Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

Comparison of automatic and physiologically-based feature selection methods for classifying physiological stress using heart rate and pulse rate variability indices

To cite this article: Marta Iovino *et al* 2024 *Physiol. Meas.* **45** 115004

IOP Publishing Physiol. Meas. 45 (2024) 115004 https://doi.org/10.1088/1361-6579/ad9234

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
21 December 2023

REVISED
18 October 2024

ACCEPTED FOR PUBLICATION
13 November 2024

PUBLISHED
25 November 2024

PAPER

Comparison of automatic and physiologically-based feature selection methods for classifying physiological stress using heart rate and pulse rate variability indices

Marta Iovino[1], Ivan Lazic[2], Tatjana Loncar-Turukalo[2], Michal Javorka[3], Riccardo Pernice[1,*] and Luca Faes[1]

[1] Department of Engineering, University of Palermo, Palermo, Italy
[2] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[3] Department of Physiology, Comenius University in Bratislava, Jessenius Faculty of Medicine, Martin, Slovakia
[*] Author to whom any correspondence should be addressed.

E-mail: riccardo.pernice@unipa.it

Keywords: stress classification, heart rate variability (HRV), pulse rate variability (PRV), machine learning (ML), feature selection (FS)

Abstract

Objective. This study evaluates the effectiveness of four machine learning algorithms in classifying physiological stress using heart rate variability (HRV) and pulse rate variability (PRV) time series, comparing an automatic feature selection based on Akaike's criterion to a physiologically-based feature selection approach. *Approach.* Linear discriminant analysis, support vector machines, *K*-nearest neighbors and random forest were applied on ten HRV and PRV indices from time, frequency and information domains, selected with the two feature selection approaches. Data were collected from 127 healthy individuals during different stress conditions (rest, postural and mental stress). *Main results.* Our results highlight that, while specific stress classification is feasible, distinguishing between postural and mental stress remains challenging. The used classifiers exhibited similar performance, with automatic Akaike Information Criterion-based feature selection proving overall better than the physiology-driven approach. Additionally, PRV-based features performed comparably to HRV-based ones, indicating their potential in outpatient monitoring using wearable devices. *Significance.* The obtained findings help to determine the most relevant HRV/PRV features for stress classification, potentially useful to highlight different physiological mechanisms involved during both challenges accompanied by a shift in the sympathovagal balance. The proposed approach may have implications for advancing stress assessment methodologies in clinical settings and real-world contexts for well-being evaluation.

1. Introduction

The American Psychological Association has defined stress as 'the pattern of specific and non-specific responses an organism makes to stimuli events that disturb its equilibrium' (Gerrig and Zimbardo 2002). Excess stress has been identified as one of the most powerful pathogenic elements of life (Umair *et al* 2021). In simple terms, psychological stress is the feeling we experience when we are overwhelmed and struggling to handle the demands placed upon us. Theoretically, stress can be beneficial as it serves as a motivator, aids survival, and can alert us to potential dangers. However, when it becomes frequent, it can harm our mental and physical health raising the risk of heart disease, accelerating ageing, and making us more susceptible to mental health issues (Jiménez-Limas *et al* 2018). On the other hand, orthostatic stress is a physical stress caused by the effects of gravity on the distribution of circulating blood volume in the body due to a position change of the body (Benditt *et al* 2004). The body's ability to react to a variety of stimuli (such as environmental and psychological stressors) affects the sympathovagal balance (SVB) and thus the complexity of cardiovascular regulation, due to the inhibition of the parasympathetic and activation of the sympathetic branches of the autonomous nervous system during stress (Shaffer and Ginsberg 2017).

A common approach for noninvasive stress assessment involves studying heart rate variability (HRV), which is the beat-to-beat variation of the heart rate (HR). HRV is typically assessed by analysing time series extracted from consecutive electrocardiographic signal (ECG) R peaks (R-R intervals). The most commonly adopted HRV analysis for practical applications consists of taking into account short-term measurements (i.e. 300 heartbeats, ∼5 min recordings), and computing time-, frequency- and information-domain measures (Shaffer and Ginsberg 2017, Pernice *et al* 2019). HRV indices are among the most dependable indicators of mental and physical stress (Shaffer and Ginsberg 2017).

Recently, there has been a growing interest in investigating whether and to what extent HRV can also be assessed through the photoplethysmographic signal (PPG), whose cardiovascular variability indices are usually referred to as pulse rate variability (PRV) and are computed from pulse-to-pulse intervals (PPIs) (Shaffer and Ginsberg 2017, Pernice *et al* 2019). PPG is an optical technique used in wearable devices that can detect changes in finger microvascular blood volume. It is simple, low-cost, safe, and minimally invasive (Mejía-Mejía *et al* 2020, Scardulla *et al* 2023). Although PPG and ECG are often considered interchangeable for measuring HRV, several pieces of evidence suggest that the beat-to-beat variability recorded with PPG is somewhat different from HRV (Pernice *et al* 2019, Mejía-Mejía *et al* 2020, Rinella *et al* 2022). PPG and blood pressure recordings can be affected by physiological factors related to pulse wave transmission through the vascular system and measurement errors due to motion-induced signal corruption and lower peak detection accuracy, thus reducing the agreement between PRV and HRV.

In recent years, the application of machine learning (ML) to medical data analysis has surged in both research and healthcare. ML algorithms are versatile in handling tasks such as prediction, classification, and decision-making, and are designed to manage the large, complex datasets typical in healthcare (Ahmad *et al* 2022). ML has been particularly effective in categorizing autonomic nervous system states related to various stress types (Awasthi *et al* 2020). Assessing HRV and PRV characteristics for stress level classification has become a significant research focus, with numerous studies using HRV (Giannakakis *et al* 2019, Dalmeida and Masala 2021) and PRV (Awasthi *et al* 2020, Panganiban and de Leon 2021) indices across time and frequency domains.

