Wearable Physiological Signals under Acute Stress and Exercise Conditions

Andrea Hongn[1,2*], Facundo Bosch[3], Lara Eleonora Prado[3], José Manuel Ferrández[4,5], María Paula Bonomini[1,2,4]

[1]Universidad de Buenos Aires, Facultad de Ingeniería, Instituto de Ingeniería Biomédica (IIBM), Buenos Aires, Argentina.
[2]CONICET, Instituto Argentino de Matemática "Alberto P. Calderón" (IAM), Buenos Aires, Argentina.
[3]Instituto Tecnológico de Buenos Aires (ITBA), Buenos Aires, Argentina.
[4]Departamento de Electrónica, Tecnología de Computadoras y Proyectos, Universidad Politécnica de Cartagena, Cartagena, Spain.
[5]European Culture and Technology Laboratory - ECT Lab+, European University of Technology.

*Corresponding author(s). E-mail(s): ahongn.ext@fi.uba.ar;

Abstract

In this work, a novel dataset containing physiological signals recorded non invasevely during structured acute stress induction, as well as aerobic and anaerobic exercise sessions is presented. The physiological data were collected using the Empatica E4, a wearable device that measures electrodermal activity, skin temperature, three-axis accelerometry and blood volume pulse, from which heart rate and heart rate variability features can be derived. A stress induction protocol was designed using mathematical and emotional tasks to elicit physiological responses. For aerobic and anaerobic exercise, a stationary bike routine was developed to distinguish between the two types of activity. The dataset includes records from 36 healthy individuals during the stress protocol, 30 during aerobic exercise, and 31 during anaerobic exercise. Several machine learning algorithms were applied to validate the dataset, with XGBoost achieving an accuracy of 93% in classifying stress versus rest, 91% in distinguishing between aerobic and anaerobic exercise, and 84% in a four-label classification task involving stress, rest, aerobic, and anaerobic activities. The dataset is publicly available for further research.

Keywords: stress, wearable monitoring, aerobic and anaerobic exercise, dataset

1 Background & Summary

1.1 Motivation

Type 1 diabetes mellitus (T1DM) is an autoimmune disease that leads to the destruction of insulin-producing pancreatic beta cells. As a result, patients require lifelong exogenous insulin administration and continuous blood glucose monitoring to prevent hypo- and hyperglycemia episodes [1, 2]. The most advanced technology for diabetes management today is the artificial pancreas (AP) system, which automates the delivery of basal insulin via continuous subcutaneous infusion. This system uses a control algorithm that receives data from interstitial glucose sensors to regulate an insulin pump [1]. However, varied factors, such as food intake, physical activity (PA), stress, and illness, can cause significant fluctuations in blood glucose levels. These fluctuations often require user intervention due to rapid and sometimes unpredictable changes in insulin requirements [2]. The current challenge is to improve the performance and autonomy of AP systems by incorporating biometric measurements of these disturbances. The first step toward this goal is the accurate detection of physical activity and acute stress using non-invasive physiological signals.

Stress is a natural human response, defined by the World Health Organization (WHO) as a state of worry or mental tension triggered by a difficult situation. Stressful events can elicit cognitive, emotional, and biological reactions, which can be assessed through self-report measures, behavioral coding, or physiological measurements [3]. Indicators such as heart rate (HR), heart rate variability (HRV), blood volume pulse (BVP), electrodermal activity (EDA), skin temperature (ST), and motion activity are widely recognized as markers of acute stress. These physiological responses are of particular interest because they can be measured in real-time using wearable sensors like photoplethysmography (PPG), accelerometers (ACC), and temperature sensors, enabling continuous stress monitoring in daily life—a critical requirement for integration into the AP system [4].

Physical activity is defined by WHO as any bodily movement produced by skeletal muscles that requires energy expenditure. PA includes both structured exercise—activities characterized by specific types, intensities, and durations—and non-exercise activities related to daily living. Aerobic, sprint, and resistance training can lead to significant variations in blood glucose levels, making it crucial to detect not only the presence of physical activity but also its type, intensity, and duration. Although wearable devices may not achieve the accuracy of laboratory techniques in quantifying exercise, they offer the advantage of continuous monitoring in real-world settings. Among the biomarkers related to PA intensity, HR provides the best estimation of energy expenditure. Other variables measurable by wearable devices that can serve as indicators of exercise include ST, near-body ambient temperature, skin electrical conductance, breathing rates/frequencies, heat flux, sweat rate, and accelerations in triaxial planes [2].

Although the physiological responses to exercise and acute stress can be described using the same indicators, a key objective of this work is to analyze the best features for detecting each condition. In pursuing this goal, we observed that, despite several

research groups working on this topic, there is a notable lack of open datasets containing wearable data collected during structured sessions of acute stress and aerobic and anaerobic exercise.

1.2 Public datasets

We conducted an extensive search for open datasets on platforms such as PhysioNet and Kaggle, as well as popular scientific repositories, focusing on wearable data collected under acute stress and structured physical activity conditions. However, none of the datasets we found met the specific needs of our research. As a result, we decided to carry out our data collection, to make this dataset publicly available in the future.

Several publicly available datasets focus on stress detection and emotional research using wearable data [5–7]. We excluded datasets from non-wearable devices, as they do not align with our research objectives. However, some datasets also include data from non-wearable sensors such as ECG, EEG, respiration, and SpO2 [8–10]. Additionally, certain datasets feature multimodal recordings, including voice and video [10].

The measurement devices, stressors, and self-reported emotional/stress assessments vary across these datasets. Common stress-inducing tasks include the Stroop test, arithmetic tasks, the Trier Social Stress Test, and stressful videos. Most of these datasets were collected in controlled environments, though some were recorded in specific contexts, such as during exams [11] and while driving [12].

