Comparison of wearable and clinical devices for acquisition of peripheral nervous system signals

Andrea Bizzego [1], Giulio Gabrieli [2],
Cesare Furlanello [3], Gianluca Esposito [1,2,4]

[1]Department of Psychology and Cognitive Science, University of Trento, Italy
[2]Psychology Program, School of Social Sciences, Nanyang Technological University, Singapore
[3]HK3 Lab, Rovereto, Trento, Italy
[4]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

Abstract

A key access point to the functioning of the Autonomic Nervous System is the investigation of peripheral signals. Wearable Devices (WDs) enable the acquisition and quantification of peripheral signals in a wide range of contexts, from personal uses to scientific research. WDs have lower costs and higher portability than medical-grade devices. But achievable data quality can be lower, subject to artifacts due to body movements and data losses. It is therefore crucial to evaluate the reliability and validity of WDs before their use in research. In this study we introduce methods and scientific software for the assessment of WDs for multivariate physiological signals. The quality of cardiac and Electrodermal Activity signals is validated with a standard set of Signal Quality Indicators, available as open source in the pyphysio Python package. We apply the indicators for the analysis of signal quality on data simultaneously recorded from a clinical-grade

device and two WDs. The dataset provides signals of 6 different physiological measures collected from 18 subjects with WDs. This study indicates the need of validating the use of WD in experimental settings for research and the importance of both technological and signal processing aspects to obtain reliable signals and reproducibility of results.

wearable devices, physiological data analysis, signal processing, multivariate analysis

1 Introduction

The quantification of peripheral physiological nervous signals is a core step to measure the functioning of the Autonomic Nervous System [19]. Being able to observe these phenomena in real-life and without constraints imposed by laboratory settings is a key reason for adopting WDs in scientific research [13]. Wearable technologies enable the acquisition and quantification of physiological signals in a wide range of contexts, from personal uses to industrial and scientific research [32, 9, 23]. In a general sense, Wearable Devices (WDs) are portable, non-invasive devices that allow the acquisition of physiological signals during daily life, with no need of external equipment [4]. The application of WDs is boosted by their improved portability (increasing miniaturization and battery life) [10], as well by new materials and layouts with reduced invasiveness, e.g.: epidermal tattoos [3], wireless ring pulse oximeter [16] and garments [11].

Besides higher portability, WDs may have an apparent lower cost than clinical-grade devices, which enabled their diffusion for commercial applications [29]. In contrast to growing market availability, the adoption of WDs in research is still limited. Wearable technologies could open to the scientific community novel tools in the study of human physiology and autonomic responses [20, 19], paving the way for a new generation of experiments in which long-term monitoring and real-life ecological acquisitions are key aspects. In particular, a decrease in obtrusiveness on subject's behaviour and daily activities is expected by adopting WDs instead of clinical devices [13]. Many scientific fields would benefit from the adoption of WDs for long-term, not obtrusive recording of physiological signals. In particular, scientific use of WDs is explored in affective computing [25], research on autism [32], interpersonal coupling [2, 6] and psychology in general.

However, due to the technical constraints imposed by improvements in miniaturization and autonomy of WDs, the data quality that can be achieved is typically lower and more subject to noise due to body movements [24, 36, 34]. Further, we need to consider the hurdles in extracting, preprocessing and analyzing the data from non-standard data interfaces: high scientific "Total Cost of Ownership", possibly requiring manual preprocessing is also associated to limited reproducibility of research.

It is therefore necessary to create a working protocol and resources to validate WDs as a reliable scientific tool, before their use in research [17]. In particular, the validation should be based on open access scientific software and include a comparison with a baseline of results from signals collected by medical-grade devices, within the same or very similar experimental settings. In practice, a battery of signal processing algorithms or their composition in a pipeline should be applied to signals collected by clinical-grade and wearable devices, with a lower level of precision accepted considering the type and purpose of application. However, despite its importance, the validation of a WD is often omitted or left to manufacturers, who rarely provide the dataset used and the details about the validation procedure. Although several datasets with physiological signals are available, such as DEAP [18], MANHOB-HCI [28], SEMAINE [22] and the PhysioBank archive [15], none of the existing resources provides signals collected from both wearable and clinical grade devices to allow the comparison. Finally the set of algorithms applied in the validation should be also available for reproducibility and further reference.

In this study we introduce scientific software and data aimed at validating the use of WD for research within different real-life contexts (resting and movement), focusing on cardiac and electrodermal activity signals. Our contributions are: (a) creation of the Wearable and Clinical Signals (WCS) dataset which allows comparing a representative set of WD with clinical-grade devices; (b) a procedure to assess the validity of signals collected from WD; and (c) quantitative results on physiological signals of interest for the affective-computing field.

Physiological signals were collected from a clinical-grade device and from two WDs to allow both the direct comparison of the signals' data quality and the investigation of the reproducibility. The adopted protocol comprises two different activities to evaluate the performance in different real-life contexts.

2 Methods

2.1 Participants

A total of 18 participants were recruited for the experiment through announcements among university students. The participation to the experiment was voluntary and granted credits for the fulfillment of an academic course. Before starting the experiment each participant was briefed about the purpose and procedure of the experiment and gave the informed consent. The experiment was approved by the Independent Review Board of the University of Trento (2017-019) and was conducted according to the principles of the Declaration of Helsinki. All the data have been de-identified by assignment of randomized numeric identifier to each subject and we removed any reference to the absolute timestamps.

2.2 Devices and architecture

The physiological signals were simultaneously collected by three devices (see Figure 1): the FlexComp unit which provided the reference signals, and two WDs: the Empatica E4 and the ComfTech HeartBand.

The FlexComp acquisition unit was connected through optical fibre to a signal converter which in turn was connected to a Windows 10 workstation where the proprietary BioGraph Infinity Software Platform managed the acquisition and collection of the signals from the unit.

The two WDs sent the data through Bluetooth connection to an Android (5.1) tablet (Samsung Tab A) where a specifically developed app gathered the signals and sent them to a database through secure connection. The app interface also allows the annotation of manual markers to allow the synchronization of the signals collected with the FlexComp and WDs.

Before starting the experiment, the signals acquired through the two WDs were visually inspected by activating the real-time plotting function.

The WDs used in the experiment were selected according to four requirements we defined to explicitly highlight the key characteristics that allow the usage in everyday life:

1. Wearability: the device should be easily worn by the subject with no need of applying conductive gels, wired electrodes or similar preparations. In fact, in everyday applications the presence of an external

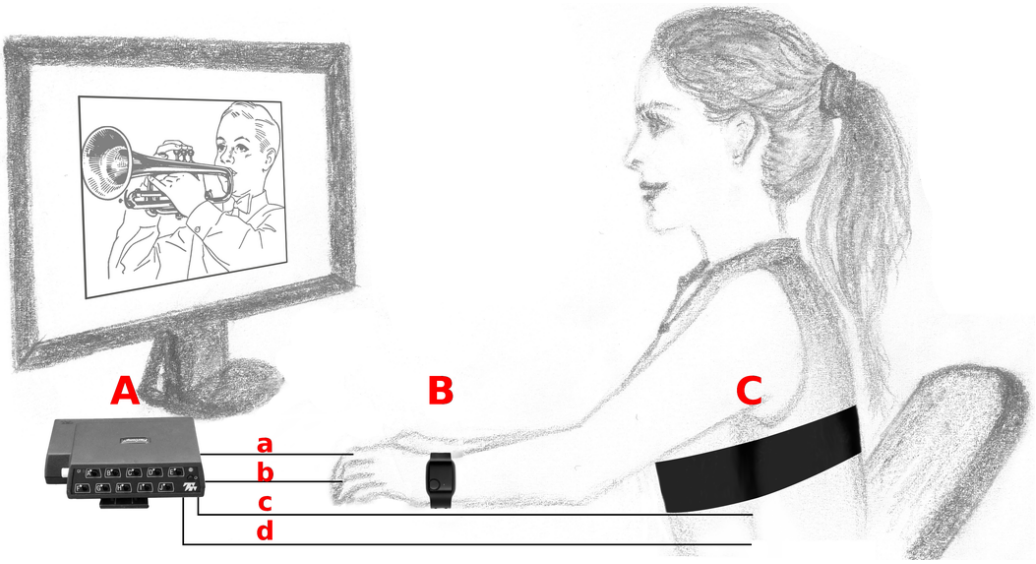


Figure 1: Illustration showing the devices used for the acquisition of physiological signals. A: Thought Technology FlexComp (picture from www.thoughttechnology.com) with four wired sensors: a) Electrodermal activity, b) Blood Volume Pulse, c) Electrocardiogram, d) Respiration; B: Empatica E4 (picture from www.empatica.com); C: ComfTech HeartBand.

supervisor over-viewing the preparation (e.g. experimenter) is not expected;

2. Streaming: the device should stream the data in real-time to an external collector through Bluetooth connection. Streaming capability is a key feature which allows real-time processing and feedback, which is fundamental in remote health monitoring applications;

3. Availability: the device should be commercially available at the moment of testing (prototypes, proof-of-concepts or custom devices were not considered). This requirement is imposed to restrict the selection to state-of-art technologies, which could be accessed by researchers with any technical and/or technological competences;

4. Performances: the final choice should favour the device with higher sampling rate and higher number of sensors to select the most advanced technological solution.

