi, J., Eusebi, L., Cavallo, F., 2017. A wearable system for stress detection through physiological data analysis. In: Ambient Assisted Living: Italian Forum 2016 7. Springer, pp. 31–50.

Acharya, U.R., Sree, S.V., Ang, P.C.A., Yanti, R., Suri, J.S., 2012. Application of non-linear and wavelet based features for the automated identification of epileptic eeg signals. Int. J. Neural Syst. 22 (02), 1250002.

Alotaiby, T., El-Samie, F.E.A., Alshebeili, S.A., Ahmad, I., 2015. A review of channel selection algorithms for eeg signal processing. EURASIP J. Adv. Signal Process. 2015, 1–21.

Andersson, D., 2017. Real-time ecg for objective stress level measurement.

Aqajari, S.A. Hossein, Naeini, E.K., Mehrabadi, M.A., Labbaf, S., Dutt, N., Rahmani, A.M., 2021. Pyeda: An open-source python toolkit for pre-processing and feature extraction of electrodermal activity. Procedia Comput. Sci. 184, 99–106, the 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops. DOI: DOI: 10.1016/j.procs.2021.03.021. URL https://www.sciencedirect.com/science/article/pii/S1877050921006438.

Aslam, M., Albassam, M., 2020. Presenting post hoc multiple comparison tests under neutrosophic statistics. J. King Saud Univ., Eng. Sci. 32 (6), 2728–2732.

Aslam, M., Smarandache, F., 2023. Chi-square test for imprecise data in consistency table. Infin. Study.

Association, R.I., et al., 1999. Industrial robots and robot systems-safety requirements, R15. 06-1999 62.

Berretz, G., Packheiser, J., Wolf, O.T., Ocklenburg, S., 2022. Acute stress increases left hemispheric activity measured via changes in frontal alpha asymmetries. Iscience 25 (2).

Bishay, M., Preston, K., Strafuss, M., Page, G., Turcot, J., Mavadati, M., 2022. Affdex 2.0: A real-time facial expression analysis toolkit. arXiv preprint arXiv:2202.12059.

Borghi, S., Zucchi, F., Prati, E., Ruo, A., Villani, V., Sabattini, L., Peruzzini, M., 2024. Unlocking human–robot dynamics: Introducing sensecobot, a novel multimodal dataset on industry 4.0. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, HRI '24. Association for Computing Machinery, New York, NY, USA, pp. 880–884. http://dx.doi.org/10.1145/3610977.3636440.

Boucsein, W., 2012. Electrodermal Activity. Springer Science & Business Media.

Bryant, R.A., Friedman, M.J., Spiegel, D., Ursano, R., Strain, J., 2011. A review of acute stress disorder in dsm-5. Focus 9 (3), 335–350.

Bussolan, A., Baraldo, S., L.M. Gambardella, A., 2023. Valente, assessing the impact of human–robot collaboration on stress levels and cognitive load in industrial assembly tasks. In: ISR Europe 2023; 56th International Symposium on Robotics. VDE, pp. 78–85.

Buzsaki, G., 2006. Rhythms of the Brain. Oxford University Press.

Cannon, W., 1932. The Wisdom of the Body. ww norton & company, Inc., New York.

Carissoli, C., Negri, L., Bassi, M., Storm, F.A., Delle Fave, A., 2023. Mental workload and human–robot interaction in collaborative tasks: A scoping review. Int. J. Hum.-Comput. Interact. 1–20.

Chandra, S., Jaiswal, A.K., Singh, R., Jha, D., Mittal, A.P., 2017. Mental stress: Neurophysiology and its regulation by sudarshan kriya yoga. Int. J. Yoga 10 (2), 67.

Cohen, S., Kamarck, T., Mermelstein, R., et al., 1994. Perceived stress scale. Measuring Stress: A Guide Health Soc. Sci. 10 (2), 1–2.

Corder, G.W., Foreman, D.I., 2011. Nonparametric statistics for non-statisticians.

De Luna, A.B., 2008. Basic Electrocardiography: Normal and Abnormal ECG Patterns. John Wiley & Sons.