In previous preliminary works (Iovino *et al* 2023a, 2023b), we evaluated the efficacy of several ML algorithms for classifying postural and mental stress, employing either short-term or ultra-short-term HRV and PRV indices. This work builds on those studies by including an initial feature selection phase.

The goals of this study are twofold. First, we aim to assess the feasibility of differentiating stress conditions from a resting state, using physiologically meaningful indices derived from HRV. Second, we aim to establish the feasibility of performing such differentiation using PRV-based indices for future automatic stress classifiers based on wearable physiological monitoring. To this end, we extract a comprehensive set of features in three domains (time, frequency, and information-theoretic) from HRV and PRV series measured in a large group of healthy subjects monitored at rest and during postural and mental stress, evoked by head-up tilt and mental arithmetic tasks, respectively. We then employ two feature selection methods to identify the most relevant features for stress discrimination, one automatic based on the Akaike criterion and one supervised by an expert physiologist. Finally, we use four widely used ML algorithms to compare the classification performance of HRV-based and PRV-based classifications, both for the automatic and the manual feature selection approaches.

2. Materials and methods

2.1. Subjects and experimental protocol

Analyses were carried out on a dataset specifically acquired to assess cardiovascular variability during different physiological states encompassing rest, orthostatic, and mental stress. Data were collected from 127 young healthy volunteers (75 females, 52 males; age: 18.63 $\pm$ 3.27 years), all normotensive and with body mass index in a normal range (BMI: 21.42 $\pm$ 2.20 kg m^{-2}). All the procedures were approved by the Ethics Committee of the Jessenius Faculty of Medicine, Comenius University, Martin, Slovakia. The participants signed written informed consent, and when the subject was a minor (age $<$ 18 years), parental or legal guardian permission was gathered to allow the child to participate in the study. The analysed physiological signals consisted of horizontal bipolar thoracic lead ECG recordings acquired through Cardiofax ECG-9620 (Nihon Kohden, Japan) and arterial blood pressure obtained through the volume-clamp photoplethysmographic method using the Finometer Pro device (FMS, Netherlands), which measures beat-to-beat arterial pressure variability. In the Finometer device, the recorded continuous blood pressure (CBP) is related to the PPG signal since a pulsation of the arterial diameter during a heartbeat produces a pulsation in the photodetected signal, e.g. timings of the CBP signal are based on plethysmographic principle (Pernice *et al* 2019). All signals were acquired synchronously with a sampling frequency of 1 kHz. The

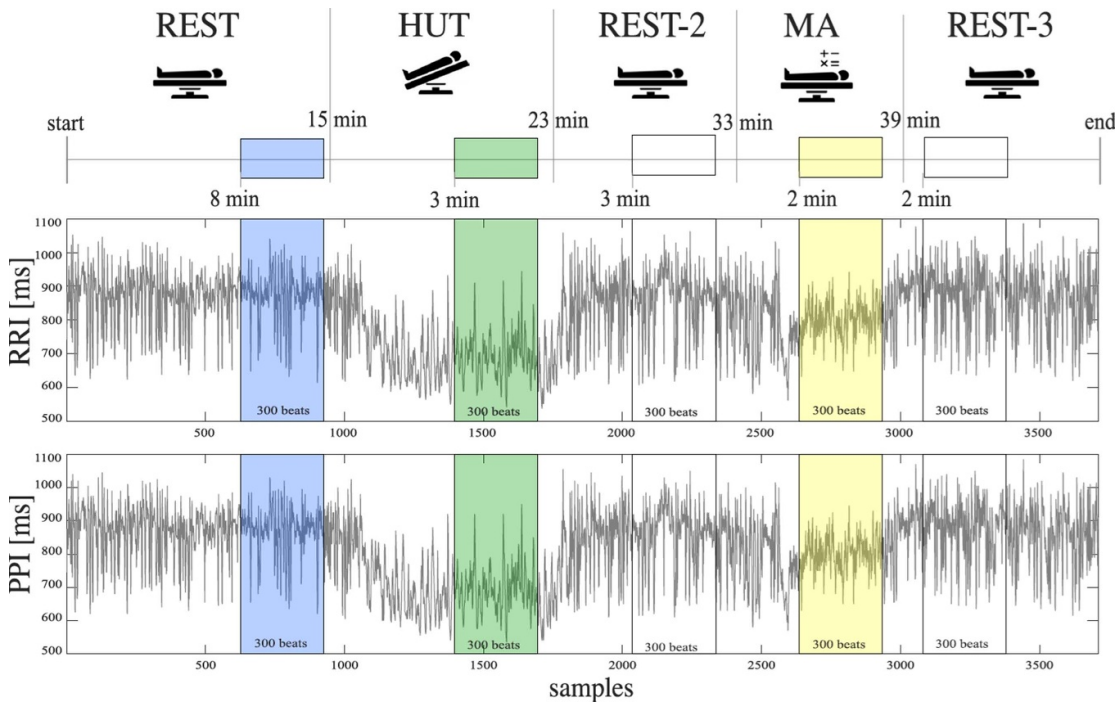


Figure 1. (a) Schematic illustration of the experimental protocol, including baseline resting (REST), orthostatic stress (HUT), second resting (REST-2), mental arithmetic (MA) and the final third recovery phase (REST-3). (b) Representative RRI and PPI time series, extracted respectively from ECG and blood pressure recordings. Colour boxes indicate the windows taken into account for short-term (300 points) analysis.

experimental protocol encompasses five phases alternating physiological stress and recovery states (schematically represented in figure 1):

- REST: subjects resting in the supine position on a motorized bed for 15 min (baseline);
- HUT: 8 min-lasting head-up tilt test to evoke orthostatic stress;
- REST-2: resting phase in the supine position for 10 min for a full recovery of physiological parameters;
- MA: subjects lying in the supine position and undergoing a mental arithmetic test for 6 min to induce cognitive load;
- REST-3: 10 min resting phase in the supine position to let the physiological parameters recover again.