After a preliminary phase in which a list of WDs was tested, we selected two devices, that full-filled all the requirements: the Empatica E4 and the Comftech HeartBand. Empatica E4 was chosen as a representative of wristband devices, while Comftech HeartBand belongs to the smart-garments category.

2.2.1 Reference: Thought Technology FlexComp

The clinical device we used as reference was the Thought Technology FlexComp unit (commercial code: T7555M). It is a customizable acquisition unit which provides up to 10 input slots that can be used to connect diverse physiological sensors. Maximal sampling frequency is 2048 Hz, which was adopted for the experiment, with 14 bits of resolution for each input. The physiological signals acquired with the FlexComp unit were:

1. Electrocardiogram (ECG): using three electrodes placed over the left and right coracoid processes and below the ribs on the left. The signal is pre-amplified and filtered by the EKG Sensor (T9306M) which returns a single channel read in millivolts. UniGel electrodes (T3425) are used as conductive mean between the sensor and the skin;

2. Electrodermal Activity (EDA): two finger bands with Ag-AgCl electrodes (SA2659) are placed on the second and fourth finger of the left hand and connected to the sensor (SA9309M). The skin conductance is measured in microSiemens (μS);

3. Blood Volume Pulse (BVP): the sensor (SA9308M) is placed on the third finger of the left hand. The relative amount of reflected infrared light is measured;

4. Respiration (RESP): a band is worn on the chest to measure the relative volumetric expansion by elongation of an elastic patch (SA9311M);

5. Trigger (TRG): a handle with a button to generate electrical impulses used to manually mark the experimental events.

2.2.2 Empatica E4

The Empatica E4 [14] is a multi-sensor wristband designed for real-life acquisitions of physiological signals. It has streaming (through Bluetooth Low Energy) and storing (internal flash memory) capacity; communication with an Android smartphone is provided by the EmpaLink Software Development Kit which is used by the PhysioREC app. Collected signals are:

1. BVP: four Light Emitting Diodes (LEDs) are used to generate light at two different wavelengths (Green and Red) and two photodiodes are used to measure reflected light. Using two wavelengths and an appropriate proprietary algorithm to preprocess the signals should reduce motion effects and sensitivity to external sources of light. Sensors are placed on the bottom of the wristband case in firm contact to the skin; signal is sampled at 64 Hz;

2. EDA: two stainless steel electrodes are placed on the band to allow positioning on the inner side of the wrist. Skin conductance is measured in microSiemens at 4 Hz sampling rate;

3. Acceleration (ACC): three axes acceleration (range $\pm$2g) is measured at 32 Hz sampling frequency;

4. Skin Temperature (ST): measured by an infrared thermopile placed on the back of the case, at 4 Hz sampling frequency.

2.2.3 ComfTech HeartBand

The Comftech HeartBand is an elastic band with embedded tissue electrodes and an acquisition unit that is connected to the band through two snaps. Streaming is based on Bluetooth 2.0 and the documentation to decode the hexadecimal messages from the device was provided by the manufacturer.

The collected signals are:

1. ECG: using two tissue electrodes placed over the chest and connected to the acquisition unit where the signal is sampled (128 Hz), pre-amplified and filtered to return a single channel read;

2. ACC: three axes acceleration is measured by a sensor embedded on the acquisition unit; sampled at 200 Hz;

2.3 Experimental protocol and procedure

After the signed informed consent was given, the experimenter instructed the participant about the placement of WDs and the experimental settings. The setup of the FlexComp unit, of the acquisition software and the positioning of the sensors was performed by the experimenter to ensure high quality signals. Regarding the WDs, the participant was asked to place himself the devices, after having received detailed instructions from the experimenter. This procedure reflects the idea that WDs are to be used in real-life context, where the presence of an experimenter is not expected. In particular, the participant was allowed to adjust the positioning of both the Empatica E4 and of the ComfTech HeartBand to favor the comfort during the experiment. However, the placement of the E4 wristband was inspected by the experimenter to ensure the device was not worn too tight (which would have prevented the physiological blood flow) or too loose (which would have prevented the smartwatch case to be in contact with the skin) to guarantee optimal conditions for a correct BVP signal acquisition.

The experiment comprises two phases: baseline and movement. During baseline the participants were asked to remain sit and still. The signals collected during this phase represent a reference for the maximal signal quality that can be achieved by WDs, as all external sources of noise that could affect the signal quality (such as moving artifacts, sensor displacement or detachment) are experimentally avoided. This condition is far from the real-life context in which the WDs are expected to be used, but it is needed to

isolate the effects of technical limitations and constraints from other causes of errors.

The effects of body movements, instead, can be better evaluated on the movement phase which concludes the experiment. During this phase the subject is asked to stand and simulate a walking on place, as the free movement was prevented by the wires connecting the medical-grade sensors with the FlexComp unit. The experimenter suggested moving naturally to replicate the intensity of usual walking. In particular, special attention was paid to movements of the left hand as, due to presence of sensors, participants tended to keep it still and near to the body. This experimental phase is expected to be more similar to real-life contexts and can be used to evaluate performances of algorithms for every-day applications, such as health monitoring, fitness and stress detection.

2.4 Preprocessing

The raw data have been preprocessed in order to appropriately format the data for the distribution and signal analysis. The first step was the synchronization of the two signal flows, from the FlexComp device and from the WDs. After the synchronization the absolute timestamps of the samples were converted into elapsed time since the start of the experiment (start of the baseline). Based on the TRG signal the instants of the beginning and end of each experimental phase were extracted and used to split the signals. Then each portion of signal was separately saved. The acceleration signals, which are three-dimensional, were converted into uni-dimensional signals by computation of the acceleration module.

Due to the sensitive nature of the data, the dataset is not public, but the collected signals are available to researchers aiming to create and test new algorithms for the automatic identification of noise and artifacts: https://doi.org/10.21979/N9/42BBFA.

2.5 Signal Quality analysis

Assessing the quality of the collected signals is crucial for reproducible research; this step is even more important when working with WDs, as signals are more sensitive to artifacts and noise due to body movements and technical constraints (see Figure 2).

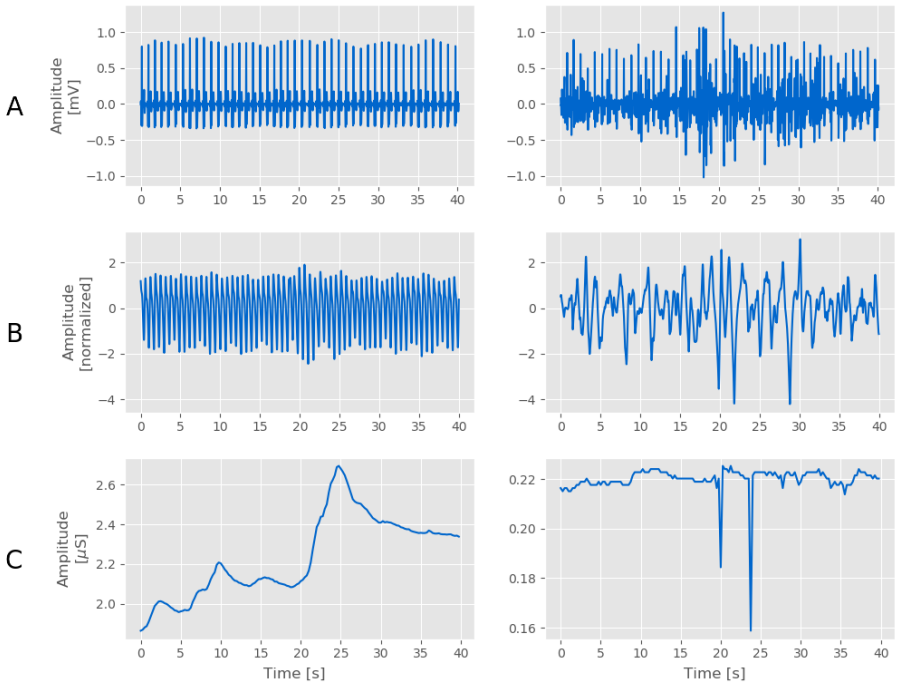


Figure 2: Portions of signals from WDs extracted from the WCS dataset to show examples of good (left column) and bad (right column) quality signals. A: ECG signals collected by the HeartBand; B) BVP signals collected by the E4; C) EDA signals collected by the E4.

