Article

Personalized PPG Normalization based on Subject Heartbeat in Resting State Condition

Francesca Gasparini 1*, Alessandra Grossi 1, Marta Giltri 1 and Stefania Bandini 2, 1

1 Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy; francesca.gasparini@unimib.it (F.G.), a.grossi6@campus.unimib.it (A.G.), m.giltri@campus.unimib.it (M.G.), stefania.bandini@unimib.it (S.B.)
2 RCAST - Research Center for Advanced Science & Technology, The University of Tokyo, Tokyo, Japan
* Correspondence: francesca.gasparini@unimib.it

Abstract: Physiological responses are nowadays widely used to recognize the affective state of subjects in real-life scenarios. However, these data are intrinsically subject-dependent, making machine learning techniques for data classification not easily applicable due to inter-subject variability. In this work, the reduction of inter-subject heterogeneity is considered in the case of PhotoPlethysmoGraphy (PPG), which is successfully used to detect stress and evaluate experienced cognitive load. To face the inter-subject heterogeneity, a novel personalized PPG normalization is here proposed. A subject-normalized discrete domain where the PPG signals are properly re-scaled is introduced, considering the subject’s heartbeat frequency in resting state conditions. The effectiveness of the proposed normalization is evaluated in comparison with other normalization procedures in a binary classification task, where cognitive load and relaxing state are considered. The results obtained on two different datasets available in the literature confirm that applying the proposed normalization strategy permits to increase classification performance.

Keywords: wearable sensors, PPG, bio-signal processing, normalization, physiological signals

1. Introduction

In recent years, sensor technology has improved significantly and wearable devices have become increasingly popular, allowing to easily register subjects’ physiological responses during their daily activities [1,2]. Physiological signals are successfully used to measure arousal. Arousal is a human uncontrolled reaction, related to attention and cognitive alertness, activated by stimuli that require high psycho-physical engagement, and thus activated in particular during cognitive tasks and stressful conditions. Although it is proved that sometimes stress can have a positive effect on a person by improving his/her alertness state or his/her ability to react [3], it is also proved that a high and continuous level of stress or cognitive load can affect the physical and mental subject’s well-being. Illnesses like depression, anxiety and sleep disorders are, indeed, often due to excessive stress or workload [4].

In view of its importance, the automatic recognition of stress and excessive cognitive load has recently become an object of studies and researches, even in different application areas. For instance, systems able to recognize emotion and, above all, stress experienced by subjects can be used in working or academic environment [5] in order to monitor and identify the emotional state of employees or students. In this regard, it is indeed proved that a high level of stress or cognitive load due to excessive workload can increase the level of fatigue, decrease the subject’s working capability and, consequently, bring physical and mental illness that can lead to absence from the workplace [6]. Similarly, automatic stress recognition systems can be used in the context of vehicle driving for the detection of excessive mental fatigue states that can reduce a person’s driving skills [7]. Finally, algorithms of stress detection can be also used in
recreational areas for the development of systems able to modify their parameters based on the user’s emotional state. Concrete examples concern music-retrieval systems able to interact with the user of a music playlist using both external inputs and his physiological signals [8] or videogames in which some internal game parameters, like the difficulty, are set on the base of player’s emotions and stressing level [9]. Moreover, systems able to recognize the subject’s emotional state can also be involved in the medical area. For instance, emotion recognition systems can be used to monitor the health state of convalescent patients [3] or to help elderly subjects during their daily activities [10,11]. In all these contexts, the development of systems able to recognize, interpret and simulate human affect can be seen as a necessary step to make technologies user-friendly and able to interact actively with people.

Despite the progress in sensor technology and the relative simplicity of acquiring physiological signals from the human body, there are still some critical issues that must be addressed to fully exploit the potential that the analysis of physiological signals can offer. Although it is always easier to acquire data, which can overcome the low cardinality of datasets, the application of machine learning techniques is still limited by inter-subject heterogeneity. Even in the same resting condition, without external stimuli, physiological signals appear significantly subject dependent.

In this work, the reduction of inter-subject heterogeneity is faced in the case of one of the most used physiological signal acquired by wearable devices: the heartbeat, mainly detected through PhotoPlethysmoGraphy (PPG) [12]. The PPG signal is one of the most used signals to measure arousal [13,14], and consequently to detect stress and evaluate the experienced cognitive load [15,16]. The PPG signal of each subject appears different, both in terms of amplitude and beat frequency. Regarding the amplitude, differences can be due to the subjects’ skin characteristics or to different sensor adherence during the acquisition phase. Concerning the diversity in terms of heartbeat, according to the American Heart Association the heart rate frequency of a resting adult can vary in a range between 60 and 100 beats per minute, and it depends on many different factors both personal (like age, sex, ethnicity, sports ability, diet, illnesses, prescribed medications, etc.) and environmental (humidity, temperature, etc.) [17]. Normalization procedures based on data rescaling are often used [18] to overcome amplitude variability within subjects. The main strategies adopted involve rescaling to the range \([0, 1]\) [19], normalizing by dividing by the maximum value of the signal [20] and applying Z-score [21]. All these methods, however, do not take into account the effective differences in the subjects’ heartbeat, which are not only related to amplitude, but also to frequency.