The passive head-up-tilt test was performed by tilting the motorized table on which the volunteers were laying to a 45° upright position. The non-verbal mental test was carried out using the WQuick software with the WIN 5 PMT test (Psycho Soft Software, s.r.o., Brno, Czech Republic) and consisted of a repetitive visualization on the room ceiling of randomly generated 3-digit numbers. Each subject was asked to read the number and mentally sum the digits as quickly as possible until a one-digit number was reached; then, the subject had to decide whether the final resulting number was even or odd by using a computer mouse to click the corresponding virtual button also projected on the ceiling.

2.2. Time series extraction and preprocessing

For each subject and condition, two time series of 300 heartbeats per the standard short-term analysis protocol were extracted from the acquired ECG and blood pressure signals, respectively. To calculate the nth RR interval (RRI), the time interval between the nth and $(n+1)$th QRS apexes was derived from the ECG data. On the other hand, for the nth PPI, the time interval between the nth and $(n+1)$th blood pressure maxima was measured. This approach ensured that both RRI and PPI time series were of the same length for each subject and physiological condition. The 300-point windows were extracted from the recorded signals starting approximately 8 min after the beginning of the REST phase, around 3 min after the start of the HUT phase, around 3 min after the start of the REST-2 phase, approximately 2 min after the start of MA phase, and around 2 min after the start of REST-3 phase (figure 1(b)). This choice has been made to allow the recovery of the physiological parameters with regard to resting states, and to avoid considering the initial transition between rest and stress states. The 300-point windows that were analysed did not contain any

artefacts, including those related to the calibration of the Finometer device. It is worth noting that the calibration process, which interrupts the continuous measurement of CBP, was only performed in the last minute of REST and REST-2 phases, disregarded in the analyses. Each time series was corrected to remove outliers beyond three standard deviations from the mean. These outliers were replaced using two different interpolation algorithms, i.e. linear for isolated outliers and cubic spline for multiple successive ones. The corrected time series were visually inspected for any issues following the interpolation procedure. Overall, only 0.6% of the samples of each time series were adjusted, allowing a more reliable analysis. In this study, only three out of the five different phases were considered, i.e. REST, HUT and MA phases. This choice has been made since the other two recovery phases (REST-2 and REST-3) are very similar to each other (Pernice *et al* 2019) and the baseline condition, and it would thus not be useful to carry out classification among them.

2.3. Feature extraction

Ten features, belonging to three domains (time, frequency and information-theoretic) were calculated on RRI and PPI series, as detailed in previous works (Pernice *et al* 2019, Volpes *et al* 2022). Before conducting frequency- and information-domain analyses, both time series underwent additional preprocessing steps, encompassing first a zero-phase IIR high-pass filtering with a cutoff frequency of 0.015 Hz to remove the mean and the slow trends (Pernice *et al* 2019). The sampling frequency was selected under the assumption that the RRI and PPI time series are evenly sampled with a sampling period equivalent to the mean RRI, according to a common practice in HR variability analysis (Faes *et al* 2013, Pernice *et al* 2021). Before computing information-theoretic measures, the time series were also normalized to unit variance.

2.3.1. Time-domain

Conventional time-domain metrics for HRV and PRV were computed on RRI and PPI time series to extract information about the beat-to-beat variability and the vagally mediated changes reflected in HRV. Specifically, the following time-domain indices were calculated: the average value (MEAN), the standard deviation (SDNN), and the root mean square of successive differences (RMSSD) in each time series (Shaffer and Ginsberg 2017).

2.3.2. Frequency-domain

A parametric spectral analysis was employed for the frequency domain, fitting the preprocessed time series with an autoregressive model. Here, model identification was carried out using the ordinary least squares method. Instead of relying on standard model order selection criteria, a fixed model order of $p = 10$ was chosen to ensure the representation of various oscillations (for further details we refer the reader to Pernice *et al* 2019). The absolute spectral power was then computed within two frequency bands: the low frequency (LF) band, ranging from 0.04 to 0.15 Hz, and the high frequency (HF) band, ranging from 0.15 to 0.4 Hz. Additionally, the normalized power within the HF band (HFn) was calculated by dividing the absolute power within the band by the sum of the absolute powers of both LF and HF bands (Shaffer and Ginsberg 2017). The LF-to-HF power ratio, used in past studies as the SVB index, was also determined. Furthermore, the respiratory peak frequency f_{HF} was identified as the peak frequency within the HF band (Shaffer and Ginsberg 2017).

2.3.3. Information-domain

Information-theoretic measures were computed on both the RRI and PPI time series to assess the amount of information conveyed and to measure their complexity. This analysis quantifies the predictability of the current sample of the time series, considering its past samples, and is thus associated with the regularity of the time series. To achieve this, two entropy measures were calculated on each time series: the entropy that can be derived from the history of the process (self entropy, SE) and the part that cannot be derived from it (conditional entropy, CE) (Xiong *et al* 2017, Valente *et al* 2018). Both CE and SE features were estimated using the model-free k-nearest neighbour approach (number of neighbours and embedding dimension set to 10 and 2 respectively), as in (Xiong *et al* 2017, Volpes *et al* 2022).