To this aim, a set of Signal Quality Indicators (SQIs) has been defined to compare the quality of signals from clinical-grade and wearable devices. The SQIs have been computed on all cardiac signals and EDA signals of the baseline and movement sessions. The processing pipeline was developed in Python, based on pyphysio [7], and it is publicly available at https://gitlab.com/abp-san-public/wearable-clinical-devices.

2.5.1 Cardiac signals

Several Signal Quality Indicators (SQIs) have been proposed in literature to evaluate the quality of cardiac signals. The SQIs can be grouped in two main categories: indicators that require a previous detection of the heart beats, and indicators that are computed directly on the signal. We first used two SQIs from the second category to obtain an overall quantification of the signal quality, then adopt a specialized beat detection algorithm to assess the possibility of effectively identify the heart beats.

We used the Kurtosis (K) and the Spectral Power Ratio (ψ) which have been proposed for both the ECG [21, 5] and the BVP [12, 33]. Good quality signals are expected to have $K > 5$ in case of ECG [21], while $K < 3.5$ in case of BVP [27].

The Spectral Power Ratio is the ratio between the spectral power in two frequency bands F_1 and F_2: $\psi = \frac{\sum_{F_1} PSD_f}{\sum_{F_2} PSD_f}$. F_1 is associated with the components in the signals that convey the information about the heart beats (the QRS complexes in the ECG and the pulses in the BVP); F_2 is the band which contains all the main frequency components of the signals. For both the ECG ($F_1 = [5, 14], F_2 = [5, 50]$, [21]) and BVP ($F_1 = [1, 2.25], F_2 = [0, 8]$ [12]) ψ is expected to have values $0.5 \leq \psi \leq 0.8$ [21].

To compute the Signal Quality Indicators (SQIs) the signal is resampled (128 Hz) and filtered (band pass filter: [0.5, 50] Hz). Then the signal is standardized and segmented with 5 s length non-overlapping windows. The SQIs K and ψ are then computed for each segment and then averaged for each signal.

We then investigated the reliability of the acquired cardiac signals in measuring the cardiac activity. To this aim we applied a beat detection algorithm to identify the heart beats and derive the Inter Beat Intervals (IBIs) series, which is the series of distances between consecutive beats. The validation metrics are then computed by comparing the IBIs series with the

reference IBIs series obtained from the ECG signal of the Flex Comp, which were manually validated. These metrics are different from the SQIs presented before, as they are computed on a derived signal and not on the original cardiac signal, and thus the results depend also on the performance of the beat detection algorithm.

The ECG signals from the Flex Comp unit were used to obtain the reference IBIs series used for the validation: heart beats were detected with the Adaptive Peak Detection (APD) algorithm (see 2.5.2) and then visually inspected to manually correct false and misdetections.

The APD algorithm was applied also to the ECG signals of the Heart Band, while the Derivative-Based beat detection algorithm [8] was used to derive the IBIs series from the BVP signals of both the Flex Comp and E4. All IBIs series were then processed to find detection errors by application of the Adaptive Outlier Detection (AOD) algorithm (see 2.5.3). In summary, for each subject we obtained the reference IBIs series (IBI_r) and those derived from the ECG signal of the Heart Band (IBI_{HB}), from the BVP signal of the Flex Comp (IBI_{FC}) and from the BVP signal of the E4 (IBI_{E4}).

The quantification of the reliability of the beat detection was assessed in terms of precision (P) and recall (R) of the detected beats and by root mean square error (e_{RMSE}), by comparing each derived IBIs series (IBI_d) with the IBI_r.

To compute these metrics, we paired each beat b_d of the IBI_d to a beat b_r of IBI_r. The pairing is considered valid if $|b_d - b_r| < 0.5$ s. The beats that were successfully paired are considered True Positives (TPs), the remaining unpaired beats in IBI_d are considered FPs and the remaining unpaired beats In IBI_r are considered as FNs. Then we counted the number of TPs (n_{TP}), the number of FPs (n_{FP}) and the number of FNs ((n_{FN})) to compute the overall precision $P = n_{TP}/(n_{TP} + n_{FP})$ and recall $R = n_{TP}/(n_{TP} + n_{FN})$.

The RMSE is computed on paired beats: $e_{RMSE} = \sqrt{\frac{\sum_{TPs}(IBI_{di} - IBI_{ri})^2}{n_{TP}}}$.

2.5.2 Adaptive Beat Detection

The Adaptive Beat Detection (ABD) algorithm applied to the ECG signal is composed of two steps:

- Computation of the local range of the signal;
- Peak detection with threshold from the local range.

In the first step the signal is segmented by windowing (windows width: 1 s, overlap: 0.5 s) and the range is computed for each segment. The result is the estimation of the local range of the signal.

In the second step the signal is scan to find local maxima; a local maximum is considered a heart beat (i.e.: R peak of the QRS complex) if it is followed by a local minimum and the difference is greater than 0.7 times the local range of the signal.

2.5.3 Adaptive Outlier Detection

The AOD uses a fixed size cache vector $IBI_c = (ibi_1, ibi_2, ..., ibi_k)$ to store the last k valid IBI values, in order to adapt to IBI variability. The size k is empirically set to 5 and the cache is initialized with the median value of the IBIs series. Outliers detection is also regulated by the sensitivity parameter ϕ, which is empirically set to 0.25.

A detected beat is considered valid if its corresponding IBI value is within the interval $[(1\text{-}\phi)ibi_{median}, (1+\phi)ibi_{median}]$, where ibi_{median} is the median of the values in IBI_c. When a new valid beat is detected, its IBI value is used to update IBI_c using the First-In-First-Out rule. IBI_c is re-initialized when k consecutive non valid beats are detected.

2.5.4 Electrodermal Activity signals

Only few studies address the problem of assessing the quality of EDA signals as this is a quite recent issue, associated with the emergence of wearable devices for EDA acquisition. The main approaches are based on machine learning to classify good and bad quality portions of the EDA starting from a list of computed metrics [30, 35], but without identifying the most appropriate that, alone, could be used as SQI. On the other hand, these models are not very useful with signals from new devices as they need to be re-trained on new data that would require a manual labelling process.

Quality metrics of electrodermal activity signals In this study we adopted two metrics proposed by Kleckner and colleagues [17]: the Ratio of Out-of-Range samples (r_{oor}) and the number of jumps, or drops, in the signal (n_j).

We define r_{oor} as the ratio between the number of samples with value exceeding 60 μS or below 0.05 μS and the total number of samples. We

empirically defined the threshold $r_{oor} = 0.05$ to identify good ($r_{oor} leq 0.05$) and bad ($r_{oor} > 0.05$) signals..

A *jump* is defined in [17] as a portion of the signal in which the absolute value of the derivative of the signal is grater than 10 $\mu S/s$. This threshold was defined on signals sampled at 32 Hz, but the EDA signal from the Empatica is acquired at 4 Hz which make the 10 $\mu S/s$ threshold too high to allow the correct identification of the jumps. In addition, the amplitude of the EDA signal, and thus its derivative, also depends on the characteristics of the device, such as: amplification, type of electrodes and positioning. For these reasons we derived a normalized version of the metric proposed by Kleckner and colleagues.

We first applied a low pass filter to the EDA signal (cutoff frequency 0.05 Hz) to remove the trend and the long-term components, obtaining the filtered signal (s_f); then computed the normalized derivative $d_n = \frac{|s_f{}'|}{\sigma_s}$, where s'_f is the derivative of the filtered signal and σ_s is the standard deviation. Then we considered as jumps the samples where d_n is greater than 10. We considered as acceptable a total number of 5 jumps over the 5 minutes sessions of the baseline and movement: therefore signals with $n_j \leq 5$ are considered as good quality.

In addition, we defined a new metric, called Signal Activity metric (α), to account for the cases in which the amplitude of the peaks associated to the Galvanic Skin Response is too low or the peaks are not present (see Figure 2C). We defined $\alpha = \sqrt{\frac{\sum_n s_f n^2}{N}}$ where s_f is the filtered signal and N is the number of samples. As a threshold value we selected $\alpha = 0.05$ μS which is the minimal amplitude for a peak in the EDA signal to be considered as part of the Galvanic Skin Response [1]. EDA signals with $\alpha < 0.05$ μS are considered of bad quality.