The aim of this work is to solve this inter-subject variability, adopting a novel personalized PPG normalization based on the heartbeat of the subject in a resting state condition. To validate the normalization procedure here presented, PPG data belonging to two different datasets are considered: the CLAWDAS (Cognitive Load and Affective Walkability in Different Age Subjects) dataset, partially introduced in [22] and in [23] and the CLAS (Cognitive Load, Affect and Stress recognition) dataset, available in the literature [24].

The proposed normalization is analyzed considering a binary classification task to discriminate cognitive load from relaxing state and compared with normalization strategies adopted in the literature.

The paper is organized as follows. In Section 2, the two considered datasets are described, while the preliminary signal preprocessing strategies applied to each of them are reported in Section 3. The novel personalized PPG normalization strategy, based on subject resting state heartbeat is presented in Section 4. Extracted features, classification strategies and the adopted cross validation approach are reported in Section 5. The comparison of classification performances on the two datasets, obtained with different normalization strategies, is then reported and analyzed in Section 6. Finally, conclusions are drawn in the last Section.
Figure 1. Example of signal acquisition in CLAWDAS dataset: Left: Math Calculation, Right: Relaxing Audio Listening

Figure 2. Sensors used to collect physiological data in the CLAWDAS dataset

2. Dataset description

CLAS [24] and CLAWDAS [22,23] datasets are here considered to validate the proposed normalization procedure.

In the CLAS dataset, the physiological signals of 60 healthy volunteers (mostly students between 20 and 27 years old, 17 women) were acquired while they were performing interactive or perceptive tasks. In particular, the interactive tasks were introduced to evaluate the level of concentration and the cognitive capacity of different individuals by solving Math Problems, Logic Problems and Stroop Tests. In the perceptive tasks, different emotions have been elicited in the participants by images and video selected from the DEAP dataset [25]. During the whole experiment, three types of physiological signals were simultaneously recorded by means of Shimmer sensors (www.shimmersensing.com): Electrocardiography (ECG), Plethysmography (PPG) and ElectroDermal Activity (EDA). The signals were acquired with a sampling rate of 256 Hz and a resolution of 16-bits per sample. In addiction, for each subject, 3D accelerometer data and metadata were also collected.

For the purpose of this work, only PPG signals collected during the interactive tasks are considered. For each participant, this phase of the experiment is characterized by the following steps:

- **1 minutes of Baseline** in resting state condition
- **3 minutes of Math problems** in which the participant solves different simple mathematical problems (MP) in a limited interval of time.
- **30 seconds of Neutral State** (NS), in which neutral audio-visual stimuli are displayed
- **3 minutes of Stroop tests** (ST) where the user is expected to match correctly the color of the text with the meaning of the word, having a strict time constraint for each assignment.
- **30 seconds of Neutral State** (NS), in which neutral audio-visual stimuli are displayed
- **5 minutes of Logic problems** (LP) consisting of several simple logical problems often used during the IQ tests.
- **30 seconds of Neutral State** (NS), in which neutral audio-visual stimuli are displayed

Moreover the 3 Neutral State repetitions of 30 seconds for each subject, in the Picture test of the perceptive session are also considered. For further details of this dataset, please refer to [24].

CLAWDAS is a dataset collected in a controlled laboratory environment at the Research Center for Advanced Science and Technology (RCAST) at The University of Tokyo. The experiments performed at RCAST are focused on finding differences in physiological responses...
Table 1. Number of instances for each task in the CLAS dataset (first 5 columns) and in the CLAWDAS dataset (last 5 columns), distinguished in CLAWDAS Young and CLAWDAS Elderly. BL = Baseline, MP = Math Problems, ST = Stroop Test, LP = Logic Problems, NS = Neutral State, MC = Math Calculation, R = Reading, C = Comprehension.

|        | Num Subj. | BL | MP | ST | LP | NS | BL | MC | AL | R  | C  |
|--------|-----------|----|----|----|----|----|----|----|----|----|----|
| CLAS   | 60        | 60 | 60 | 60 | 60 | 360| -  | -  | -  | -  | -  |
| CLAWDAS Young | 16        | -  | -  | -  | -  | -  | 46 | 96 | 96 | 32 | 32 |
| CLAWDAS Elderly | 20        | -  | -  | -  | -  | -  | 60 | 120| 120| 40 | 40 |

related to different ages of the subjects, involved in several tasks, from cognitive, to listening and walking ones. The experiments involved two different groups of subjects: a population of 16 Japanese young adults with average age = 24.7 years old (4 women) and a population of 20 Japanese elderly people with average age = 65.15 years old (10 women). During the whole experiment, the heartbeat of each participant has been collected though Photoplethysmography (PPG) using the Shimmer3 GSR+ Unit (www.shimmersensing.com) with a sampling frequency of 128 Hz. In addition to PPG, the ElettroDermal Activity (EDA) of each subject has been acquired using the same sensor. The Shimmer3 GSR+ Units are non-invasive and completely painless sensors that could be easily worn by the participants as shown in Figure 2.