Dinno, A., 2015. Nonparametric pairwise multiple comparisons in independent groups using dunn's test. Stata J. 15 (1), 292–300.

Favre-Félix, J., Dziadzko, M., Bauer, C., Duclos, A., Lehot, J.-J., Rimmelé, T., Lilot, M., 2022. High-fidelity simulation to assess task load index and performance: a prospective observational study. Turkish J. Anaesthesiol. Reanim. 50 (4), 282.

Fay, M.P., Proschan, M.A., 2010. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. Stat. Surv. 4, 1.

Föhr, T., Tolvanen, A., Myllymäki, T., Järvelä-Reijonen, E., Rantala, S., Korpela, R., Peuhkuri, K., Kolehmainen, M., Puttonen, S., Lappalainen, R., et al., 2015. Subjective stress, objective heart rate variability-based stress, and recovery on workdays among overweight and psychologically distressed individuals: a cross-sectional study. J. Occup. Med. Toxicol. 10, 1–9.

Franklin, C.S., Dominguez, E.G., Fryman, J.D., Lewandowski, M.L., 2020. Collaborative robotics: New era of human–robot cooperation in the workplace. J. Saf. Res. 74, 153–160.

Fuentes-García, J.P., Villafaina, S., Collado-Mateo, D., Cano-Plasencia, R., Gusi, N., 2020. Chess players increase the theta power spectrum when the difficulty of the opponent increases: an eeg study. Int. J. Environ. Res. Public Health 17 (1), 46.

Gatzke-Kopp, L.M., Jetha, M.K., Segalowitz, S.J., 2014. The role of resting frontal eeg asymmetry in psychopathology: Afferent or efferent filter? Dev. Psychobiol. 56 (1), 73–85.

George, P., Cheng, C.-T., Pang, T.Y., Neville, K., 2023. Task complexity and the skills dilemma in the programming and control of collaborative robots for manufacturing. Appl. Sci. 13 (7), 4635.

Ghanavati, F.K., Choobineh, A., Keshavarzi, S., Nasihatkon, A.A., Roodbandi, A.S.J., 2019. Assessment of mental workload and its association with work ability in control room operators. La Medicina del lavoro 110 (5), 389.

Ghasemi, A., Zahediasl, S., 2012. Normality tests for statistical analysis: a guide for non-statisticians. Int. J. Endocrinol. Metab. 10 (2), 486.

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., Tsiknakis, M., 2019. Review on psychological stress detection using biosignals. IEEE Trans. Affect. Comput. PP 1.

Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P.G., Marias, K., Tsiknakis, M., 2017. Stress and anxiety detection using facial cues from videos. Biomed. Signal Process. Control 31, 89–101.

Hansen, A., Villani, V., Pupa, A., Lassen, A.H., 2023. Introducing Novice Operators To Collaborative Robots: A Hands-on Approach for Learning and Training. Authorea Preprints.

Hart, S.G., (tlx), Nasa.task.load.index., 1986. Human Performance Research Group NASA Ames Research Center Moffett Field. California.

Hayashi, T., Okamoto, E., Nishimura, H., Mizuno-Matsumoto, Y., Ishii, R., Ukai, S., 2009. Beta activities in eeg associated with emotional stress. Int. J. Intell. Comput. Med. Sci. Image Process. 3 (1), 57–68. http://dx.doi.org/10.1080/1931308X.2009.10644171.

Herman, J.P., Figueiredo, H., Mueller, N.K., Ulrich-Lai, Y., Ostrander, M.M., Choi, D.C., Cullinan, W.E., 2003. Central mechanisms of stress integration: hierarchical circuitry controlling hypothalamo–pituitary–adrenocortical responsiveness. Front. Neuroendocrinol. 24 (3), 151–180.

Hoehn-Saric, R., McLeod, D.R., Zimmerli, W.D., 1989. Somatic manifestations in women with generalized anxiety disorder: Psychophysiological responses to psychological stress. Arch. Gen. Psychiatry 46 (12), 1113–1119.