Before computing the SQIs, the EDA signals are filtered (low-pass filter, cutoff frequency $f_{lp} = 1.5$ Hz). The EDA signals from the Flex Comp device are first downsampled to 4 Hz to allow the comparison with the signals from the Empatica E4. The SQIs are computed on segments of the signals obtained from non-overlapping 20 s windows. The SQIs r_{oor} and α computed on each segment are then averaged for each signal, while the n_j values are summed to obtain the total number of jumps for each session.

Table 1: Kurtosis for each signal and subject. In bold are the values that fail to meet the requirements for good quality.

Subject	ECG				BVP			
	Baseline		Movement		Baseline		Movement	
	FC	HB	FC	HB	FC	HB	FC	HB
1	10.88	11.27	7.38	9.11	0.53	-0.95	-0.23	0.11
2	19.53	9.90	11.13	9.16	-0.64	-1.05	-0.10	-0.28
3	8.01	17.33	**1.99**	7.53	0.06	-0.54	2.54	0.32
4	7.60	8.75	**0.03**	7.65	0.62	-0.51	0.50	0.35
5	17.71	7.24	12.32	8.72	-0.64	-0.80	-0.23	0.53
6	21.26	11.21	14.42	10.81	-0.62	-0.73	0.02	0.31
7	23.41	12.04	18.65	8.80	1.77	-0.66	-0.10	0.44
8	5.20	12.49	**3.77**	5.40	-0.29	-0.84	-0.89	-0.26
9	18.40	14.37	7.09	**0.03**	-1.04	-0.98	1.51	0.32
10	12.78	10.86	5.76	9.30	-0.29	-0.68	0.91	1.09
11	11.28	11.18	7.17	9.51	-1.17	-0.91	-0.63	0.54
12	19.40	15.48	**4.44**	11.78	-0.76	-0.71	0.09	0.52
13	13.73	12.56	5.79	11.56	-1.10	-0.53	-0.69	1.16
14	18.77	12.05	11.08	6.69	-0.61	-0.11	0.31	0.94
15	22.75	13.97	6.06	7.32	1.82	-0.90	0.03	0.54
16	13.15	9.52	9.60	7.12	-0.59	-0.89	0.19	0.94
17	18.11	13.66	15.68	13.07	-0.75	-0.44	0.12	0.09
18	7.55	11.41	**0.36**	8.61	-1.25	-0.91	-0.54	-0.05

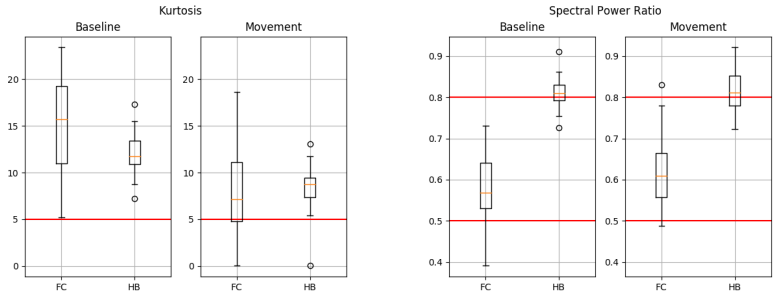
3 Results

3.1 Cardiac signals

In general, the ECG signals collected with both the Flex Comp and the Heart Band showed good quality in terms of Kurtosis (Table 1 and Figure 3a) although body movements cause a degradation; ECG signals from the HeartBand seems to contain spurious frequency components as indicated by the higher SPR (see Table 2 and Figure 3b). The BVP signals from both the Flex Comp and the E4 showed a very good Kurtosis; however the low SPR for the movement session indicates that this signal is highly affected by body

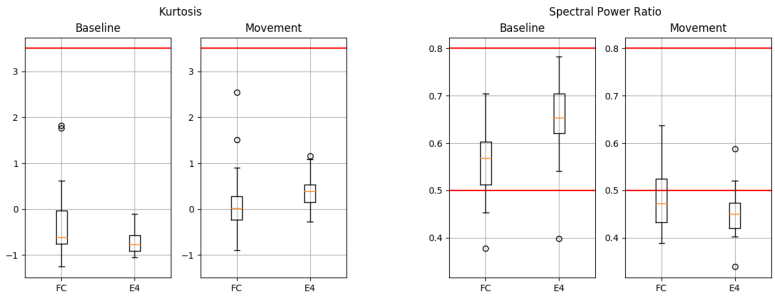
Table 2: Signal Power Ratio for each signal and subject. In bold are the values that fail to meet the requirements for good quality.

Subject	ECG				BVP			
	Baseline		Movement		Baseline		Movement	
	FC	HB	FC	HB	FC	HB	FC	HB
1	0.65	0.86	0.63	0.86	0.50	0.62	**0.49**	**0.42**
2	0.54	0.76	0.56	0.74	0.64	0.78	**0.40**	**0.41**
3	0.55	0.82	0.61	0.84	**0.38**	**0.40**	**0.43**	**0.42**
4	**0.39**	0.73	0.83	0.72	0.52	0.64	**0.49**	0.51
5	0.59	0.91	0.60	0.87	0.59	0.75	0.53	0.46
6	0.50	0.76	**0.49**	0.72	0.57	0.70	**0.45**	**0.42**
7	0.53	0.82	0.55	0.83	0.50	0.69	0.51	**0.48**
8	0.66	0.82	0.65	0.87	0.57	0.63	0.56	0.52
9	0.54	0.79	0.60	0.83	0.61	0.66	**0.39**	**0.46**
10	0.73	0.82	0.76	0.80	0.54	0.78	**0.39**	**0.34**
11	0.62	0.85	0.61	0.84	0.66	0.71	0.64	**0.47**
12	0.61	0.79	0.64	0.78	0.51	0.55	**0.46**	**0.40**
13	0.50	0.80	0.52	0.78	0.57	0.65	0.54	**0.43**
14	0.70	0.83	0.71	0.92	0.61	0.63	**0.46**	**0.48**
15	0.59	0.84	0.67	0.86	**0.45**	0.61	**0.46**	**0.43**
16	0.53	0.80	0.56	0.78	0.59	0.73	**0.49**	**0.46**
17	0.48	0.80	**0.49**	0.79	0.57	0.54	**0.41**	**0.44**
18	0.71	0.79	0.78	0.79	0.70	0.65	0.62	0.59



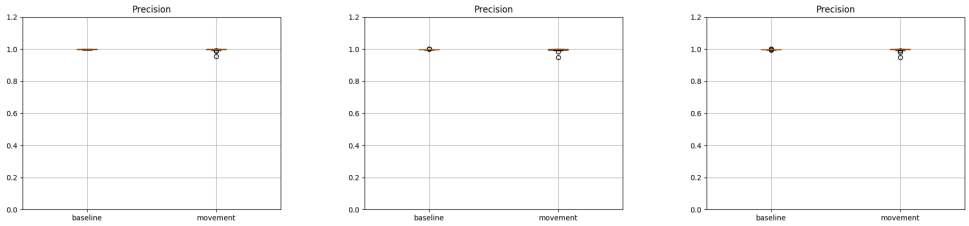
(a) Kurtosis (b) Spectral Power Ratio

Figure 3: SQIs of the ECG signals, a) Kurtosis, b) SPR. The red lines indicate the threshold values for good quality signals: the Kurtosis (a) is expected to be greater than 5; the SPR (b) is expected to have values between 0.5 and 0.8. FC: Flex Comp, HB: Heart Band



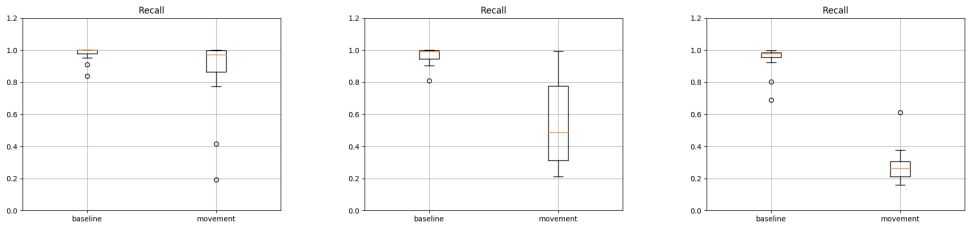
(a) Kurtosis (b) Spectral Power Ratio

Figure 4: SQIs of the BVP signals: a) Kurtosis, b) SPR. The red lines indicate the threshold values for good quality signals: the Kurtosis (a) is expected to be less than 3.5; the SPR (b) is expected to have values between 0.5 and 0.8. FC: Flex Comp, E4: Empatica E4



(a) ECG from HeartBand (b) BVP from FlexComp (c) BVP from E4

Figure 5: Precision of the heart beat detection showing good performances for all the signals and sessions.