CLAWDAS data acquired during cognitive and listening tasks are here considered. The experimental protocol is composed of the following steps:

- **3 minutes** of questionnaires to collect the personal details and the current emotional state of each participant using STAI Questionaries.
- **1 minute** of Baseline (BL) in resting state condition.
- **6 minutes** of Reading (R) and Comprehension (C) tasks composed by two repetitions (trials) of 2 minutes of R followed by 1 minute of self assessment and C questions.
- **1 minute** of Baseline (BL) in resting state condition.
- A **15 minutes** sequence composed of six repetitions of the following two tasks:
  1. **2 minutes** of Audio Listening (AL). In this task, relaxation is induced by natural sounds (Figure 1 right).
  2. **30 seconds** of cognitive load, induced by mental Math Calculations (MC) that involve sums, subtractions and multiplications (Figure 1 left).

Each repetition has a different audio track and math calculation.

- **1 minute** of Baseline (BL) in resting state condition

The experimental protocol has been reviewed and approved by the Research Ethics Committee at The University of Tokyo, Japan (No. 19-283 and 19-376). The CLAWDAS dataset is partitioned into two distinct subsets, according to the age of the participants: CLAWDAS Young, that includes all the signals acquired from young adults and the CLAWDAS Elderly, that groups all the signals collected from the elderly. In the following analysis the two groups are considered separately.

Table 1 reports the number of instances for each task in the CLAS and CLAWDAS datasets, keeping distinct the two subsets of CLAWDAS related to subjects’ age (Young and Elderly).
3. Signal preprocessing: denosing and amplitude normalization

In this section the preprocessing operations applied to raw PPG data are detailed and differentiated for CLAS and CLAWDAS respectively.

3.1. Denoising strategies

The raw PPG signals are usually corrupted by noise and motion artifacts that can undermine their interpretation and use [26]. In the CLAS dataset, the signals have been already preprocessed by the authors during the acquisition phase ([24]) and thus no further denosing procedure is applied.

Concerning the CLAWDAS, the PPG raw signals of each subject are preprocessed by a multiresolution wavelet denoising strategy, as suggested by [26,27]. The signal is divided in frequency sub-bands using Stationary Wavelet Transform (SWT) [28] with mother wavelet Fejer-Korovkin [29], and four levels of decomposition. A soft thresholding is applied to the detail coefficients of each sub-band. The universal threshold calculated by the formula $T_k = \sqrt{2\log(N_j)}$ is adopted, where $N_j$ is the length of the $j-th$ wavelet coefficient, and $k$ is the sub-band, [30]. The SWT is implemented with the algorithm a-trous [31]. A preliminary operation of replicate padding is applied to the signal in order to obtain a length divisible by $2^\text{level}$ [28], with $\text{level} = 4$.

3.2. Amplitude normalization

In order to normalize the signals with respect to the amplitude, a Z-score operation, defined by the formula $Z = (x - \mu) / \sigma$, is applied on the PPG recordings after the denoising procedure.

In CLAWDAS the amplitude normalization, ($\text{AmpN}$), as well as the denoising preprocessing, is applied to the signal of each subject, before splitting it into the different experimental trials.

A similar procedure is also applied to the CLAS dataset signals. However, in this case, the authors have already split the data into single trials, according to their experimental protocol (see Section 2), with no preliminary amplitude normalization. Thus, in order to apply a similar procedure to both the datasets, the segmented trials of each subject are concatenated to re-build the original acquired signal. Then, the Z-score amplitude normalization is applied to each subject signal. Finally, the amplitude normalized signals are split back into the trials, related to single tasks, using the markers properly defined during the previous phase of concatenation.

4. Personalized PPG normalization based on subject resting state heartbeat

The American Heart Association has underlined that the heart rate frequency of an adult in a resting state can vary in a range between 60 and 100 beats per minute. This inter-subject variability depends on many different factors both personal and environmental. In case of CLAS and CLAWDAS datasets, the subjects’ heartbeat range of the baseline recordings belongs to the one reported by the literature, as depicted in Figure 3, where the average heartbeat distribution for each dataset is plotted.

In order to get rid of this inter-subject variability, the core idea of our normalization procedure is to map PPG signals, defined in the Discrete Time Domain (DTD), into a new Subject Normalized discrete Domain (SND), applying a mapping procedure based on the resting state heartbeat frequency. In this SND, all the subjects have the same resting state heart frequency. For each subject, a personal resampling frequency can be calculated, so that for all the subjects the heart frequency of the resting state in the SND is equal, despite the original subject-peculiar frequency in the Continuous Time Domain (CTD).
Figure 3. Distributions of the average heartbeat of the resting state PPG signals in CLAS (left), CLAWDAS Young (middle) and CLAWDAS Elderly (right) datasets.