The analyses were entirely conducted on MATLAB 2022b (The MathWorks, Inc. Natick, MA, USA) and the 'Information Theory for the Analysis of Physiological Time Series—ITS Toolbox v 2.1' http://www.lucafaes.net/its.html was used to compute information-theoretic indices (Faes *et al* 2015, 2016).

2.4. Feature selection

Feature selection reduces computation, mitigates high dimensionality, and enhances predictor performance by identifying a subset of variables that effectively describe the dataset while minimizing noise (Chandrashekar and Sahin 2014, Ying 2019, Aggrawal and Pal 2020). This is crucial for our study as it helps

highlight the most physiologically meaningful HRV and PRV features, improving the classification of stress conditions and providing deeper insights into the underlying physiological mechanisms.

2.4.1. Akaike feature selection method

In this study, feature selection based on the Akaike information criterion (AIC) was performed (Vetrov *et al* 2009). AIC is a widely used metric for model selection that penalizes model complexity to prevent overfitting, aiming to find a model that provides a good fit with the least number of predictors (Bishop and Nasrabadi 2006, Vetrov *et al* 2009). For each combination of features, a multinomial logistic regression model (Bishop and Nasrabadi 2006, Vetrov *et al* 2009) was trained using continuous features and categorical target variables, computing the corresponding AIC value. The algorithm systematically evaluated all 1023 possible combinations of features to identify the subset that minimized the AIC, which was the primary criterion for model selection. This exhaustive search, although time-consuming, was preferred because it ensures an exact solution by avoiding local minima. This precision is crucial since feature selection is a key goal of our analysis, providing a basis for interpreting the physiological significance of the selected features. The combination of features yielding the minimum AIC value was used for the subsequent analysis of physiological state classification.

2.4.2. Physiologically-based feature selection method

In addition to the automatic AIC-FS approach for feature selection, an alternative feature selection method was employed. In this case, the focus was to exclusively select features according to the highest physiological relevance in a medical context, based on current literature. This shift from conventional methods aimed to prioritize features intrinsically linked to physiological mechanisms. The considered HRV/PRV features set encompassed the MEAN as the basic cardiovascular measure expressing the overall balance between parasympathetic and sympathetic control, the SDNN that describes the overall magnitude of HRV (Shaffer and Ginsberg 2017), the information-theoretical CE as an independent measure of signal complexity related to sympathetic activation (Porta *et al* 2016), and both frequency-domain indices, i.e. LF (containing relatively independent information including vascular control reflecting sympathetic activity and baroreflex influence) and HF (classical parasympathetic activity index) absolute spectral power in the respective bands, being them mutually more independent and relatively easy to interpret from a physiological perspective.

2.5. Classification algorithms

Four classic ML classifiers (linear discriminant analysis (LDA), support vector machines (SVM), *k*-nearest neighbors (kNN), random forest (RF)) (Lim *et al* 2016, Ahmad *et al* 2022, Rani *et al* 2022) were selected to evaluate and compare their performance in discriminating between rest, orthostatic stress and mental stress. The classifiers were chosen given their widespread use for various classification tasks, including those involving ECG signals (Lim *et al* 2016). Each of the four classifiers was trained by inputting the HRV or PRV features. The classes were balanced since each of the 127 subjects went through the three conditions (REST, HUT, and MA). The integrated 'Statistics and Machine Learning Toolbox' of MATLAB 2022b was used to apply the classifiers and the feature selection algorithm. All classifiers were evaluated by using a *k*-fold cross-validation (CV), with a nested CV for optimal hyperparameter search (outlined in detail in section 2.6) and leaving the remaining hyperparameters at their default MATLAB 2022b values.

In this work, the following hyparameters of the LDA classifier were optimized: gamma or the amount of regularization to apply when estimating the covariance matrix of the predictors, and Delta or the linear coefficient threshold. Delta was searched among positive values, by default log-scaled in the range $[1 \times 10^{-6}, 1 \times 10^{3}]$, while gamma was searched among real values in the range [0,1] (Bishop and Nasrabadi 2006).

For the SVM classifier, it was decided to optimize all the possible hyparameters, including the soft margin penalty (i.e. the box constraint hyparameter in the used MATLAB function) [positive values log-scaled in the range from 1×10^{-3} to 1×10^{3}], the kernel function [linear or polynomial kernel function of order 2] used to compute the elements of the Gram matrix and the Kernel Scale (the appropriate kernel norm to compute the Gram matrix) [positive values log-scaled in the range from 1×10^{-3} to 1×10^{3}] (Bishop and Nasrabadi 2006).

The following hyparameters of the kNN classifier were optimized: the distance metric, such as the city-block or the Chebychev, and the number of nearest neighbours to search, for classifying each point when predicting, which the algorithm usually searches among positive integer values, by default log-scaled in the range between 1 and the maximum number of observations depending on the CV approach (please refer to section 2.6 for the details) (Bishop and Nasrabadi 2006).