A few datasets include continuous monitoring in daily life, capturing spontaneous stress events and unstructured physical activities like chores, walking, and climbing stairs. For example, the Multilevel Monitoring of Activity and Sleep in Healthy people (MMASH) dataset provides psychophysiological data of individuals in daily life, but the activities related to physical activity are grouped into 13 categories, introducing inter-participant variability (n=22) [13] .

In terms of physical activity monitoring, Poon et al. developed a PPG dataset from wearable devices as participants engaged in sitting, stationary walking, and running [14]. The dataset presented by Urbanek et al. focuses on ACC data collected during semi-structured outdoor activities, such as walking, stair climbing, and driving [15]. Similarly, the REALDISP dataset includes a wide range of physical activities (e.g., warm-up, cool-down, and fitness exercises), sensor modalities (including acceleration, rate of turn, magnetic field, and quaternions), and participants (n=17) [16].

There are also several open datasets focused on Human Activity Recognition, which involve participants performing various activities, including exercise [17].

The datasets developed by Birjandtalab et al. and Falk et al., which include acute stress and exercise, are the most closely related to our work[18, 19]. The first one explores neurological status assessments and includes a 5-minute exercise routine as a physical stressor measured with a wearable device. In the second one, participants performed tasks of varying stress levels at three different activity levels on a stationary bike (0 km/h,18 km/h,24 km/h) and provided quantitative ratings of their perceived stress and fatigue levels. However, to the best of our knowledge, no publicly available dataset captures wearable data during structured, well-defined, induced stress sessions alongside long-duration aerobic and anaerobic exercise protocols.

1.3 Key contribution

We are sharing a dataset of healthy young men and women (n=36), which includes physiological signals from a wearable device. These signals were collected during structured sessions of induced acute stress, as well as during exercise sessions. In addition, self-reported stress levels are provided. All activities followed standardized and reproducible protocols.

The main contributions of this work are as follows:

- To the best of our knowledge, this dataset fills the gap between stress and structured physical activity assessment, unifying both in a single database. This enables the analysis of variables influenced by stress, as well as aerobic and anaerobic exercise, improving the accurate detection of these states. This is essential, for example, to incorporate biometric measurements into an AP control algorithm due to different effects of these disturbances on glycemic dynamic in T1DM patients.
- For stress research, we provide an efficient and reproducible protocol for stress inducement, supported by self-reported stress levels throughout the protocol. This protocol alternates periods of acute induced stress with rest periods, providing a comprehensive view of physiological responses.
- Exercise sessions were conducted on different days to ensure that the effects do not overlap. Aerobic and anaerobic exercise sessions are of considerable length (20-30 minutes), reflecting typical exercise routines in daily life. focusing on differentiating between aerobic and anaerobic activities.
- Protocols are well-documented, and signals are labeled to accurately identify and segment each phase. Additionally, we provide a script to open, read, and visualize the data.
- This work also presents machine learning classification results to validate the potential of the presented data.

2 Methods

2.1 Measurement Device

The Empatica E4 wristband, a class IIa Medical Device in the EU, is a wearable wireless device designed for comfortable, continuous, real-time data acquisition in daily life. This wristband is intended for use in research settings. The E4 contains four sensors, each with its own sampling frequency: (1) Photoplethysmography to provide blood volume pulse, from which HR, HRV, and other cardiovascular features may be derived; (2) Electrodermal activity, used to measure sympathetic nervous system arousal; (3) 3-axis accelerometer, to capture motion-based activity; (4) Infrared thermopile, reading skin temperature.

2.2 Data collection

The data collection process was conducted in two stages. Initially, a group of 18 volunteers (v1) followed the protocol. A few months later, a second group of 18 volunteers

(v2) participated using an improved protocol based on initial experience. We will outline these modifications below. For all protocols, the E4 device was placed on the subject's non-dominant hand to minimize motion artifacts during the tests.

2.2.1 Population

Participants were males and females between 19 and 30 years. Demographic information is presented on Table 1.

Table 1: Demographic participants data

ID	Gender	Age	Height (cm)	Weight (kg)	Trains	Protocol	Stress Inducement	Aerobic Exercise	Anaerobic Exercise
S01	m	21	192	84	Y	v1	Y	Y	Y****
S02	m	20	185	95	N	v1	Y****	Y	Y
S03	m	20	175	62	Y	v1	Y	Y***	Y
S04	m	21	174	70	Y	v1	Y	Y	Y
S05	m	21	173	72	Y	v1	Y	Y	Y
S06	m	21	172	70	Y	v1	Y	Y	Y***
S07	m	19	184	88	Y	v1	Y	Y***	Y
S08	m	20	174	67	Y	v1	Y	Y	Y
S09	m	19	174	63	Y	v1	Y	Y	Y
S10	m	21	180	80	Y	v1	Y	Y	Y
S11	m	21	183	64	Y	v1	Y	Y**	Y
S12	m	18	176	79	Y	v1	Y	-	Y
S13	m	21	175	65	Y	v1	Y	Y	Y
S14	m	19	182	85	Y	v1	Y	Y	Y
S15	m	21	176	77	Y	v1	Y	Y	Y
S16	m	20	168	61	Y	v1	Y	Y	Y**
S17	m	22	173	78	Y	v1	Y	Y	Y
S18	m	21	183	80	Y	v1	Y	Y	Y
f01	f	25	152	61	N	v2	Y	Y	Y
f02	f	29	164	80	N	v2	Y	Y	Y
f03	f	26	160	61	Y	v2	Y	Y	Y
f04	f	29	168	56	N	v2	Y	Y	Y
f05	f	21	165	55	Y	v2	Y	Y	Y
f06	f	21	169	58	N	v2	Y	Y	Y
f07	f	21	163	47	N	v2	Y*	Y	Y
f08	f	22	158	50	N	v2	Y	Y	Y
f09	f	21	170	56	Y	v2	Y	Y	Y
f10	f	21	172	65	Y	v2	Y	Y	Y
f11	f	31	170	84	N	v2	Y	Y	Y
f12	f	30	158	97	N	v2	Y	Y	Y
f13	f	29	154	56	N	v2	Y	Y	Y
f14	f	-	-	-	-	v2	Y**	-	-
f15	f	-	-	-	-	v2	Y	-	-
f16	f	-	-	-	-	v2	Y	-	-
f17	f	-	-	-	-	v2	Y	-	-
f18	m	-	-	-	-	v2	Y	-	-