Table 2. Correspondences between the three domains

| Domain              | Continuous Time | Discrete Time | Subject Normalized |
|---------------------|-----------------|---------------|--------------------|
| Acronym             | CTD             | DTD           | SND                |
| Heartbeat           | $f$ [beat second] | $f_N$ [beat sample] | $f_{SN}$ [beat SNsample] |
| Resting state Heartbeat | $f_b$ [beat second] | $f_{Nb}$ [beat sample] | $f_{SNb}$ [beat SNsample] |
| Sampling frequency  | –               | $f_c$ [sample second] | $f_{SNc}$ [SNsample sample] |

Then, the PPG data acquired during all the experimental tasks are also mapped into this new domain, applying the calculated personal resampling frequency and obtaining subject normalized PPG signals that can be considered for population-based analysis.

Defining as $f_c$ [sample second] the sampling frequency of the PPG signal, given by the acquisition device, and $f_b$ [beat second] the heartbeat frequency in the resting state condition in CTD, the corresponding normalized heartbeat frequency $f_{Nb}$ [beat sample] in the DTD is:

$$f_{Nb} = \frac{f_b \text{ [beat second]}}{f_c \text{ [second]}}$$

(1)

We now define the Subject Normalized heartbeat frequency of the resting state in the SND as $f_{SNb}$ [beat SNsample], where SNsample stands for Subject Normalized sample, that is the independent variable of the SND.
The personal resampling frequency that permits to map the DTD PPG signal into the SND one is defined as $f_{SNc} = \frac{f_{Nb_{\text{beat}}}}{f_{SN_{\text{beat}}}}$, and can be calculated as follows:

$$f_{SNc} = \frac{f_{Nb_{\text{beat}}}}{f_{SN_{\text{beat}}}}$$ \hspace{1cm} (2)

As our goal is to obtain a domain where the inter-subject variability is discounted, the $f_{SNb}$ for all subjects’ baseline should be the same. Once this constant value is chosen, the $f_{SNc}$ resampling frequency for each subject can be calculated from Eq. 2. Then, all the PPG data of the same subject can be resampled accordingly and mapped into the SND, making the SN-data reliable for population based analysis. In Table 2, the notation introduced is summarized for the sake of clarity. Taking into account Eq. 1, Eq. 2 can be rewritten as follows:

$$f_{SNc} = \frac{f_b}{f_c} \cdot \frac{1}{f_{SNb}}$$ \hspace{1cm} (3)

The resting state heartbeat in the SND ($f_{SNb}$) can be arbitrarily chosen, only paying attention to possible aliasing effects. In our calculations we set $f_{SNb} = \frac{1}{128} \text{beat}_{\text{SNsample}}$, that corresponds to one beat on 128 SNsamples in the SND.

To analyze the effect of our normalization proposal, let’s consider some numerical examples. In the case of a sampling frequency $f_c = 128 \text{Hz}$ as in the case of the CLAWDAS dataset, we can observe from Eq.3 that, in the case of a subject with a baseline heartbeat frequency of $60 \text{beat}_{\text{minute}}$, corresponding to $1 \text{beat}_{\text{second}}$, the personalized resampling frequency is $f_{SNc} = \frac{1}{128} \text{beat}_{\text{SNsample}}$, meaning that there are no differences between the signal in the DTD and in the SND. Note that $60 \text{beat}_{\text{minute}}$ is generally considered as the minimum value for normal people. For heartbeat frequencies higher than $60 \text{beat}_{\text{minute}}$ the mapping from DTD to SND implies an over-sampling, while for lower frequencies, the consequent under-sampling does not introduce aliasing as $128$ samples are guaranteed between two consecutive peaks.

In Figure 4, PPG signals corresponding to the first baseline in the CLAWDAS Elderly dataset of subjects 11 and 18 respectively are considered. In the first row, the signals in the DTD are reported, showing the difference between the two subjects’ heartbeat frequency: for subject 11 $f_{Nb} = \frac{1}{78} \text{beat}_{\text{sample}}$, corresponding to $84 \text{beat}_{\text{minute}}$, while for subject 18 is $f_{Nb} = \frac{1}{94} \text{beat}_{\text{sample}}$ corresponding to about $82 \text{beat}_{\text{minute}}$. In the second row the same two signals resampled in the SND are shown. Note that we assumed in defining our procedure that the heartbeat during a resting state is a stationary and periodic signal, however this is not the case in real life, justifying not having $f_{Nb}$ strictly equal to $\frac{1}{128} \text{beat}_{\text{SNsample}}$ for both the subjects in the SND.