For RF classifier, it was decided to optimize all the possible hyparameters including the maximum number of decision splits (or trees), integers log-scaled in the range between 1 and the maximum number of

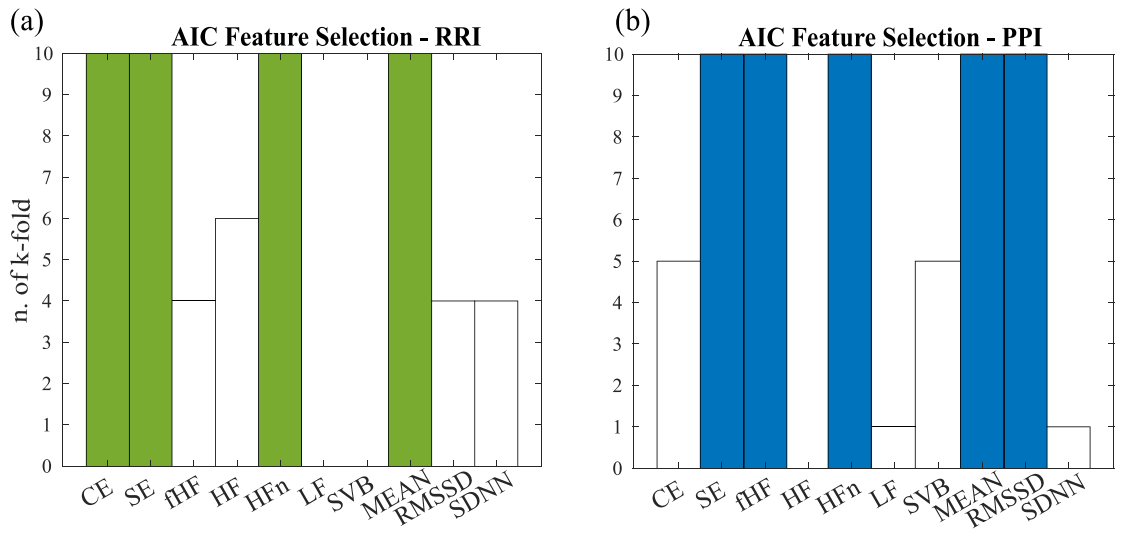


Figure 2. Frequency of selected features using Akaike information criterion across 10-fold cross-validation. The bar plot illustrates the number of times each feature was selected during the feature selection process. The coloured bars represent the features that were selected more frequently (>70%) for the RRI and PPI time series.

observations, the minimum number of leaf node observations, integers log-scaled in the range between 1 and the maximum number of observations, and the split criteria (Gini's diversity index) (Bishop and Nasrabadi 2006).

2.6. Evaluation pipeline

Due to the relatively small amount of data, a 10-fold CV was applied splitting the dataset into training and test sets. In each iteration of the outer CV, the training folds were subjected to the described FS methods and to an additional inner 10-fold CV to determine the optimal model hyparameters. This process was carried out before generating the final results on the outer CV test sets.

The performance of the various classifiers in discriminating among the various physiological states was assessed employing conventional evaluation metrics (Shah *et al* 2018, Aggrawal and Pal 2020, Ahmad *et al* 2022). Specifically, in multi-class classification, accuracy, recall, precision, and F1 score were first computed for each class, allowing to assess the classifier performance on each class, and are then averaged across the classes.

The statistical significance of the obtained results was evaluated through statistical tests, to assess differences among the classifiers and between the time series. In detail, the statistical tests were applied to the distributions of accuracy, recall, precision, and F1 score values extracted across the initial 10-fold CV step. To assess statistical significance between classifiers, the non-parametric Kruskal–Wallis test was applied to the distributions of each metric, separately for the RRI and PPI time series, followed by a post-hoc unpaired ranksum test with Bonferroni–Holm correction for multiple comparisons ($n = 6$, i.e. the number of pairwise combinations between the classifiers). Additionally, the Wilcoxon paired signed-rank test was used to compare the performance metrics between the RRI and PPI time series for each classifier, to highlight any significant differences due to the time series characteristics. The same statistical tests were also employed to assess any differences between the two feature selection methods.

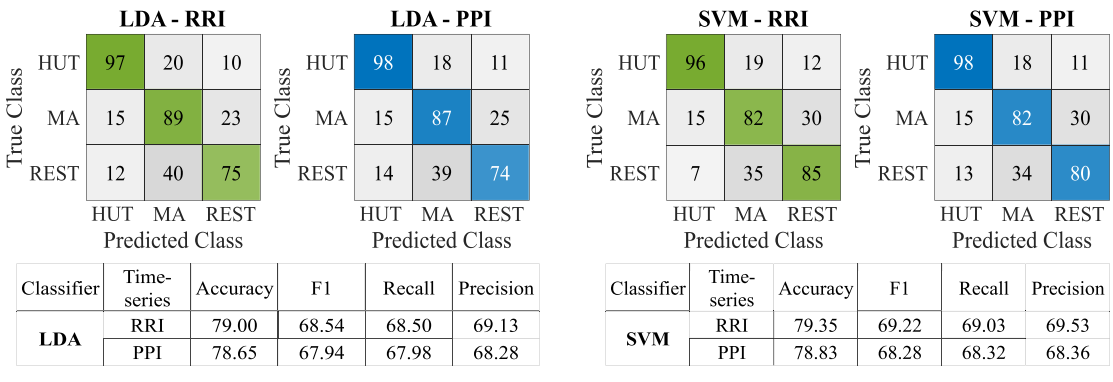
3. Results

In the following subsections, the results will be presented, focusing on classification according to the AIC-FS algorithm (section 3.1) or the physiologically-based feature selection method (section 3.2).

3.1. Akaike feature selection method

The results in figure 2 show the frequency of selection for each feature across the 10 CV folds for (a) RRI and (b) PPI time series using the AIC-FS algorithm. Regarding the RRI time series, features such as 'CE', 'SE', 'HFn', and 'MEAN' demonstrated higher selection frequencies, more than 70% over the 10-fold, while for the PPI time series features like 'SE', 'fHF', 'HFn', 'MEAN', and 'RMSSD' were more commonly selected.