References: *m(male); f(female); Y(Yes); N(No); v1(First Stage); v2(Second Stage); * Wrong wristband placement; ** Connection Loss; *** Protocol was not fully completed; **** Files with Constraints*

Data was collected under the supervision of the Ciencias de la Vida Department at the Instituto Tecnológico de Buenos Aires (ITBA) following institutional ethical standards. Enrollment in the study was facilitated through an online form. The exclusion criteria included individuals with chronic illnesses and those with a family history of sudden death during exercise, as well as participants undergoing psychiatric treatment or taking medications that could potentially affect physiological responses. Before conducting the tests, each participant signed an informed consent form. This document was sent to participants in advance, allowing them sufficient time to review the information and ask any questions they might have. Participation in the trial was completely voluntary, and participants were informed that they could withdraw at any time without providing justification.

2.2.2 Protocols

Stress Inducement

The participant was welcomed by the researcher and led to a quiet room where the protocol was conducted. The room contained two computers: one displaying a guide outlining all the steps, and another where the interactive tasks were performed. The participants wore earphones to cancel out environmental noise and to enhance the sound stimuli during a specific task. The researcher guided the procedure but left the participant alone during rest periods to promote relaxation. The original protocol(v1) started with a 3-minute baseline recording, to be used as a reference. The first stress test was an adaptation of the widely used Stroop Test [20, 21], adapted from PsyToolkit [22, 23]. Afterwards, a 5-minute-rest period was imposed, followed by a modified version of the Trier Mental Challenge Test [24], obtained through Millisecond Software, LLC. This test involved a series of mathematical tasks within a 5 seconds time limit while an annoying sound stimulus was played in the background. Participants were also instructed to vocalize their responses aloud, which further added to the cognitive load and performance demands of the task. Once again, a 5-minute-period came before the final block. In the latter, participants were asked to express their opinion about controversial topics and suddenly were instructed to defend the opposite opinion over the same subject. Finally, participants were tasked with counting backward from 1022 in decrements of 13, providing the answers aloud. Each of these tests had a time limit of 30 seconds.

Before and after every task and rest period, participants were required to verbally express their self-perception stress level (SL) on a scale ranging from 1 to 10. A summary of the protocol is shown in Figure 1.

Anaerobic Activity

The anaerobic exercise protocol was adapted from the Wingate Anaerobic Test [25]. It began with a 3-minute baseline recording during which the subject pedaled without resistance to warm up. This was followed by three cycles, each consisting of 30 seconds of maximal effort, where the subject pedaled at their highest intensity against high resistance, followed by a 4-minute cool-down period without resistance. Finally, a 2-minute recording was made while the subject remained still.

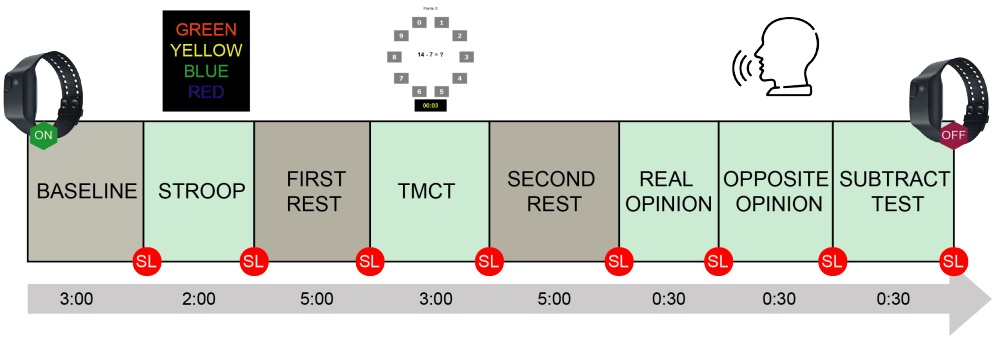

Fig. 1: Stress Induction Original Protocol (*v1*). *SL (Stress Level)*

Aerobic Activity

The aerobic exercise test was adapted from the Storer-Davis Maximal Bicycle Protocol and involved continuous stationary cycling for approximately 35 minutes [26]. First, we determined the maximum resistance for each participant by identifying the point at which they could no longer pedal at maximum effort. The protocol began with a 3-minute baseline recording during which the subject pedaled without resistance to warm up. After completing the baseline, the subject cycled in sync with a metronome, with each beat corresponding to one foot down (or one knee up), meaning one revolution was completed every two beats.

Starting with low resistance (20% of maximum), the subject went through three 3-minute periods at increasing paces of 60, 70, and 75 revolutions per minute (rpm), gradually raising the resistance up to a medium level (30% of maximum). This was followed by four periods, the first two lasting 3 minutes each and the second two lasting 2 minutes each, at paces of 80, 85, 90, and 95 rpm, respectively, with a gradual increase in resistance (5% per stage). With a final fixed medium-high resistance (50% of maximum), the last three periods consisted of 2 minutes each, at paces of 100, 105, and 110 rpm, respectively.

Once completed, a 4-minute cool down period without resistance was conducted, followed by 2 minutes of remaining still.

2.2.3 Protocol improvements

The second version of the protocol incorporated several modifications based on previous experience. For stress induction,the Stroop Test was removed, and the second rest period was relocated between the opinion tasks and the subtraction test. Rest periods were extended, and a relaxing video was shown. Additionally, the protocol was conducted remotely, creating a more relaxed environment during rest times.