In case of multiple baseline signals, the heartbeat frequency is evaluate as the average of the heartbeat frequency of all of them. This procedure is applied, for example, during the normalization of the CLAWDAS signals. In this case, in fact, three different baseline signals have been acquired from each subject.

5. Classification Setting

A binary classification task is here proposed on CLAS, CLAWDAS Young and CLAWDAS Elderly, to evaluate the performance of the personalized PPG normalization presented in this work.

Two classes are considered: the class corresponding to signals collected during high cognitive load tasks: High CL, and the class related to low cognitive load tasks: Low CL. The tasks used as representative of each class change according to the dataset considered.
In particular, according to [24], in the CLAS dataset the PPG data collected during the three cognitive tasks (MP, ST and LP) are labeled as *High CL*, while the NS data are labeled as *Low CL*. In order to make the two classes balanced, MP, ST and LP signals are split into two non-overlapped segments of equal length, reaching a cardinality of 360 instances for both classes.

In the CLAWDAS Young and CLAWDAS Elderly, the data collected during the MC task are selected for the *High CL* class, while the data collected during the AL task are chosen as representatives of the *Low CL* class. Thereby, the two classes are equally balanced, with 96 instances each for the CLAWDAS Young, and 120 instances each for the CLAWDAS Elderly.

In all the performed analysis, seven handcrafted features are extracted as characteristics useful to describe the PPG signals:

- **Minima, Maxima, Mean and Standard Deviation** of the signal
- **Peak Rate**, which represents the mean number of peaks
- **Inter Beat Interval (IBI)**, which represents the mean distance between two peaks in a row
- **Root Mean Square of Successive Distance (RMSSD)** which represents the variance of the distance between two consecutive peaks [32].

The last three features are evaluated in the discrete domain and reported with respect to *samples*. For the sake of clarity, it is recalled that the meaning of *samples* changes according to the type of normalization strategy adopted. In particular, *samples* refer to subject normalized samples when the features is evaluated on signals with personalized normalization based on resting state heartbeat, while it refers to discrete time samples in all the other cases. All the features so evaluated are also standardized applying z-score before being used as input to the different classifiers.

Using the the seven features above introduced, three binary classification experiments are performed for each of the two datasets considered, comparing the following three normalization strategies:
Table 3. Summary of the classification settings

| Types of Normalization considered | AmpN, SubjFeatN, PersFreqN |
|----------------------------------|---------------------------|
| Features Used                    | Maximum, Minimum, Mean, Standard deviation, Peak Rate, IBI, RMSSD |
| Classifier Involved              | SVM Linear, SVM Cubic, SVM Gauss and Cart |
| Performance Evaluation method    | LOSO                       |
| Evaluation Metrics               | Accuracy, single class F1-score |

| Dataset                  | CL High Class Task | Num. of signals | CL Low Class Task | Num. of signals |
|--------------------------|--------------------|-----------------|------------------|-----------------|
| CLAS                     | Math, Stroop and Logic Test | 360 | Neutral State | 360 |
| CLAWDAS Young            | Math Calculation   | 96 | Audio Listening | 96 |
| CLAWDAS Elderly          | Math Calculation   | 120 | Audio Listening | 120 |

- **AmpN**: Amplitude normalization, as described in 3.2;
- **SubjFeatN**: Amplitude normalization followed by a subject feature normalization. This feature normalization is performed with respect to the subject baseline on peak rate, IBI and RMSSD features as follows:
  \[
  \text{featNorm}_i = \frac{\text{feature}_i - \text{featureBL}_i}{\text{featureBL}_i}
  \]
  where \(i \in \text{PeakRate, IBI, RMSSD}\); \(\text{feature}_i\) represents the feature value before the normalization; \(\text{featNorm}_i\) the new normalized value; and \(\text{featureBL}_i\) the mean value of the \(i\)–th feature evaluated on the subject resting state;
- **PersFreqN**: Amplitude normalization followed by the personalized normalization based on resting state heartbeat, described in chapter 4.

For each analysis, four different classification models are tested: a Classification and Regression Tree (Cart) with Gini’s diversity index as criterion of splitting and 100 as max number of decision split, and three Support Vector Machine (SVM) with different kernels: Linear (SVM Linear), Gaussian (SVM Gaussian) and polynomial cubic (SVM Cubic). In particular, for the gaussian kernel SVM, the kernel scale is set to 3.3 in order to consider a Medium Gaussian SVM.