Holm, A., Lukander, K., Korpela, J., Sallinen, M., Müller, K.M., 2009. Estimating brain load from the eeg. ScientificWorld J. 9, 639–651.

Holmes, T.H., Rahe, R.H., 1967. The social readjustment rating scale. J. Psychosom. Res..

Hong, S., Zhang, W., Sun, C., Zhou, Y., Li, H., 2020. Practical lessons on 12-lead ecg classification: meta-analysis of methods from physionet/computing in cardiology challenge. Front. Physiol. 12, 2505.

Huberty, S., Carter Leno, V., van Noordt, S.J., Bedford, R., Pickles, A., Desjardins, J.A., Webb, S.J., Team, B., Elsabbagh, M., 2021. Association between spectral electroencephalography power and autism risk and diagnosis in early development. Autism Res. 14 (7), 1390–1403.

Hyvärinen, A., 2013. Independent component analysis: recent advances. Phil. Trans. R. Soc. A 371 (1984), 20110534.

Irastorza, X., Milczarek, M., Cockburn, W., 2016. Second European Survey of Enterprises on New and Emerging Risks (ESENER-2): Overview Report: Managing Safety and Health At Work. Publications Office of the European Union.

Javaid, M., Haleem, A., Singh, R.P., Rab, S., Suman, R., 2022. Significant applications of cobots in the field of manufacturing. Cogn. Robotics 2, 222–233.

J.R. Crawford, J.D. Henry, 2004. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. Br. J. Clin. Psychol. 43 (3), 245–265.

Karpiel, I., Kurasz, Z., Kurasz, R., Duch, K., 2021. The influence of filters on eeg-erp testing: Analysis of motor cortex in healthy subjects. Sensors 21 (22), 7711.

Kassinopoulos, M., Harper, R.M., Guye, M., Lemieux, L., Diehl, B., 2021. Altered relationship between heart rate variability and fmri-based functional connectivity in people with epilepsy. Front. Neurol. 12, 671890.

Kleiger, R.E., Stein, P.K., Bigger Jr., J.T., 2005. Heart rate variability: measurement and clinical utility. Ann. Noninvasive Electrocardiol. 10 (1), 88–101.

Koolhaas, J.M., Bartolomucci, A., Buwalda, B., de Boer, S.F., Flügge, G., Korte, S.M., Meerlo, P., Murison, R., Olivier, B., Palanza, P., et al., 2011. Stress revisited: a critical evaluation of the stress concept. Neurosci. Biobehav. Rev. 35 (5), 1291–1301.

Levine, S., 2005. Stress: An historical perspective. In: Techniques in the Behavioral and Neural Sciences, 15. Elsevier, pp. 3–23.

Macias, E., Parent-Thirion, A., j. hurley, Vermeylen, G., 2007. Fourth European Working Conditions Survey. Eurofound.

Mariotti, A., 2015. The effects of chronic stress on health: new insights into the molecular mechanisms of brain–body communication. Future Sci. OA 1 (3).

Mariscal, M.A., Ortiz Barcina, S., García Herrero, S., López Perea, E.M., 2023. Working with collaborative robots and its influence on levels of working stress. Int. J. Comput. Integr. Manuf. 1–20.

Mayapur, P., 2018. A review on detection and performance analysis on rr interval methods for ecg. Int. J. Innov. Res. Sci. Eng. Technol 7, 11019–11025.

Melo, H.M., Martins, T.C., Nascimento, L.M., Hoeller, A.A., Walz, R., Takase, E., 2018. Ultra-short heart rate variability recording reliability: The effect of controlled paced breathing. Annals Noninvasive Electrocardiol. 23 (5), e12565.

Memar, M., Mokaribolhassan, A., 2021. Stress level classification using statistical analysis of skin conductance signal while driving. SN Appl. Sci. 3 (1), 64.

Moser, B., et al., 2022. A systematic literature review of user experience evaluation scales for human–robot collaboration. In: Transdisciplinarity and the Future of Engineering: Proceedings of the 29th International Society of Transdisciplinary Engineering (ISTE) Global Conference. IOS Press, Cambridge, MA, USA, 28, p. 13, July 5–July 8 2022.