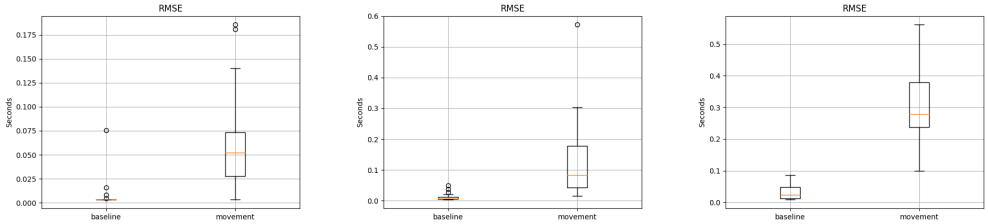


(a) ECG from HeartBand (b) BVP from FlexComp (c) BVP from E4

Figure 6: Recall of the heart beat detection: the body movements affect the performance. The effects are more evident for the BVP signals from the FlexComp (b) and the E4 (c).

movements, as already reported in literature [34].

During the baseline sessions all devices show good performances in terms of P and R, although the recall of the IBIs series derived from the BVP signals is inferior (see Figures 5, 6 and 7). The e_{RMSE} is also sensibly higher for the IBIs series derived from the BVP signals. This result is expected as the beat detection in BVP signals is in general more difficult due to posture and individual physiological characteristics [26]. As expected, all metrics worsen during the movement sessions, in particular the IBIs series derived from BVP signals of the Empatica E4 show very low recall (below 0.4) and high e_{RMSE}.



(a) ECG from HeartBand (b) BVP from FlexComp (c) BVP from E4

Figure 7: RMSE of the heart beat detection. The body movements affect the performance.

Table 3: Ratio of Out-of-Range values for each signal and subject.

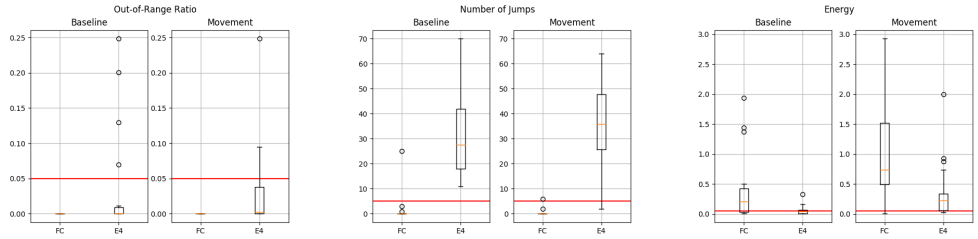
	Baseline		Movement	
Subject	FC	E4	FC	E4
1	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.03
4	0.00	**0.07**	0.00	0.00
5	0.00	0.00	0.00	0.04
6	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.06
8	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.02
10	0.00	**0.25**	0.00	0.00
11	0.00	**0.13**	0.00	0.00
12	0.00	**0.20**	0.00	0.00
13	0.00	0.01	0.00	**0.25**
14	0.00	0.00	0.00	0.02
15	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	**0.10**
17	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	**0.07**

Table 4: Ratio of Out-of-Range values for each signal and subject.

	Baseline		Movement	
Subject	FC	E4	FC	E4
1	3	**42**	0	**32**
2	0	**70**	0	**15**
3	0	**11**	0	**36**
4	0	**18**	0	**63**
5	0	**51**	0	**57**
6	0	**18**	0	**28**
7	0	**23**	0	**23**
8	0	**13**	0	**30**
9	0	**24**	0	**37**
10	1	**31**	2	**63**
11	0	**18**	0	2
12	0	**11**	0	**38**
13	0	**14**	0	**20**
14	0	**31**	0	**38**
15	0	**51**	0	**51**
16	0	**42**	0	**25**
17	0	**42**	0	**64**
18	**25**	**60**	**6**	**36**

Table 5: Signal Activity values for each signal and subject.

	Baseline		Movement	
Subject	FC	E4	FC	E4
1	**0.03**	0.07	0.62	0.22
2	0.46	**0.00**	1.37	**0.02**
3	0.05	0.06	1.37	0.87
4	1.44	0.07	2.92	0.06
5	**0.04**	**0.01**	0.16	0.06
6	0.22	**0.01**	1.66	0.07
7	1.37	0.07	2.68	2.00
8	0.18	0.07	1.78	**0.03**
9	1.93	**0.01**	1.52	0.74
10	**0.01**	**0.02**	0.54	**0.03**
11	0.31	0.33	0.45	**0.03**
12	0.30	0.16	1.52	0.22
13	**0.01**	**0.02**	0.19	0.27
14	0.50	0.12	0.86	0.36
15	0.05	**0.02**	0.58	0.23
16	**0.02**	0.06	0.47	0.93
17	0.29	**0.01**	0.58	0.08
18	**0.00**	**0.01**	**0.01**	0.28



(a) Out-of-Range Ratio (b) Number of Jumps (c) Signal Activity

Figure 8: SQIs of the EDA signals, a) Ratio of Out-of-Range samples, b) Number of Jumps and c) Signal Activity. The red lines indicate the threshold values for good quality signals: the Ratio of Out-of-Range samples (a) is expected to be less than 0.05; the Number of Jumps (b) is expected to be less than 5 and the Signal Activity (c) is expected to be greater than 0.05. FC: Flex Comp, E4: Empatica E4

3.2 Electrodermal activity

None of the EDA signals collected with the Flex Comp presented Out-of-Range samples while few signals collected by the Empatica E4 appear corrupted, with a worsening effect due to the body movements (Figure 8a).

In terms of number of jumps, the Flex Comp unit provided good quality signals for both the baseline and the movement sessions, with few exceptions. Instead, the signals collected with the Empatica E4 presented a low quality with an average $n_j = 31.6$ for the baseline and $n_j = 36.7$: more than one jump every 10 s.

The Signal Activity index of the signals collected with the Empatica E4 is low, despite the high number of jumps, in particular for the baseline session.

4 Discussion

We adopted a set of SQIs derived from the literature to assess the overall quality of the cardiac signals and EDA signals provided in the WCS dataset.

In general, the signals collected with the clinical device showed an optimal signal quality on the baseline session, indicating that the experimental design and settings were appropriate for the acquisition of scientific-level data. During the movement session, the induced artifacts slightly worsen the signal quality, in particular for the BVP signal, which is known to be more sensi-

tive to body movements. During the baseline, the wearable devices collect cardiac signals with quality comparable to the signals from the FlexComp unit but are more affected by the body movements as shown by the worse signal quality of the movement session.

The signal quality of the EDA signals collected by the E4 is not appropriate for scientific research, in particular due to the high number of jumps. The lower quality of signals collected from WDs can be attributed to a number of technical limitations. For instance, the performances of some sensor components are reduced to decrease battery consumption and increase autonomy. In particular, the sampling frequency is kept lower than medical-grade devices, usually below 256 Hz with consequent issues in the accuracy of the physiological indicators computed. To favour wearability, no aids are used to fix the sensors to the body (e.g. adhesive ECG electrodes). This allow for minimal shifts of the sensors that can introduce noise on the acquired signal [24, 36, 34]. Finally, although some anatomical loci are more appropriate than others for measuring a physiological signal, the need of embedding the sensor in a wearable and comfortable support might constrain the positioning of the sensor into a non-optimal locus where the signal magnitude might be lower, such as the wrist instead of hand palm in the case of EDA [31].

This study evidenced the need of testing and validating the use of WD in experimental settings for research and the importance of both technological and signal processing aspects to obtain reliable signals. The validation should reflect the experimental context in which the data are collected; in particular, since the WD are employed to favor the freedom of movement of the subject, the influence of artifacts due to movements should be carefully considered. The use of proper SQI represents a quantifiable and reproducible method to identify the signals with good quality.

The collected data and the developed pipeline are offered to the scientific community to encourage the development and validation of new SQI and algorithms. The adoption of standardized metrics and procedures should favor the implementation of experiments with multivariate physiological signals, based on WDs.

As a final note, it is important to remark that the results of this study are not to be interpreted in terms of technological validation of the WDs used for the experiment. The technological development is still in progress and manufactures of WDs are constantly improving the sensing and acquisition technologies. Therefore results reflect the state of the technology at the time of the experiments. Nevertheless, the validation and quantification of the

quality of the collected data is a critical step that should be performed each time the technology or the WDs are updated.

Author contribution

Conceptualization, A.B., C.F. and G.E.; methodology, A.B.; software, A.B. and G.G.; resources, G.E.; data curation, G.G.; writing–original draft preparation, A.B.; writing–review and editing, A.B. and C.F. and G.E.

Conflicts of interest

The authors declare no conflict of interest.

Funding

A.B. was supported by a Post-doctoral Fellowship within MIUR programme framework "Dipartimenti di Eccellenza" (DiPSCO, University of Trento). G.E. was supported by NAP SUG 2015, Singapore Ministry of Education ACR Tier 1 (RG149/16 and RT10/19).