As shown in the confusion matrices of figure 3, the HUT class consistently ranks best, achieving higher True Positive values than the other two classes. On the other hand, the high values shown in grey in the



Classifier	Time-series	Accuracy	F1	Recall	Precision
LDA	RRI	79.00	68.54	68.50	69.13
	PPI	78.65	67.94	67.98	68.28
SVM	RRI	79.35	69.22	69.03	69.53
	PPI	78.83	68.28	68.32	68.36
kNN	RRI	78.83	68.32	68.24	68.63
	PPI	77.08	65.42	65.62	65.50
RF	RRI	79.88	69.85	68.82	69.90
	PPI	79.00	68.57	68.50	68.68

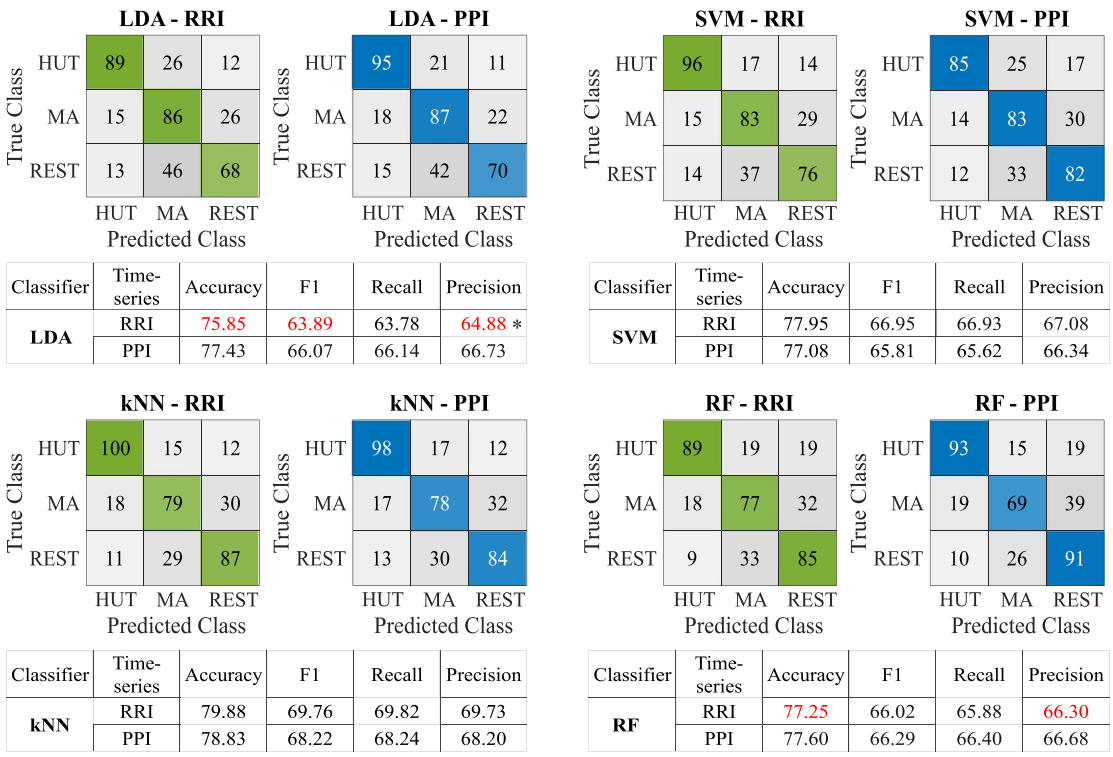
Figure 3. Confusion matrices, tables of the classification performance results (accuracy, recall, precision and F1 score [%]) for each classifier (LDA, SVM, kNN, RF) and each time series (RRI and PPI). No statistically significant differences were observed in the comparisons between RRI and PPI, or across the classifiers.

confusion matrix suggest that both MA and REST phases have been more often incorrectly classified. Across the four classifiers (LDA, SVM, KNN, and RF) the RF generally achieves the highest results for both time series (figure 3). In contrast, kNN showed the weakest performance across all computed metrics compared to the other classifiers. However, the Wilcoxon signed-rank test for the four metrics (accuracy, recall, precision, and F1 score) across the classifiers for each time series (RRI and PPI) indicated that the differences were not statistically significant. Additionally, the Kruskal–Wallis test revealed no significant overall differences for each measure between the RRI and PPI time series. These findings suggest that, despite the observed differences in performance, the classifiers do not significantly differ in their performance when applied to the RRI and PPI time series.

3.2. Physiologically-based feature selection method

Figure 4 presents the classification values achieved by the four ML classifiers, taking into account the proposed most informative features from a physiological perspective (CE, HF, LF, MEAN, SDNN). The results of the confusion matrices show that the HUT class regularly gets the highest rating, as previously seen in figure 3. It routinely earns higher True Positive values than the other two states. The latter appears to have been more often wrongly categorised, as shown by the high values given in grey in the confusion matrix of figure 4. The classifier achieving the best performance considering the most physiologically informative features is the kNN classifier which outperforms the others. Nonetheless, the results of statistical tests indicated once again no significant overall differences between classifiers, suggesting that, when considering each metric individually across both time series, the classifiers do not significantly differ in their performance. When comparing values obtained using RRI or PPI time series for each classifier and metric, significant differences ($p < 0.05$) were found for precision when considering the LDA classifier ($p = 0.031$, $*$ in figure 4).