Pollak, A., Paliga, M., Pulopulos, M.M., Kozusznik, B., Kozusznik, M.W., 2020. Stress in manual and autonomous modes of collaboration with a cobot. Comput. Hum. Behav. 112, 106469.

Raufi, B., Longo, L., 2022. An evaluation of the eeg alpha-to-theta and theta-to-alpha band ratios as indexes of mental workload. Front. Neuroinform. 16, 44.

Ruo, A., Villani, V., Sabattini, L., 2022. Use of eeg signals for mental workload assessment in human–robot collaboration. In: International Workshop on Human-Friendly Robotics. Springer, pp. 233–247.

Said, S., Gozdzik, M., Roche, T.R., Braun, J., Rössler, J., Kaserer, A., Spahn, D.R., Nöthiger, C.B., Tscholl, D.W., 2020. Validation of the raw national aeronautics and space administration task load index (nasa-tlx) questionnaire to assess perceived workload in patient monitoring tasks: pooled analysis study using mixed models. J. Med. Internet Res. 22 (9), e19472.

Selye, H., 1946. The general adaptation syndrome and the diseases of adaptation. J. Clin. Endocrinol. 6 (2), 117–230.

Shaffer, F., Ginsberg, J.P., 2017. An overview of heart rate variability metrics and norms. Front. Public Health 5, 258.

T. F. o. t. E. S. o. C. t. N. A. S. o. P. Electrophysiology, 1996. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Circulation 93 (5), 1043–1065. http://dx.doi.org/10.1161/01.CIR.93.5.1043.

Taesi, C., Aggogeri, F., Pellegrini, N., 2023. Cobot applications—recent advances and challenges. Robotics 12 (3), 79.

Teo, G., Matthews, G., Reinerman-Jones, L., Barber, D., 2020. Adaptive aiding with an individualized workload model based on psychophysiological measures. Human-Intell. Syst. Integr. 2, 1–15.

Tompkins, W.J., 1993. Biomedical Digital Signal Processing. Editorial Prentice Hall.

Tsang, P.S., Velazquez, V.L., 1996. Diagnosticity and multidimensional subjective workload ratings. Ergonomics 39 (3), 358–381.

Ursin, H., Eriksen, H.R., 2010. Cognitive activation theory of stress (cats). Neurosci. Biobehav. Rev. 34 (6), 877–881.

Valente, A., Avram, O., 2023. From cobots to human–robot synergy–overview and future trends. In: Frontiers in Robotics and AI. vol. 10.

Vidulich, M.A., Tsang, P.S., 2012. Mental workload and situation awareness. In: Handbook of Human Factors and Ergonomics. pp. 243–273.

Villani, V., Righi, M., Sabattini, L., Secchi, C., 2020. Wearable devices for the assessment of cognitive effort for human–robot interaction. IEEE Sens. J. 20 (21), 13047–13056.

Warne, R., 2019. A primer on multivariate analysis of variance (manova) for behavioral scientists. Pract. Assess. Res. Eval. 19 (1).

Xavier, G., Su Ting, A., Fauzan, N., 2020. Exploratory study of brain waves and corresponding brain regions of fatigue on-call doctors using quantitative electroencephalogram. J. Occup. Health 62 (1), e12121.

Zakeri, Z., Arif, A., Omurtag, A., Breedon, P., Khalid, A., 2023. Multimodal assessment of cognitive workload using neural, subjective and behavioural measures in smart factory settings. Sensors 23 (21), 8926.

Zou, Z., Ergan, S., 2021. Evaluating the effectiveness of biometric sensors and their signal features for classifying human experience in virtual environments. Adv. Eng. Inform. 49, 101358.

Zulkurnaini, N.A., Kadir, R.S.S.A., Murat, Z.H., Isa, R.M., 2012. The comparison between listening to al-quran and listening to classical music on the brainwave signal for the alpha band. In: 2012 Third International Conference on Intelligent Systems Modelling and Simulation. IEEE, pp. 181–186.