References

[1] David M Alexander, Chris Trengove, P Johnston, Tim Cooper, JP August, and Evian Gordon. Separating individual skin conductance responses in a short interstimulus-interval paradigm. *Journal of neuroscience methods*, 146(1):116–123, 2005.

[2] Atiqah Azhari, Mengyu Lim, Andrea Bizzego, Giulio Gabrieli, Marc H Bornstein, and Gianluca Esposito. Physical presence of spouse enhances brain-to-brain synchrony in co-parenting couples. *Scientific reports*, 10(1):1–11, 2020.

[3] Amay J Bandodkar, Denise Molinnus, Omar Mirza, Tomás Guinovart, Joshua R Windmiller, Gabriela Valdés-Ramírez, Francisco J Andrade, Michael J Schöning, and Joseph Wang. Epidermal tattoo potentiometric

sodium sensors with wireless signal transduction for continuous non-invasive sweat monitoring. *Biosensors and bioelectronics*, 54:603–609, 2014.

[4] David R Bassett. Device-based monitoring in physical activity and public health research. *Physiological measurement*, 33(11):1769, 2012.

[5] Joachim Behar, Julien Oster, Qiao Li, and Gari D Clifford. Ecg signal quality during arrhythmia and its application to false alarm reduction. *IEEE transactions on biomedical engineering*, 60(6):1660–1666, 2013.

[6] Andrea Bizzego, Atiqah Azhari, Nicola Campostrini, Anna Truzzi, Li Ying Ng, Giulio Gabrieli, Marc H Bornstein, Peipei Setoh, and Gianluca Esposito. Strangers, friends, and lovers show different physiological synchrony in different emotional states. *Behavioral Sciences*, 10(1):11, 2020.

[7] Andrea Bizzego, Alessandro Battisti, Giulio Gabrieli, Gianluca Esposito, and Cesare Furlanello. pyphysio: A physiological signal processing library for data science approaches in physiology. *SoftwareX*, 10:100287, 2019.

[8] Andrea Bizzego and Cesare Furlanello. DBD-RCO: Derivative based detection and reverse combinatorial optimization to improve heart beat detection for wearable devices. *bioRxiv*, page 118943, 2017.

[9] Marie Chan, Daniel Estève, Jean-Yves Fourniols, Christophe Escriba, and Eric Campo. Smart wearable systems: Current status and future challenges. *Artificial intelligence in medicine*, 56(3):137–156, 2012.

[10] Lisa Y Chen, Benjamin C-K Tee, Alex L Chortos, Gregor Schwartz, Victor Tse, Darren J Lipomi, H-S Philip Wong, Michael V McConnell, and Zhenan Bao. Continuous wireless pressure monitoring and mapping with ultra-small passive sensors for health monitoring and critical care. *Nature communications*, 5, 2014.

[11] Toan Dinh, Thanh Nguyen, Hoang-Phuong Phan, Nam-Trung Nguyen, Dzung Viet Dao, and John Bell. Stretchable respiration sensors: Advanced designs and multifunctional platforms for wearable physiological monitoring. *Biosensors and Bioelectronics*, page 112460, 2020.

[12] Mohamed Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4):21, 2016.

[13] Richard R Fletcher, Ming-Zher Poh, and Hoda Eydgahi. Wearable sensors: opportunities and challenges for low-cost health care. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 1763–1766. IEEE, 2010.

[14] M. Garbarino, M. Lai, D. Bender, Rosalind W. Picard, and S. Tognetti. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 EAI 4th International Conference on Wireless Mobile Communication and Healthcare*, pages 39–42, 2014.

[15] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[16] Cheng-Yang Huang, Ming-Che Chan, Chien-Yue Chen, and Bor-Shyh Lin. Novel wearable and wireless ring-type pulse oximeter with multi-detectors. *Sensors*, 14(9):17586–17599, 2014.

[17] Ian R Kleckner, Rebecca M Jones, Oliver Wilder-Smith, Jolie B Wormwood, Murat Akcakaya, Karen S Quigley, Catherine Lord, and Matthew S Goodwin. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. *IEEE Transactions on Biomedical Engineering*, 2017.

[18] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[19] Sylvia D Kreibig. Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421, 2010.

[20] Robert W Levenson. The autonomic nervous system and emotion. *Emotion Review*, 6(2):100–112, 2014.

[21] Qiao Li, Roger G Mark, and Gari D Clifford. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. *Physiological measurement*, 29(1):15, 2007.

[22] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[23] Alexandros Pantelopoulos and Nikolaos G Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1):1–12, 2010.

[24] Michael T Petterson, Valerie L Begnoche, and John M Graybeal. The effect of motion on pulse oximetry and its clinical significance. *Anesthesia & Analgesia*, 105(6):S78–S84, 2007.

[25] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[26] Axel Schäfer and Jan Vagedes. How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram. *International journal of cardiology*, 166(1):15–29, 2013.

[27] Nandakumar Selvaraj, Yitzhak Mendelson, Kirk H Shelley, David G Silverman, and Ki H Chon. Statistical approach for the detection of motion/noise artifacts in photoplethysmogram. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 4972–4975. IEEE, 2011.

[28] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[29] Melanie Swan. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks*, 1(3):217–253, 2012.

[30] Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. Automatic identification of artifacts in electrodermal activity data. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 1934–1937. IEEE, 2015.

[31] Marieke van Dooren, J. J. G. de Vries, and Joris H Janssen. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & behavior*, 106(2):298–304, 2012.

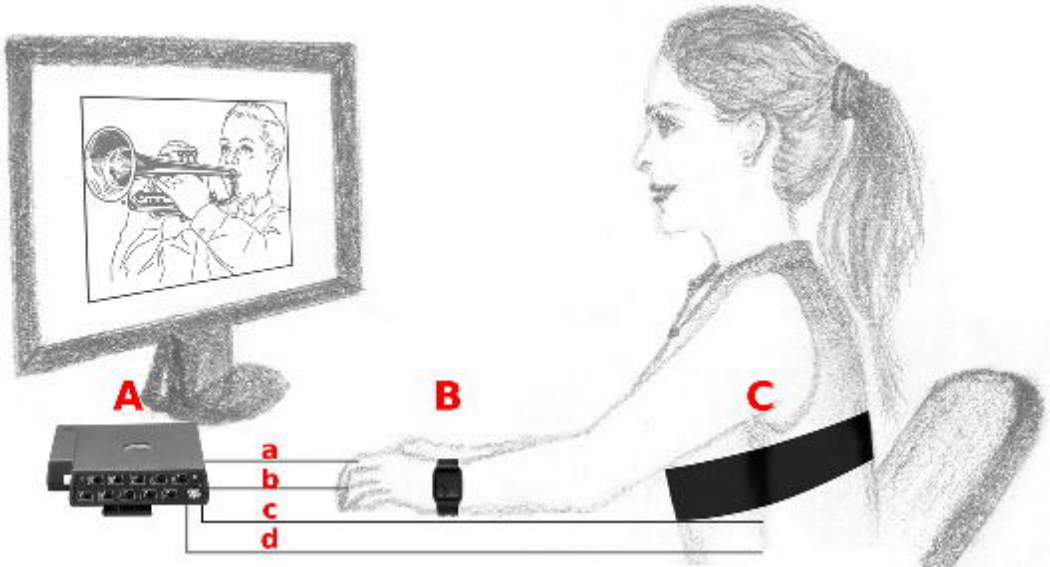
[32] Angelina Vernetti, Frederick Shic, Laura Boccanfuso, Suzanne Macari, Finola Kane-Grade, Anna Milgramm, Emily Hilton, Perrine Heymann, Matthew S Goodwin, and Katarzyna Chawarska. Atypical emotional electrodermal activity in toddlers with autism spectrum disorder. *Autism Research*, 2020.