In the updated exercise protocols, participants attended in groups to a spinning room. The aerobic protocol was modified as follows: a baseline was introduced, followed by a 2:15-minute warm-up. This was succeeded by three 1:30-minute intervals at 70, 75, and 80 rpm, respectively. An 11:15-minute session at 85 rpm was conducted, leading into a final 4:30-minute period at 90/95 rpm (depending on the participant's

condition). The session concluded with a 3-minute cool down, followed by a rest period where participants sat on the bike without movement.

For the anaerobic protocol, a fourth maximum power sprint was added, with the sprints extended to 45 seconds each, followed by a corresponding cool down period.

3 Data Records

All raw data recorded during the experiment are publicly available for further research and analysis. The Jupyter Notebook containing the source code for data visualization, along with complementary files such as stress levels and demographic data, is provided alongside the raw signal files. Data can be downloaded by visiting the URI: https://doi.org/10.5281/zenodo.13993658.

The dataset is organized into three main categories: STRESS, AEROBIC exercise, and ANAEROBIC exercise. Each category contains subfolders specific to individual subjects, where raw sensor reading *.csv* files downloaded from the Empatica E4 Connect are stored. The exceptions are the *IBI.csv* and *HR.csv* files, which are generated by Empatica's algorithm. Tags related to unintentionally pressed buttons have been deleted from the tags file to improve protocol understanding. These tags mark the beginning and end of protocol segments, which facilitates signal segmentation.

In accordance with Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor De-Identification guidelines, each participant is assigned a unique ID. The session dates and event marks during the protocols (in *tags.csv*) have been modified by a random number of days. Time samples have been shifted consistently across all records to maintain signal alignment. Participants from the first stage are labeled as "Sxx," while those from the second stage are labeled as "fxx".

Each subject folder contains the raw signal *.csv* files provided by Empatica. These files follow this format: the first row is the initial time of the session expressed in UTC (Empatica provides time in Unix timestamp format, but files are already converted to UTC). The second row is the sample rate expressed in Hz.

- TEMP.csv: Data from temperature sensor expressed degrees on the Celsius ($^{\circ}C$) scale.
- EDA.csv: Data from the electrodermal activity sensor expressed as microsiemens (μS) .
- BVP.csv: Data from photoplethysmography sensor.
- ACC.csv: Data from 3-axis accelerometer sensor. The accelerometer is configured to measure acceleration in the range [-2g, 2g]. Therefore the unit in this file is 1/64g. Data from x, y, and z axis are respectively in first, second, and third column.
- IBI.csv: Time between individuals heart beats extracted from the BVP signal. No sample rate is needed for this file. The first column is the time (respect to the initial time) of the detected inter-beat interval expressed in seconds (s). The second column is the duration in seconds (s) of the detected inter-beat interval (i.e., the distance in seconds from the previous beat).
- HR.csv: Average heart rate values computed in spans of 10 seconds extracted from the BVP signal. The first row is the initial time of the session expressed in UTC. The second row is the sample rate expressed in Hz.

- tags.csv: Event mark times. Each row corresponds to a physical button press on the device; the same time as the status LED is first illuminated. The time is expressed in UTC and it is synchronized with initial time of the session indicated in the related data files from the corresponding session

Activities performed and demographic data such as age, weight, and height are provided in the *subject-info.csv* file.

Additionally, a file containing all self-reported stress levels for each stage is provided (*Stress_level_v1.csv* and *Stress_level_v2.csv*).

Some participants experienced issues such as incorrect wristband placements, incomplete protocols, and connection problems. Details about these issues can be found in the *data_constraints.txt* file. These register have not taken in count for ML classifications but are provided in the dataset for research purpose.

4 Technical Validation

4.1 Signal Preprocessing and Feature Extraction

For signal filtering and preprocessing, we utilized Python libraries such as Pandas, SciPy and NumPy, along with open-source tools like NeuroKit2.

Tags were used to segment the signal according to the protocol. From each segment, we extracted relevant features to characterize the signals, aiming to distinguish between different conditions. These features were then used as input for training and evaluating different machine learning models.

The files we worked with included *BVP.csv, HR.csv, EDA.csv,* and *ACC.csv*, as well as the *tags.csv* file for segmentation purposes. In this work, we did not perform any temperature analysis. Additionally, we chose not to use the *IBI.csv* file provided by Empatica due to missing data, particularly for exercise recordings. A summary of the entire process, from data collection to classification, is illustrated in Figure 2.

4.1.1 Signal Preprocessing

Blood Volume Pulse refers to the variation in arterial blood volume under the skin resulting from the heart cycle. To remove noise, we applied a 4th-order Chebyshev Type II digital filter with a bandpass of 0.5–10 Hz.

Electrodermal Activity measures the variation in skin conductance in response to sweat gland activity, which is controlled by the sympathetic nervous system. This signal includes both a tonic component (skin conductance level, SCL) and a rapid phasic component (skin conductance responses, SCRs), which result from sympathetic neuronal activity. The raw EDA signal was filtered using a 5th-order Butterworth low-pass filter with a cutoff frequency of 1.99 Hz. Once filtered, a Savitzky-Golay filter was applied to separate the tonic and phasic components.

Heart rate variability is considered an indicator of mental stress and physical fitness. The standard method of obtaining HRV is through ECG, using the time interval between two consecutive R-peaks (*Inter Beat Interval, IBI)*. An alternative method is through Pulse Rate Variability, which measures the time interval between successive peaks in the PPG signal. BVP was filtered, and wave peaks were detected to