A Leave One Subject Out (LOSO) Cross Validation ([33]) is applied to evaluate the performance of the trained classifiers. At each iteration, the data used to train the classifier consists of the signals collected from all the subjects except one, whose instances are instead used to test the performance of the model. An overall confusion matrix is finally generated, joining the single confusion matrices resulting from each iteration. From this confusion matrix, several well-known evaluation metrics are extracted. In particular, we have selected the accuracy to evaluate the general performance of the classifier and the single class F1-score ([34]) to
Table 4. Performance comparison on CLAS dataset, varying the normalization strategies (columns) and classification models (rows). Two performance metrics are evaluated using a LOSO validation strategy: accuracy (Acc) and single class F1-Score. The best performances reached for each type of normalization are underlined, while the the highest accuracy value of all is highlighted in bold.

| Classifier | AmpN | SubjFeatN | PersFreqN |
|------------|------|-----------|-----------|
|            | High CL | Low CL | High CL | Low CL | High CL | Low CL | High CL | Low CL |
| SVM Linear | Acc 66% | 0.65 | 0.67 | 67% | 0.68 | 0.67 | 76% | 0.77 | 0.76 |
| SVM Cubic  | 74% | 0.74 | 0.73 | 73% | 0.74 | 0.72 | 81% | 0.81 | 0.81 |
| SVM Gauss  | 72% | 0.72 | 0.71 | 73% | 0.72 | 0.73 | 78% | 0.78 | 0.77 |
| Cart       | 72% | 0.71 | 0.73 | 66% | 0.65 | 0.66 | 73% | 0.72 | 0.73 |

assess, instead, the goodness of the classification model in recognizing the single classes. The classification settings described above are summarized in Table 3.

6. Results and discussion

The performance of the classification settings described in Table 3 are reported in Tables 4, 5 and 6, for CLAS, CLAWDAS Young and CLAWDAS Elderly, datasets respectively. In particular, the results of the different normalization strategies are reported in term of accuracy and single class F1-score generated using the LOSO Cross Validation approach.

In all the experiments carried out, the normalization strategy here proposed, PersFreqN, outperforms the other two normalization procedures, for all the datasets. This observation is further supported by a visual comparison of the performance of the classifiers reported in the bar plot of Figure 5, varying the normalization strategy and the involved dataset.

Comparing the datasets, the best performance is observed on the CLAS dataset. In this case, the proposed PersFreqN allows to reach an accuracy of 81% adopting the SVM classifier with polynomial cubic kernel. This result significantly outperforms the accuracy of 74% reached using AmpN and 73% obtained using SubjFeatN.

In the CLAWDAS datasets, the highest accuracy achieved are 79% for CLAWDAS Young and 80% for CLAWDAS Elderly, both obtained with the proposed PersFreqN and SVM with linear kernel. These values show a significant improvement with respect to the other normalization strategies AmpN and SubjFeatN, that are always lower than 66%.

Another consideration regards the classifier that allows to reach the best results. In general, from the three Tables 4, 5 and 6 it emerges that the highest accuracy values are usually achieved by the SVM classifiers with linear or cubic kernel, whereas the lowest ones are generally obtained by the Cart classifier.

Finally, a last consideration should be done on the SubjFeatN normalization strategy. In general, this normalization produces performance higher than the AmpN one, even if it appears less effective compared to the proposed PersFreqN strategy. These results confirm that a normalization strategy that takes into account not only amplitude normalization but also subject’s characteristics, should be adopted to remove inter-subject variability.

As a final remark, all the adopted classification settings are able to classify with comparable performance both the two classes, as indicated by the values of the single class F1-score in all the Tables. However, the introduction of the proposed PersFreqN seems to produce even more balanced classification results.
Table 5. Performance comparison on CLAWDAS Young dataset, varying the normalization strategies (columns) and classification models (rows). Two performance metrics are evaluated using a LOSO validation strategy: Accuracy (Acc) and single class F1-Score. The best performances reached for each type of normalization are underlined, while the the highest accuracy value of all is highlighted in bold.

| Classifier | AmpN | SubjFeatN | PersFreqN |
|------------|------|-----------|-----------|
|            | High CL | Low CL | High CL | Low CL | High CL | Low CL | High CL | Low CL | Acc | F1-Score | F1-Score | Acc | F1-Score | F1-Score | Acc | F1-Score | F1-Score |
| SVM Linear | 66% 0,62 0,69 | 68% 0,63 0,72 | 79% 0,77 0,80 |
| SVM Cubic  | 66% 0,66 0,66 | 68% 0,67 0,68 | 72% 0,72 0,72 |
| SVM Gauss  | 63% 0,59 0,66 | 66% 0,61 0,70 | 76% 0,75 0,76 |
| Cart       | 56% 0,59 0,51 | 68% 0,67 0,69 | 64% 0,63 0,64 |

Table 6. Performance comparison on CLAWDAS Elderly, varying the normalization strategies (columns) and classification models (rows). Two performance metrics are evaluated using a LOSO validation strategy: Accuracy (Acc) and single class F1-Score. The best performances reached for each type of normalization are underlined, while the the highest accuracy value of all is highlighted in bold.