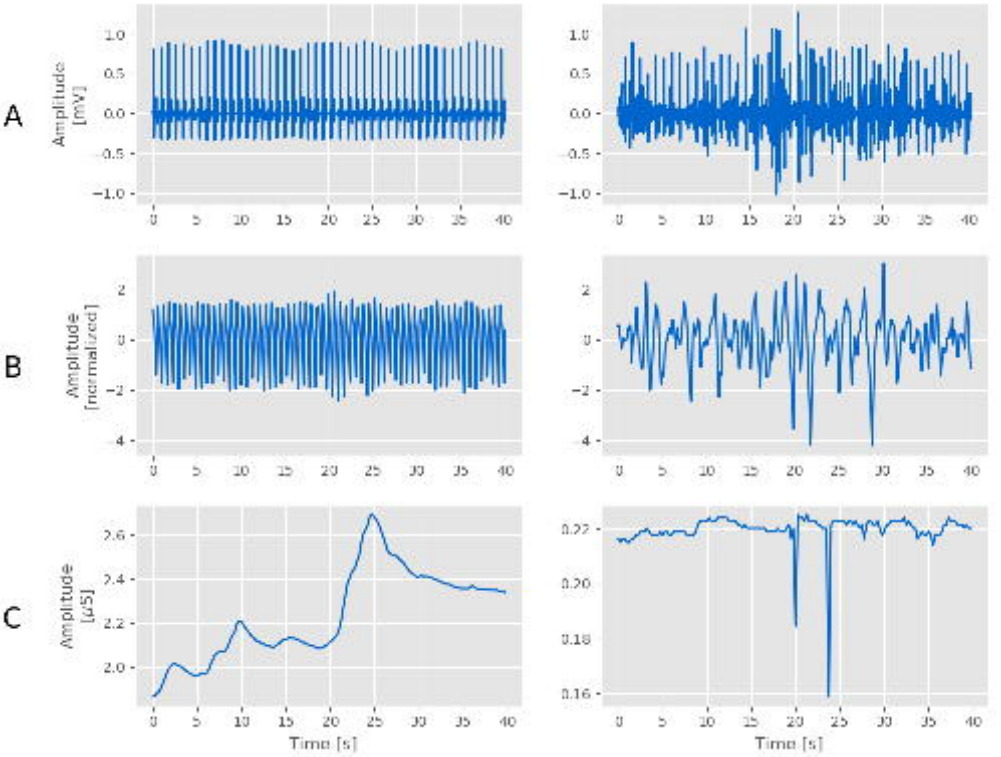
[33] JD Wander and D Morris. A combined segmenting and non-segmenting approach to signal quality estimation for ambulatory photoplethysmography. *Physiological measurement*, 35(12):2543, 2014.

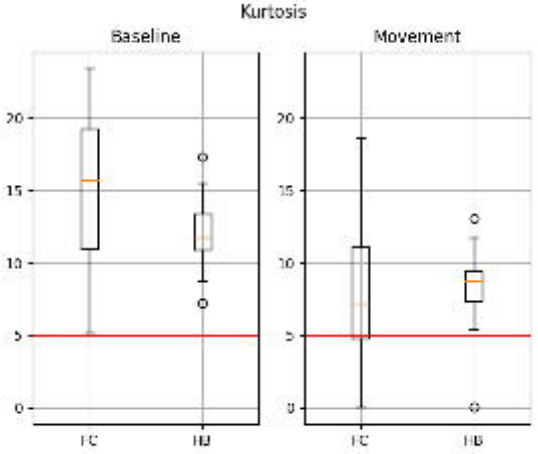
[34] Kristen M Warren, Joshua R Harvey, Ki H Chon, and Yitzhak Mendelson. Improving pulse rate measurements during random motion using a wearable multichannel reflectance photoplethysmograph. *Sensors*, 16(3):342, 2016.

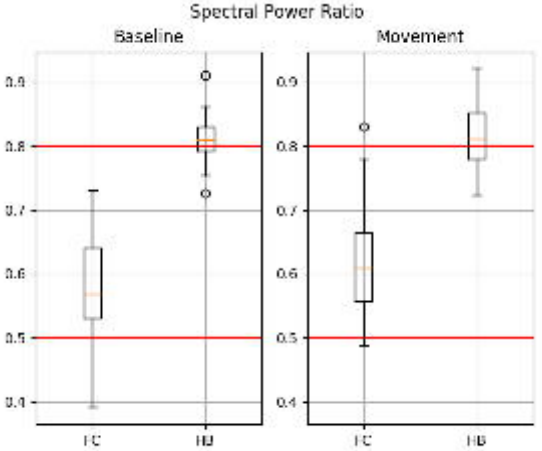
[35] Victoria Xia, Natasha Jaques, Sara Taylor, Szymon Fedor, and Rosalind Picard. Active learning for electrodermal activity classification. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2015 IEEE*, pages 1–6. IEEE, 2015.

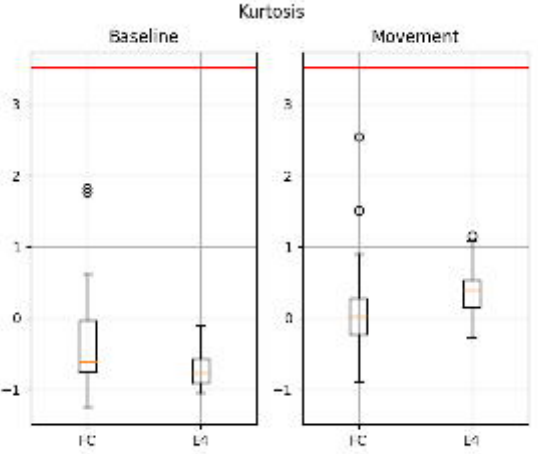
[36] SR Yadhuraj and H Harsha. Motion artifact reduction in photoplethysmographic signals: A review. *International Journal of Innovative Research and Development*, 2(3):626–640, 2013.

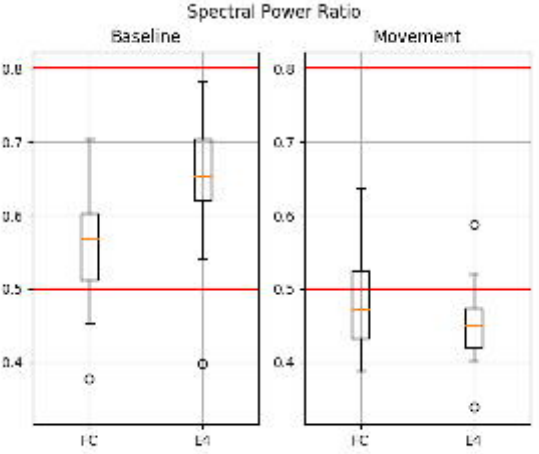

A
B
C
a
b
c
d

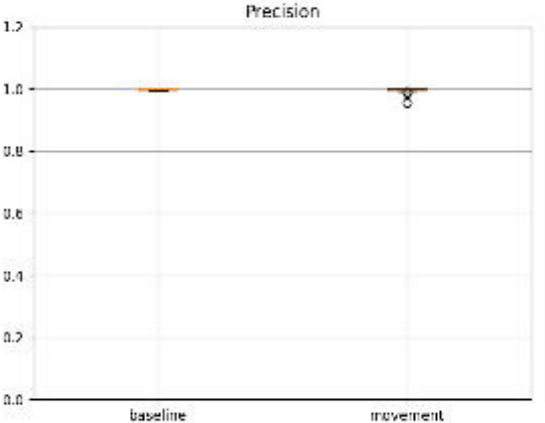

A
B
C
Amplitude [mV]
Amplitude [normalized]
Amplitude [µS]
Time [s]
Time [s]