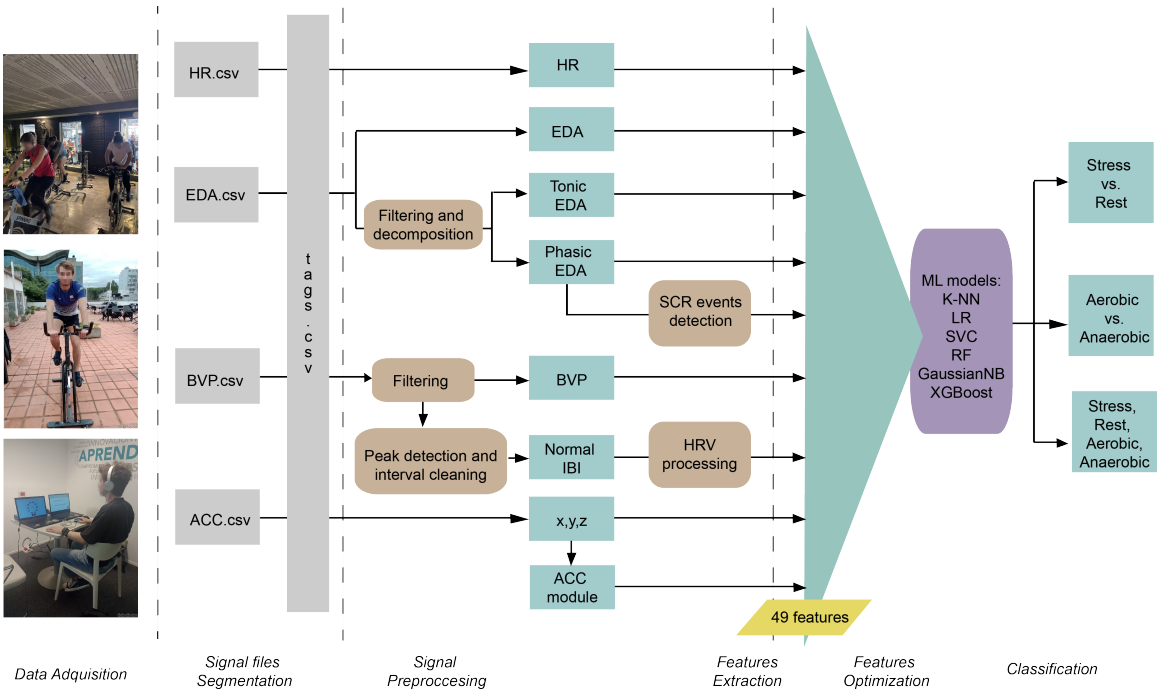

Fig. 2: Feature Extraction

obtain pulse intervals , equivalent to the RR intervals in ECG [27]. We then obtained the normal pulse-to-pulse intervals, discarding ectopic and incorrectly detected beats. Once we got the clean interval series, time domain and frequency domain features were computed.

No filtering was applied to the 3-axis accelerometer signals or the heart rate data provided by Empatica.

4.1.2 Feature Extraction

The features can be categorized into general statistical properties (mean and standard deviation), minimum and maximum change ratios(obtained from the 5% and 95% percentile of the first derivative, respectively), and specific signal characteristics.

- For the filtered BVP, the mean and standard deviation were calculated.
- For HR, the mean and standard deviation were obtained, along with the maximum and minimum change ratios.
- Statistical features were derived from the raw, phasic, and tonic EDA signals. The minimum and maximum change ratios was computed for the tonic component. For the phasic component, the Neurokit2 tool [28] was used to extract information from Skin Conductance Responses (SCR) events, including SCR amplitude, the samples at which SCR onsets and peaks occurred, SCR rise time (time from onset to peak), SCR recovery time (time from peak to the midpoint of the fall), and SCR height

(value from onset to peak). Additionally, SCR density was derived by dividing the number of SCR events by the segment length.

- Regarding HRV, time-domain features included mean, maximum and minimum IBI. The HR mean was computed from the IBI mean, which differs from the HR data provided for Empatica. Further features included SDNN (standard deviation of normal-to-normal intervals), RMSSD (root mean square of successive differences between normal beats), pNN50, and pNN20 (percentage of the difference associated with NN intervals differing by more than 50 ms and 20 ms, respectively) were calculated. In the frequency domain, the power and peaks of each frequency band — Very Low Frequency (VLF: 0-0.04 Hz), Low Frequency (LF: 0.04-0.15 Hz), High Frequency (HF: 0.15-0.4 Hz), and Very High Frequency (VHF: 0.4-2 Hz) — were calculated. Total power, normalized LF and HF, and the LF/HF ratio were also computed.
- For the accelerometer data, the mean and standard deviation of each axis, as well as the magnitude, were calculated. Additionally, the maximum and minimum change ratios derived from the acceleration magnitude were obtained.
  For a better comprehension, all extracted features are presented in Table 2.

Table 2: Features extracted

E4 Signal	Type	Features
BVP (2)	Statistical	*bvp_mean, bvp_std*
HR (4)	Statistical	*hr_mean, hr_std, hr_ratio_down, hr_ratio_up*
HRV (20)	Time Domain	*max_ibi, min_ibi, mean_ibi, hr_mean_ibi, pnn20, pnn50, rmssd, sdnn*
	Frequency Domain	*total_power, ratio, VLF_power, VLF_peak, LF_power, LF_peak, LH_n, HF_power, HF_peak, HF_n, VHF_power, VHF_peak*
EDA (13)	Statistical	*mean_raw_eda, std_raw_eda, mean_tonic_eda, std_tonic_eda, mean_phasic_eda, std_phasic_eda, tonic_ratio_down, tonic_ratio_up*
	SCR Events	*peaks_density, scr_mean_amp, scr_mean_height, scr_mean_risetime, scr_mean_recoverytime*
ACC (10)	Statistical	*x_mean, x_std, y_mean, y_std, z_mean, z_std*
	Magnitude	*acc_mean, acc_std, acc_ratio_down, acc_ratio_up*

4.2 Classification

4.2.1 Data preprocessing

To reduce inter-subject variability, intra-participant normalization was applied to all blocks within the stress protocol. No normalization was applied for exercise sessions and multiple classification as they were conducted on different days.

Stress blocks (Stroop, TMCT, Real, and Opposite Opinion, Subtract) were labeled as 1, and rest blocks (Baseline, First Rest and Second Rest) were labeled as 0. We used the second half of both the First Rest and Second Rest blocks to minimize the influence of any residual stress effects from the previous stress blocks. The resulting

data was imbalanced due to differences in protocol versions: 154 corresponding to stress blocks and 102 for rest periods.