Comparing the results from the AIC feature selection method and the physiologically-based feature selection method, overall it is possible to notice higher accuracy values and F1 scores when using the AIC-FS method. Overall, AIC feature selection appears to enhance the performance of most classifiers on both time series. Also in this case, the Wilcoxon signed-rank test was employed between distributions of values computed using each approach to statistically validate the results and compare the performance of classifiers using the two feature selection methods. When considering the RR time series, significant differences were observed in accuracy for LDA ($p = 0.027$) and RF ($p = 0.039$) (highlighted in red in figure 4) with the



Classifier	Time-series	Accuracy	F1	Recall	Precision
LDA	RRI	75.85	63.89	63.78	64.88 *
	PPI	77.43	66.07	66.14	66.73

Classifier	Time-series	Accuracy	F1	Recall	Precision
SVM	RRI	77.95	66.95	66.93	67.08
	PPI	77.08	65.81	65.62	66.34

Classifier	Time-series	Accuracy	F1	Recall	Precision
kNN	RRI	79.88	69.76	69.82	69.73
	PPI	78.83	68.22	68.24	68.20

Classifier	Time-series	Accuracy	F1	Recall	Precision
RF	RRI	77.25	66.02	65.88	66.30
	PPI	77.60	66.29	66.40	66.68

Figure 4. Confusion matrices, tables of the classification performance results (accuracy, recall, precision and F1 score [%]) using only the features with the most informative physiological meaning (CE, HF, LF, MEAN, SDNN), for each classifier (LDA, SVM, kNN, RF) and each time series (RRI and PPI). Statistical test: *, $p < 0.05$, pairwise Wilcoxon test between the distribution of metrics computed for RRI vs PPI across the 10-fold CV. Red values, $p < 0.05$, unpaired Wilcoxon test between distribution of metrics computed using the AIC-FS vs physiological-based FS across the 10-fold CV.

physiological-based feature selection method showing lower values. Precision demonstrated significant differences for LDA ($p = 0.027$) and RF ($p = 0.004$), again with lower values for the physiologically-based approach, while recall showed no significant differences for any classifier. The F1 score indicated a significant difference for LDA ($p = 0.020$), while no other classifiers showed significant differences in this metric. It is worth noting that concerning the PPI time series, the performance of classifiers using the two feature selection methods did not exhibit statistically significant differences in accuracy, recall, and F1 score.

4. Discussion

This study aimed to comprehensively compare the results of an automatic feature selection method using the Akaike Information Criterion and a physiologically-based feature selection approach in classifying different physiological states from cardiovascular variability time series. The objective was to gain insights into their performance and to identify the most suitable HRV and PRV features (in time-, frequency- and information-theoretic domain) for the best discrimination among physiological states, highlighting the different physiological mechanisms involved during postural and mental stress accompanied by a shift in the SVB.

The feature selection process using the AIC was repeated for each CV fold to ensure a selection specifically tailored to each training set, thus minimizing potential overfitting and enhancing the model's performance. Our results shown in figure 2 indicate that certain features are consistently chosen more frequently than others. Specifically, for the RRI time series features such as 'CE', 'SE', 'HFn' and 'MEAN' demonstrated higher selection frequencies, suggesting their greater relevance in predicting the categorical target variable. Different features like 'SE', 'fHF', 'HFn', 'MEAN' and 'RMSSD' were more commonly selected concerning PPI time series. It is interesting to underline that each set includes features belonging to all three domains, i.e. time, frequency, and information, highlighting the diverse nature of the relevant features for each time series. The different selection of the features suggests that RRI and PPI time series present specific aspects that can be better explained from a peculiar features set.

In the time domain, the MEAN was always selected across the CV folds when considering both RRI and PPI time series. These results are expected, given that the MEAN feature is widely recognized as the basic cardiovascular measure expressing the overall balance between parasympathetic and sympathetic control, as

the SE features of the information domain (Porta *et al* 2016). Moreover, the selection of RMSSD using the PPI time series can be related to the fact that RMSSD is more responsive to the parasympathetic nervous system activity being, in turn, less influenced by respiratory frequency than HF power (Penttilä *et al* 2001, Quintana *et al* 2016). The frequency-domain analysis reveals differential behaviour of the various indices across the two types of time series. For the PPI time series, the fHF and the HFn features are always selected across the 10-folds, while only the HFn is always selected for the RRI time series. This different behaviour is difficult to physiologically interpret, even if we can hypothesize that the selection of fHF for PPI is related to the strong influence of respiration on pre-ejection period (PEP) that affects PPI time series (and not RRI series). It is widely known that information-theoretic metrics reveal the complexity of physiological processes and thus cardiovascular time series, varying across different physio-pathological conditions (Goldberger *et al* 2002, Porta *et al* 2006, 2014). Entropy-based measures, i.e CE and SE, have been shown to respond to different degrees of sympathetic activity during postural stress and can effectively describe variations in the cardiac system activity linked to diverse physiological and clinical conditions (Xiong *et al* 2017, Pernice *et al* 2019). The CE is a measure of complexity that quantifies the dynamics of a time series by assessing the information in the current data point that is not explained by previous points, while the SE is the portion of information which can be derived from its past history (Javorka *et al* 2017). A rise in CE reflects an increase in series complexity, while higher SE indicates a more regular and predictable series (Bari *et al* 2015). Our results demonstrate that both CE and SE indices are consistently selected across all folds for RRI time series, suggesting that complexity measures represent features that can effectively explain differences among physiological states. The same remarks apply to the PPI time series, but with the difference that only SE was consistently chosen.

The difference between HRV and PRV selected indices may be related to the distorting effect of the non-constant PEP, depending mainly on left ventricular contractility and on pulse transit time (PTT), which has also been shown to exhibit physiological variability (Chan *et al* 2007). Respiration may also affect ventricular loading and thus PEP, as well as non-physiological factors like an inaccurate detection of blood pressure maxima (Pernice *et al* 2019). The confounding effect of PEP and PTT may also have a role in the complexity of the PPI time series, which in turn could be responsible for the non-selection of the SE index.