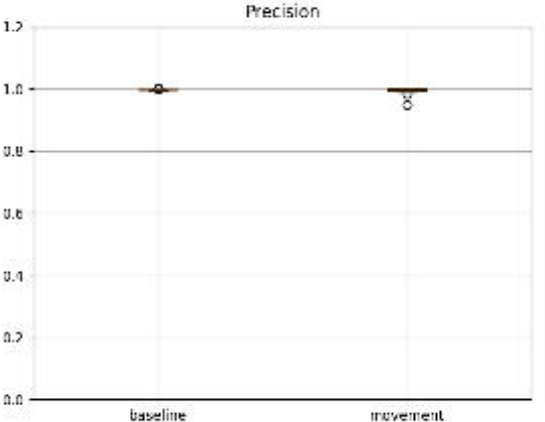

Kurtosis
Baseline
Movement
20
15
10
5
0
FC
HB

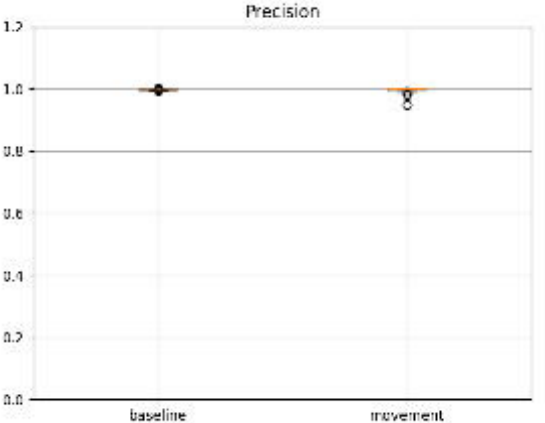

Spectral Power Ratio
Baseline
Movement
0.9
0.8
0.7
0.6
0.5
0.4
FC
HB

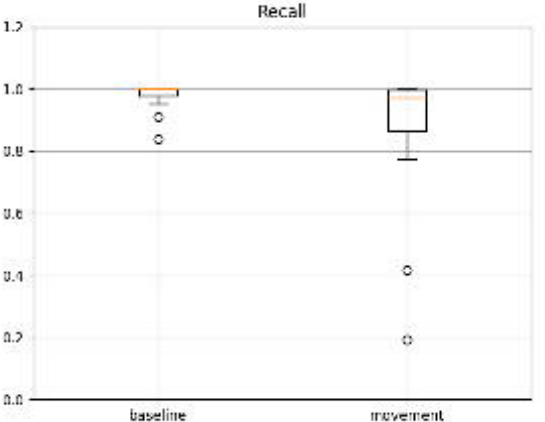

Kurtosis
Baseline
Movement
3
2
1
0
−1
FC
LM

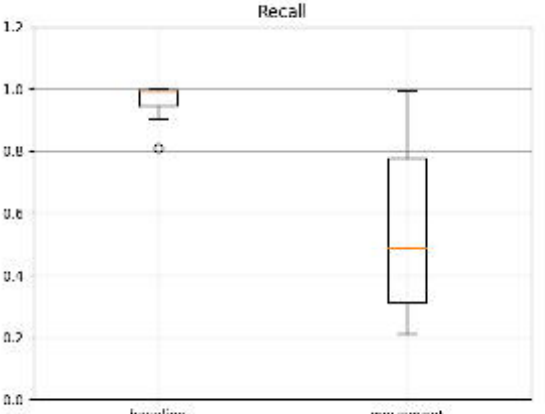

Spectral Power Ratio
Baseline
Movement
0.8
0.7
0.6
0.5
0.4
FC
LM

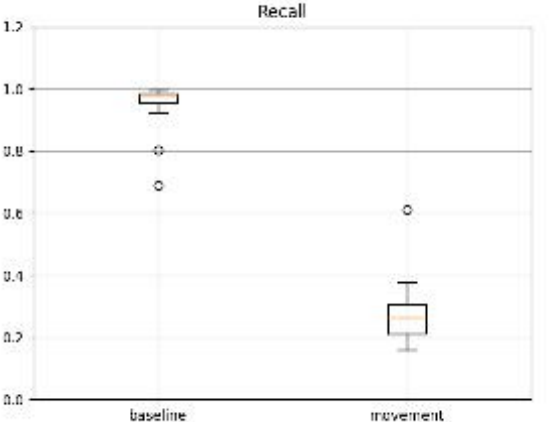

Precision
1.2
1.0
0.8
0.6
0.4
0.2
0.0
baseline
movement

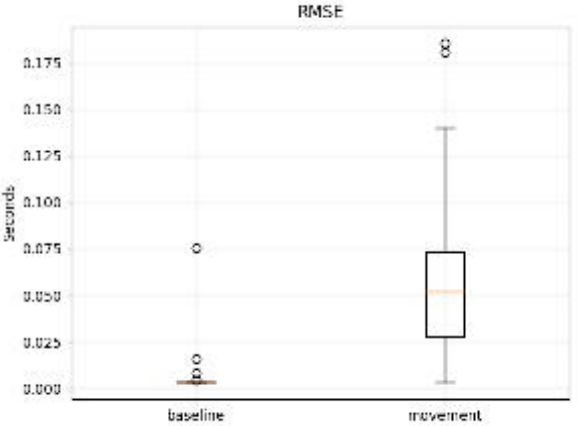

Precision
1.2
1.0
0.8
0.6
0.4
0.2
0.0
baseline
movement

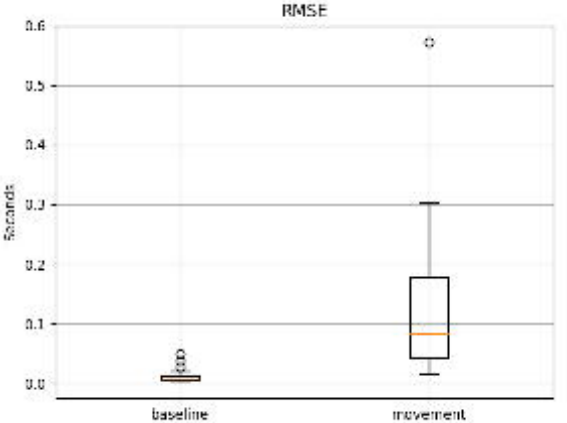

Precision
1.2
1.0
0.8
0.6
0.4
0.2
0.0
baseline
movement

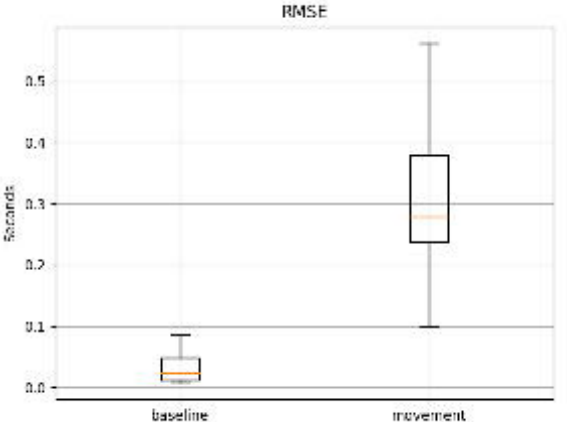

Recall
1.2
1.0
0.8
0.6
0.4
0.2
0.0
baseline
movement

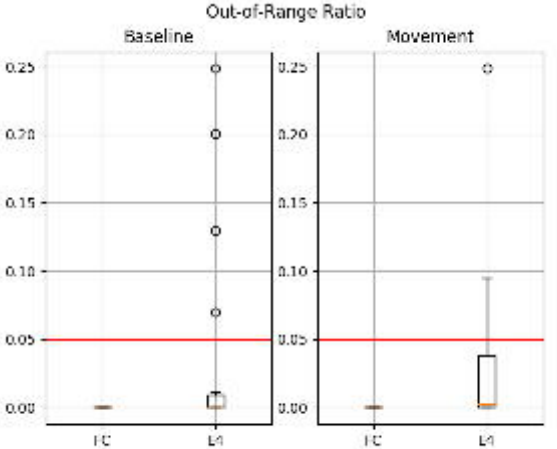

Recall
1.2
1.0
0.8
0.6
0.4
0.2
0.0
baseline
movement

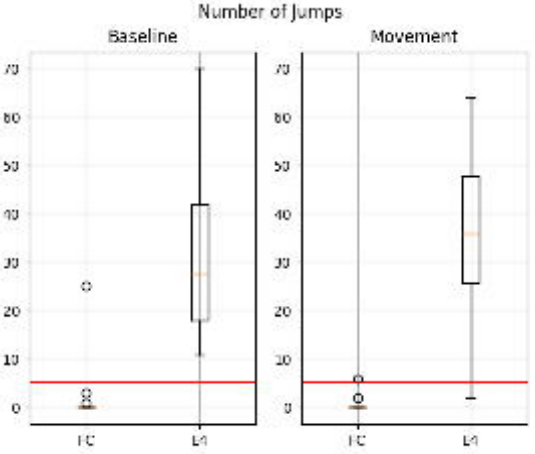

Recall
1.2
1.0
0.8
0.6
0.4
0.2
0.0
baseline
movement

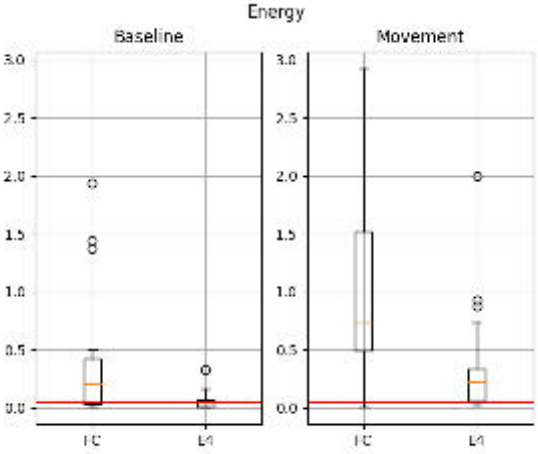

RMSE
seconds
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
baseline
movement


RMSE
seconds
0.6
0.5
0.4
0.3
0.2
0.1
0.0
baseline
movement


RMSE
seconds
0.0
0.1
0.2
0.3
0.4
0.5
baseline
movement


Out-of-Range Ratio
Baseline
Movement
0.25
0.20
0.15
0.10
0.05
0.00
FC
L4


Number of Jumps
Baseline
Movement
70
60
50
40
30
20
10
0
FC
L4


Energy
Baseline
Movement
3.0
2.5
2.0
1.5
1.0
0.5
0.0
FC
L4