| Classifier | AmpN | SubjFeatN | PersFreqN |
|------------|------|-----------|-----------|
|            | High CL | Low CL | High CL | Low CL | High CL | Low CL | High CL | Low CL | Acc | F1-Score | F1-Score | Acc | F1-Score | F1-Score | Acc | F1-Score | F1-Score |
| SVM Linear | 59% 0,58 0,60 | 69% 0,65 0,72 | 80% 0,80 0,81 |
| SVM Cubic  | 59% 0,55 0,62 | 64% 0,61 0,66 | 72% 0,71 0,73 |
| SVM Gauss  | 63% 0,61 0,64 | 75% 0,74 0,76 | 78% 0,78 0,78 |
| Cart       | 54% 0,51 0,57 | 68% 0,68 0,67 | 75% 0,75 0,74 |

7. Conclusions

While considering physiological data, signal normalization not only with respect to amplitude but also with respect to personal characteristics is mandatory to perform subject based analysis, especially if machine learning techniques should be applied. Personalized normalization on PPG data, both at the feature level, SubjFeatN, and with respect to heartbeat frequency, PersFreqN, introduces an increase in the classification performance, considering different classification models and datasets. In particular, the personalized PPG normalization based on subject heartbeat here proposed, PersFreqN, outperforms the other strategies and permits to significantly reduce inter-subject heterogeneity. Moreover, the proposed normalization could be also useful for intra-subject analysis, especially when comparing the physiological responses of the same subject, in different days or even in different moment of the day: it is well known, in fact, that the physiological responses not only depend on external stimuli, but also on physical and internal conditions that can significantly vary for the same subject with respect to time.

Author contributions

F.G: Conceptualisation, methodology, formal analysis, investigation, data curation, writing original draft preparation; A.G: software, validation, formal analysis, investigation, data curation, writing original draft preparation, visualisation; M.G: investigation, writing review
Figure 5. Bar plot comparison of the accuracy obtained using different classifiers and normalization procedures. Three dataset are considered: CLAS, CLAWDAS Young and CLAWDAS Elderly.

and editing; S.B.: project leader, supervision. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research is partially supported by the FONDAZIONE CARIPLO “LONGEVICITY-Social Inclusion for the Elderly through Walkability” (Ref. 2017-0938) and by the Japan Society for the Promotion of Science (Ref. L19513). We want to give our thanks to Prof. Katsuhiro Nishinari and his staff, in particular Kenichiro Shimura and Daichi Yanagisawa for their indispensable support during the experiment held at RCAST - The University of Tokyo.

Abbreviations

The following abbreviations are used in this manuscript:
| Acronym   | Description                                                                 |
|-----------|-----------------------------------------------------------------------------|
| PPG       | Photoplethysmography                                                         |
| CLAS dataset | Cognitive Load, Affect and Stress Recognition dataset                      |
| CLAWDAS dataset | Cognitive Load and Affective Walkability in Different Age Subjects dataset |
| BL        | Baseline                                                                    |
| MP        | Math Problems                                                               |
| ST        | Stroop Test                                                                 |
| LP        | Logic Problems                                                              |
| NS        | Neutral State                                                              |
| MC        | Math Calculation                                                           |
| R         | Reading                                                                     |
| C         | Comprehension                                                               |
| SWT       | Stationary Wavelet Transform                                                |
| CTD       | Continuous Time Domain                                                      |
| DTD       | Discrete Time Domain                                                        |
| SND       | Subject Normalized discrete Domain                                          |
| High CL   | High Cognitive Load                                                         |
| Low CL    | Low Cognitive Load                                                          |
| IBI       | Inter Beat Interval                                                        |
| RMSSD     | Root Mean of Successive Distance                                           |
| SVM       | Support Vector Machine                                                      |
| SVM Linear | Support Vector Machine with Linear Kernel                                 |
| SVM Cubic | Support Vector Machine with Polynomial Cubic Kernel                         |
| SVM Gaus  | Support Vector Machine with Gaussian Kernel                                 |
| Cart      | Classification and Regression Tree                                          |
| LOSO Cross Validation | Leave One Subject Out Cross Validation                                   |
| AmpN      | Amplitude Normalization                                                     |
| SubjFeatN | Amplitude normalization followed by a Subject Feature Normalization.         |
| PersFreqN | Personalized Normalization based on baseline heartbeat Frequency            |
References

1. Can, Y.S.; Arnrich, B.; Ersoy, C. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics* **2019**, *92*, 103139.

2. Han, H.J.; Labba, S.; Borelli, J.L.; Dutt, N.; Rahmani, A.M. Objective stress monitoring based on wearable sensors in everyday settings. *Journal of Medical Engineering & Technology* **2020**, *44*, 177–189.

3. Choi, J.; Ahmed, B.; Gutierrez-Osuna, R. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE transactions on information technology in biomedicine* **2011**, *16*, 279–286.

4. Umematsu, T.; Sano, A.; Taylor, S.; Picard, R.W. Improving students’ daily life stress forecasting using LSTM neural networks. 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2019, pp. 1–4.