For the exercise sessions, a binary classification between common aerobic slots and maximum power sprints was performed, due to variations in protocol versions. For aerobic protocols, a 1-minute middle segment from 70 rpm, 75 rpm, 80 rpm, and 85 rpm blocks was extracted for analysis. For anaerobic protocols, sprints were used. The resulting data was also imbalanced: 120 samples from aerobic blocks and 106 from high-intensity blocks.

4.2.2 Features selection

From an initial set of 49 features, we applied a high-correlation filter, discarding features with a Pearson correlation coefficient above 0.8. The Python open-source library *Sweetviz* was used for data exploration, providing insights into categorical associations, multicolinearity analysis and distribution of the data. This helped in selecting features that provided the most informative value for the target labels. Different sets of features were tested based on these criteria.

4.2.3 ML algorithms

For model training and testing, we used the Python Sklearn library. Binary classification was performed between stress vs. rest states, and aerobic segments vs. sprints. Additionally, a multi-class classification task was conducted with four labels: rest, stress, aerobic, and maximum power states.

The data were imputed and scaled (subtracting the mean and dividing by the standard deviation), and several machine learning algorithms were tested, including K-Nearest Neighbors, Logistic Regression, Gaussian Naive Bayes, Support Vector Classifier, Random Forest, and XGBoost. These models were evaluated using 10-fold cross-validation, following an 80/20 train-test split. Since the data were imbalanced, different resampling techniques were applied. The performance of the models was assessed using metrics such as accuracy, precision, recall, and F1 score.

4.3 Results

Out of the 36 volunteers participating in the study, all 36 completed the stress protocol, 31 completed the anaerobic session, and 30 completed the aerobic session. Some participants were unable to finish the exercise sessions. For classification purposes, 2 stress records were discarded due to incorrect wearable device placement for participant f07 and bad fit of the wristband for participant f13. Despite some incomplete exercise sessions, all participants performed the stages selected for classification. Although these stress records were excluded from the classification analysis, they are included in dataset with their corresponding comments and constraints.

4.3.1 Protocols

The stress induction protocol proved to be effective in generating stress during the tasks and relaxation during the rest periods. This is supported by the self-reported perceived stress levels shown in Figure 3 (v1) and Figure 4 (v2). To control for individual

differences in stress perception, we normalized the reported stress levels for each participant. For all statistical tests, a p-value (p) threshold of 0.05 was used to determine significance. Tests yielding $p < 0.05$ were considered statistically significant.

A Shapiro-Wilk test was used to assess the normality of the data for each block. Several blocks showed a non-normal distribution: Baseline ($p = 0.037$) and First_Rest ($p = 0.047$) in v1, and Real_Opinion ($p = 0.022$) in v2. Therefore, we applied the non-parametric Kruskal-Wallis test. Both stages demonstrated statistically significant differences between the blocks ($p = 6.24e\text{-}06$ for v1, and $p = 9.96e\text{-}16$ for v2).

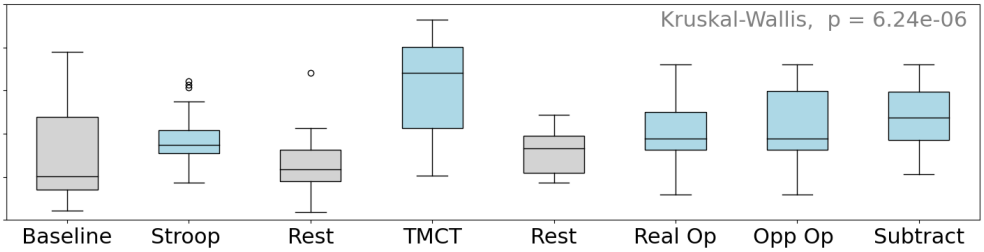

Fig. 3: Normalized Stress Auto Reported Level - First Stage *(v1)*

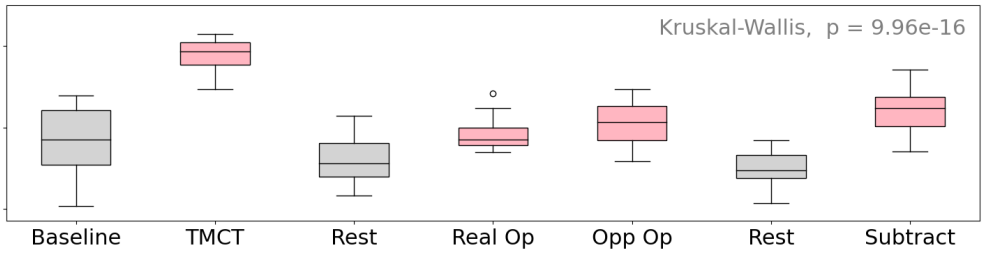

Fig. 4: Normalized Stress Auto Reported Level - Second Stage *(v2)*

Following this, Dunn's post-hoc test was applied to determine which groups were significantly different, as presented in Table 3. The Trier Mental Challenge Test resulted as the primary stress-inducing block in both data collection stages, followed by the Subtract Test, with both tasks involving mathematical challenges.

Aerobic exercise sessions involved sustained moderate activity for at least 20 minutes (excluding warm-up and cool-down periods). In contrast, anaerobic sessions consisted of at least three maximum-effort sprints, scattered with cool-down periods, reaching about 15 minutes of exercise. Both sessions are representative of common daily exercise routines.

4.3.2 Classification

After the feature optimization process, the original 48 features were reduced to 12, 11, and 18 features for binary stress classification, binary exercise classification, and four-label classification, respectively. XGBoost was the best-performing model across all tasks, achieving a maximum accuracy of 93% for the binary stress/rest classification using 10-fold cross-validation. In all cases, oversampling proved to be the most effective technique for handling the data imbalance. The classification performance metrics are detailed in Table 4.