5. Durán-Acevedo, C.M.; Carrillo-Gómez, J.K.; Albarracín-Rojas, C.A. Electronic Devices for Stress Detection in Academic Contexts during Confinement Because of the COVID-19 Pandemic. *Electronics* **2021**, *10*, 301.

6. Setz, C.; Arnrich, B.; Schumm, J.; La Marca, R.; Tröster, G.; Ehler, U. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine* **2009**, *14*, 410–417.

7. Saeed, A.; Trajanovski, S.; Van Keulen, M.; Van Erp, J. Deep physiological arousal detection in a driving simulator using wearable sensors. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017, pp. 486–493.

8. Bandini, S.; Gasparini, F.; Giltri, M. Personalized music experience for the wellbeing of elderly people. International Conference on Internet Science. Springer, 2019, pp. 335–340.

9. Vachiratamporn, V.; Moriyama, K.; Fukui, K.I.; Numao, M. An implementation of affective adaptation in survival horror games. 2014 IEEE conference on computational intelligence and games. IEEE, 2014, pp. 1–8.

10. Fonseca, D.; da Silva, K.C.N.; Rosa, R.L.; Rodriguez, D.Z. Monitoring and Classification of Emotions in Elderly People. 2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). IEEE, 2019, pp. 1–6.

11. Bartolomé-Tomás, A.; Sánchez-Reolid, R.; Latorre, A.F.S.J.M.; Fernández-Caballero, A. Arousal Detection in Elderly People from Electrodermal Activity Using Musical Stimuli. *Sensors* **2020**, *20*, 4788.

12. Costadopoulos, N.; Islam, M.Z.; Tien, D. Using Z-score to Extract Human Readable Logic Rules from Physiological Data. *NPJ digital medicine* **2019**, *2*, 1–10.

13. Goh, C.H.; Tan, L.K.; Lovell, N.H.; Ng, S.C.; Tan, M.P.; Lim, E. Robust PPG motion artifact detection using a 1-D convolution neural network. *Computer methods and programs in biomedicine* **2020**, *196*, 105596.

14. Lee, S.W.; Woo, D.K.; Son, Y.K.; Mah, P.S. Wearable Bio-Signal (PPG)-Based Personal Authentication Method Using Random Forest and Period Setting Considering the Feature of PPG Signals. *JCP* **2019**, *14*, 283–294.

15. Nemati, S.; Ghassemi, M.M.; Ambai, V.; Isakadze, N.; Levantsevych, O.; Shah, A.; Clifford, G.D. Monitoring and detecting atrial fibrillation using wearable technology. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016, pp. 3394–3397.

16. Costadopoulos, N.; Islam, M.Z.; Tien, D. Using Z-score to Extract Human Readable Logic Rules from Physiological Data. 2019 11th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2019, pp. 1–6.

17. Gasparini, F.; Grossi, A.; Nishinari, K.; Bandini, S. Age-Related Walkability Assessment: A Preliminary Study Based on the EMG. *AIxIA 2020 – Advances in Artificial Intelligence; Baldoni, M.; Bandini, S., Eds.; Springer International Publishing: Cham, 2021*; pp. 423–438.

18. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis, using physiological signals. *IEEE transactions on affective computing* **2011**, *3*, 18–31.
26. Biswas, A.; Roy, M.S.; Gupta, R. Motion artifact reduction from finger photoplethysmogram using discrete wavelet transform. In Recent Trends in Signal and Image Processing; Springer, 2019; pp. 89–98.
27. Raghuram, M.; Madhav, K.V.; Krishna, E.H.; Reddy, K.A. On the performance of wavelets in reducing motion artifacts from photoplethysmographic signals. 2010 4th international conference on bioinformatics and biomedical engineering. IEEE, 2010, pp. 1–4.
28. Nason, G.P.; Silverman, B.W. The stationary wavelet transform and some statistical applications. In Wavelets and statistics; Springer, 1995; pp. 281–299.
29. Li, W. Wavelets for electrocardiogram: overview and taxonomy. IEEE Access 2018, 7, 25627–25649.
30. Donoho, D.L.; Johnstone, J.M. Ideal spatial adaptation by wavelet shrinkage. biometrika 1994, 81, 425–455.
31. Holschneider, M.; Kronland-Martinet, R.; Morlet, J.; Tchamitchian, P. A real-time algorithm for signal analysis with the help of the wavelet transform. In Wavelets; Springer, 1990; pp. 286–297.
32. Stein, P.K.; Bosner, M.S.; Kleiger, R.E.; Conger, B.M. Heart rate variability: a measure of cardiac autonomic tone. American heart journal 1994, 127, 1376–1381.
33. Schmidt, P.; Reiss, A.; Dürichen, R.; Laerhoven, K.V. Wearable-based affect recognition—A review. Sensors 2019, 19, 4079.
34. Bishop, C.M. Pattern recognition and machine learning; Springer, 2006.