5 Usage Notes

This data can be used to develop ML models for stress and exercise detection and classification, as well as for signal processing and feature extraction. For information on expected signals and recommended tools for signal processing, visit the Empatica website.

Pilot results on stress binary classification and feature extraction from the signals have been presented at conferences [29, 30]. The multimodal classification presented in this work extends classifciation to detect and differentiate among various physiological states, including aerobic and anaerobic exercise, based on the collected data.

Limitations (details about these issues can be found in the *data_constraints.txt*):

- Stress Session: S02 has duplicated data; f07 did not remove the wrist-band protection cover, so not all signals are valid; f14's data is split into two parts.
- Aerobic Session: S03 and S07 could not complete the procedure; S11's data is split into two parts; S12 did not perform this test.
- Anaerobic Session: S06 could not complete the procedure; S16's data is split into two parts.

6 Code Availability

A Jupyter Notebook (*Wearable_Dataset.ipynb*) is provided to open, read, and visualize the data. This can be downloaded from https://doi.org/10.5281/zenodo.13993658. To

Table 3: Significance Differences of Normalized Stress Levels Between Blocks. *(v1: First Stage; v2: Second Stage)*

	v1	**v2**
Baseline vs TCMT	*p=6.21e-4*	*p=7.00e-6*
First_Rest vs TMCT	*p=1.70e-5*	*p=9.53e-11*
Second_Rest vs TMCT	*p=7.82e-3*	*p=9.87e-14*
Real_Opinion vs TMCT		*p=5.20e-5*
Opposite_Opinion vs TMCT		*p=7.17e-3*
Second_Rest vs Real_Opinion		*p=3.76e-2*
First_Rest vs Opposite_Opinion		*p=1.78e-2*
Second_Rest vs Opposite_Opinion		*p=4.43e-4*
First_Rest vs Subtract Test	*p=6.23e-3*	*p=7.86e-4*
Second_Rest vs Subtract Test		*p=1.10e-5*

Table 4: XGBoost performance metrics with 10-fold cross validation

	Feature Set	Accuracy	Precision	Recall	F1
Stress / Rest	hr_mean, mean_raw_eda, mean_tonic_eda, std_tonic_eda, mean_recoverytime, max_ibi, ibi_mean, rmssd, ratio, LF_peak, x_std, y_std, z_std	93%	93%	92%	92%
Aerobic / Sprints	hr_std, std_phasic_eda, tonic_ratio_down, peaks_density, mean_recoverytime, min_ibi, pnn50, VHF_power, LF_n, z_std, acc_mean, acc_ratio_down	91%	92%	92%	91%
Stress / Rest / Aerobic / Sprints	hr_mean, hr_std, mean_tonic_eda, std_tonic_eda, mean_recoverytime, std_phasic_eda, tonic_ratio_down, peaks_density, max_ibi, ibi_mean, rmssd, min_ibi, LF_peak, LF_n, x_std, y_std, z_std, acc_mean, acc_ratio_down	84%	85%	84%	84%

execute the notebook, ensure that basic Python libraries such as pandas, os, numpy, time, and matplotlib are installed.

7 References

References

[1] Tagougui, S., Taleb, N., Molvau, J., Nguyen, É., Raffray, M., Rabasa-Lhoret, R.: Artificial Pancreas Systems and Physical Activity in Patients with Type 1 Diabetes: Challenges, Adopted Approaches, and Future Perspectives. Journal of Diabetes Science and Technology **13**(6), 1077–1090 (2019) https://doi.org/10.1177/1932296819869310

[2] Riddell, M., Zaharieva, D., Yavelberg, L., Cinar, A., Jamnik, V.: Exercise and the Development of the Artificial Pancreas: One of the More Difficult Series of Hurdles. Journal of Diabetes Science and Technology **9** (2015) https://doi.org/10.1177/1932296815609370

[3] D Crosswell, A., G Lockwood, K.: Best practices for stress measurement: How to measure psychological stress in health research. Health Psychology Open **7**(2) (2020) https://doi.org/10.1177/2055102920933072 https://doi.org/10.1177/2055102920933072

[4] Iqbal, T., Elahi, A., Redon, P., Vazquez, P., Wijns, W., Shahzad, A.: A Review of Biophysiological and Biochemical Indicators of Stress for Connected and Preventive Healthcare. Diagnostics **11** (2021) https://doi.org/10.3390/diagnostics11030556

[5] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect

Detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. ICMI '18, pp. 400–408. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3242969.3242985 . https://doi.org/10.1145/3242969.3242985

[6] Anwar, T., Zakir, S.: Machine Learning Based Real-Time Diagnosis of Mental Stress Using Photoplethysmography. In: Journal of Biomimetics, Biomaterials and Biomedical Engineering, vol. 55, pp. 154–167 (2022). https://doi.org/10.4028/p-01r9mn . Trans Tech Publ

[7] Izzah, N., Sutarto, A.P., Hariyadi, M.: Machine Learning models for the Cognitive Stress Detection Using Heart Rate Variability Signals. Jurnal Teknik Industri: Jurnal Keilmuan dan Aplikasi Teknik Industri **24**(2), 83–94 (2022) https://doi.org/10.9744/jti.24.2.83-94

[8] Markova, V., Ganchev, T., Kalinkov, K.: CLAS: A Database for Cognitive Load, Affect and Stress Recognition. In: 2019 International Conference on Biomedical Innovations and Applications (BIA), pp. 1–4 (2019). https://doi.org/10.1109/BIA48344.2019.8967457

[9] Beh, W.-K., Wu, Y.-H., An-Yeu, Wu: MAUS: A Dataset for Mental Workload Assessmenton N-back Task Using Wearable Sensor . arXiv (2021). https://doi.org/10.48550/ARXIV.2111.02561 . https://arxiv.org/abs/2111.02561

[10] Chaptoukaev, H., Strizhkova, V., Panariello, M., Dalpaos, B., Reka, A., Manera, V., Thummler, S., Ismailova, E., Evans, N.W., Bremond, F., Todisco, M., Zuluaga, M.A., Ferrari, L.M.: StressID: a Multimodal Dataset for Stress Identification. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023). https://openreview.net/forum?id=qWsQi9DGJb

[11] Amin, M., Wickramasuriya, D., Faghih, R.: A Wearable Exam Stress Dataset for Predicting Grades using Physiological Signals. In: Healthcare Innovations and Point of Care Technologies Conference, HI-POCT 2022. Healthcare Innovations and Point of Care Technologies Conference, HI-POCT 2022, pp. 30–36 (2022). https://doi.org/10.1109/HI-POCT54491.2022.9744065

[12] Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems **6**(2), 156–166 (2005) https://doi.org/10.1109/TITS.2005.848368

[13] Rossi, A., Da Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sîrbu, A., Clifton, D.A., Martini, C., Morelli, D.: A Public Dataset of 24-h Multi-Levels Psycho-Physiological Responses in Young Healthy Adults. Data **5**(4) (2020) https://doi.org/10.3390/data5040091

[14] Mehrgardt, P., Matloob, K., Poon, S., Withana, A.: Pulse Transit Time PPG

Dataset (version 1.1.0). PhysioNet. https://doi.org/10.13026/jpan-6n92 (2022)

[15] Karas, M., Urbanek, J., Crainiceanu, C., Harezlak, J., Fadel, W.: Labeled raw accelerometry data captured during walking, stair climbing and driving(version 1.0.0). PhysioNet. https://doi.org/10.13026/51h0-a262 (2021)

[16] Banos, O., Toth, M., Amft, O.: REALDISP Activity Recognition Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5GP6D (2014)

[17] Cleland, I., Donnelly, M.P., Nugent, C.D., Hallberg, J., Espinilla, M., Garcia-Constantino, M.: Collection of a Diverse, Realistic and Annotated Dataset for Wearable Activity Recognition. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 555–560 (2018). https://doi.org/10.1109/PERCOMW.2018.8480322

[18] Birjandtalab, J., Cogan, D., Pouyan, M.B., Nourani, M.: A Non-EEG Biosignals Dataset for Assessment and Visualization of Neurological Status. In: 2016 IEEE International Workshop on Signal Processing Systems (SiPS), pp. 110–114 (2016). https://doi.org/10.1109/SiPS.2016.27

[19] Parent, M., Albuquerque, I., Tiwari, A., Cassani, R., Gagnon, J.-F., Lafond, D., Tremblay, S., Falk, T.H.: PASS: A Multimodal Database of Physical Activity and Stress for Mobile Passive Body/ Brain-Computer Interface Research. Frontiers in Neuroscience **14** (2020) https://doi.org/10.3389/fnins.2020.542934

[20] Stroop, J.: Studies of Interference in Serial Verbal Reactions. Journal of Experimental Psychology: General **121**, 15–23 (1992) https://doi.org/10.1037/0096-3445.121.1.15

[21] Macleod, C.: Half A Century of Research on the Stroop Effect - An Integrative Review. Psychological bulletin **109**, 163–203 (1991) https://doi.org/10.1037/0033-2909.109.2.163

[22] Stoet, G.: PsyToolkit - A software package for programming psychological experiments using Linux. Behavior Research Methods **42**, 1096–1104 (2010) https://doi.org/10.3758/BRM.42.4.1096

[23] Stoet, G.: PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. Teaching of Psychology **44**, 24–31 (2017) https://doi.org/10.1177/0098628316677643

[24] Pruessner, J.C., Hellhammer, D.H., Kirschbaum, C.: Low self-esteem, induced failure and the adrenocortical stress response. Personality and Individual Differences **27**(3), 477–489 (1999) https://doi.org/10.1016/S0191-8869(98)00256-6

[25] Inbar, O., Bar-Or, O., Skinner, J.S.: The Wingate Anaerobic Test. Human Kinetics, Champaign, IL (1996)

[26] Harris, G.D.: In: Evans, C.H., White, R.D. (eds.) Exercise Testing Special Protocols, pp. 45–54. Springer, New York, NY (2009). https://doi.org/10.1007/978-0-387-76597-6_3

[27] Kavsaoğlu, A., Polat, K., Bozkurt, M.: An innovative peak detection algorithm for photoplethysmography signals: An adaptive segmentation method. Turkish Journal of Electrical Engineering and Computer Sciences **24**, 1782–1796 (2016) https://doi.org/10.3906/elk-1310-177

[28] Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C., Chen, S.H.A.: NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods **53**(4), 1689–1696 (2021) https://doi.org/10.3758/s13428-020-01516-y

[29] Hongn, A., Prado, L.E., Bosch, F., Bonomini, M.P.: Wearable device dataset for stress detection. In: Ferrández Vicente, J.M., Val Calvo, M., Adeli, H. (eds.) Bioinspired Systems for Translational Applications: From Robotics to Social Engineering, pp. 518–527. Springer, Cham (2024)

[30] Hongn, A., Bosch, F., Prado, L.E., Ferrández, J.M., Bonomini, M.P.: Non-invasive recording of physiological variables under stress conditions and aerobic and anaerobic physical activity. In: Ballina, F.E., Armentano, R., Acevedo, R.C., Meschino, G.J. (eds.) Advances in Bioengineering and Clinical Engineering, pp. 30–39. Springer, Cham (2024)

8 Author Contributions

All authors participated in the design of the study protocol. **A.H., F.B. and L.E.P.** conducted the data collection and performed the technical validation. **A.H.** made data publicy available and drafted the manuscript. **J.M.F. and M.P.B.** supervised the project and reviewed the manuscript.

9 Acknowledgements

This research has been funded by European Commission HORIZON-MSCA-2022-SE-01 EPISTEAM.

10 Competing Interests

The authors declare no competing interests.