The results of the AIC-FS method are in quite good accordance with the proposed more physiologically-related feature subset (figure 4) which includes the MEAN, the information-theoretical CE, and both frequency-domain LF and HF, being the latter mutually more independent. This suggested subset also contains the SDNN, which conversely did not appear to be a feature selected by the automatic AIC-based method. Consistent feature combinations are identified in the studies of (McDuff *et al* 2014, Tsunoda *et al* 2017, Huang *et al* 2018, Castaldo *et al* 2019, Posada-Quintero and Bolkhovsky 2019, Wang and Guo 2019, Izzah *et al* 2022) which focus on stress and rest detection protocols similar to ours. Nevertheless, they encompass additional features such as pNN50 (i.e. the percentage of differences between duration of successive RRIs > 50 ms), providing avenues for further exploration in subsequent research of our study.

Our results evidence (cfr. figures 3 and 4) that the comparison across the four distinct classifiers reveals no statistically significant differences in their performance. Moreover, the consistency between RRI and PPI outcomes in all except for one metric, alongside the uniformity across various classifiers, underscores the robustness and reliability of the analysis and demonstrates the viability of substituting features derived from HRV by the ones from PRV, confirming previous findings obtained in other works with classical statistical analyses (Schäfer and Vagedes 2013, Hernando *et al* 2018, Pernice *et al* 2019). Moreover, our results (both regarding the AIC-based approach, figure 3 and the physiological-based selection, figure 4) show that when considering all the classifiers and both time series, the class with the best classification performance is HUT, followed by REST. This confirms previous findings evidencing that postural stress is easier to be discriminated than mental workload and the different nature of the two stressors (Pernice *et al* 2019, Pinto *et al* 2022).

Overall, our results suggest that SVM and kNN classifiers produce similar results either using the automatic or the physiologically-based feature selection approaches, while RF and LDA are more feature-dependent, with lower performance metrics when considering physiologically-based features computed from RRI time series. This suggests that automatic feature selection enhances performance for some metrics and classifiers, but not for the PPI-based analysis. Interestingly, while the automatically selected features differ from the physiologically selected ones (except for the MEAN feature), they hold similar meanings. For instance, both CE and SE indicate sympathetic tone, while SDNN and RMSSD reflect vagal control. These findings suggest that incorporating physiological knowledge into ML approaches yields results comparable to blind automatic classification, while also enhancing explainability. The importance of employing a feature selection phase was also emphasized in a previous study (Iovino *et al* 2023a) considering

a smaller version of the dataset and the backward-type Minimal Redundancy Maximal Relevance (mRMR) feature selection algorithm (Ding and Peng 2005). The results are in agreement and evidence that the inclusion of the feature selection phase (either AIC-FS or mRMR) leads to better classification accuracy.

4.1. Limitations

A limitation of our study consists in the fact that the differences between the suggested set of most informative physiologically-based features and the automatic feature selection set could stem from distinctive traits in the subjects, i.e. inter-subject variability of the features, and the relatively low amount of available data. Subsequent research endeavours should involve the application of the identified feature set classification to more 'realistic' datasets, i.e. recorded from real-life or within clinical settings, rather than within controlled experimental environments as in this work, so as to validate the study's potential for practical applications. Moreover, the obtained results from the automatic feature selection could be improved by performing a more sophisticated feature interaction analysis. Additionally, our study specifically evaluates the effects of postural stress induced by head-up tilt and mental stress elicited by mental arithmetic tasks, while further studies should take into account different types of stressors and stress levels to prove the generability of our findings.

5. Conclusions

This study demonstrates that specific stress classification is shown to be feasible, especially for postural stress compared to the rest state, being however more difficult to distinguish between postural and mental stress since both stressors evoke similar sympathetic activation. Our results highlight that the choice of the classifier did not impact classification performances. The automatic AIC-based feature selection method overall achieved slightly better classification than the physiology-driven approach for LDA and RF algorithms. Notably, PPI-based classification, even if with a slightly different set of features, demonstrated similar performance to RRI-based one for almost all metrics and classifiers. In perspective, these results support the feasibility of implementing PPI-based algorithms on wearable devices for outpatient settings monitoring. Additionally, our findings underscore the importance of incorporating physiological knowledge into ML models, enhancing explainability while maintaining robust classification performance. The outcome of our investigation holds potential implications for advancing stress assessment methodologies, not only within clinical settings but also for broader contexts of health monitoring and well-being evaluation. Future activities may envisage the use of feature explainability in larger real-world datasets, to understand the importance of each selected feature. Furthermore, we will explore the application of alternative feature selection algorithms, and consider different training strategies more suitable for lower data amounts, taking into account features extracted from shorter cardiovascular time series, i.e. Ultra Short-Term analysis (Castaldo *et al* 2019, Volpes *et al* 2022).

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

This work was supported by 'DARE—DigitAl lifelong pRevEntion initiative' project (funded by MUR, PNC D.D. 931 06/06/2022, code PNC0000002, CUP: B53C22006460001), by PRIN 2022 project 'HONEST-High-Order Dynamical Networks in Computational Neuroscience and Physiology: an Information-Theoretic Framework' (funded by MUR, code 2022YMHNPY, CUP B53D23003020006), and SiciliAn MicronanOTecH Research And Innovation CEnter 'SAMOTHRACE' (MUR, PNRR-M4C2, ECS 00000022). R P was partially supported by European Social Fund (ESF) Complementary Operational Programme (POC) 2014/2020 of the Sicily Region. I L, T L T and this research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project 'Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad' (No. 01-3394/1). M J is supported by Grant VEGA 1/0283/21.

ORCID iDs

Marta Iovino https://orcid.org/0009-0005-9659-9381
Ivan Lazic https://orcid.org/0000-0001-8613-0049
Tatjana Loncar-Turukalo https://orcid.org/0000-0002-3582-8073

Michal Javorka https://orcid.org/0000-0001-7562-5193
Riccardo Pernice https://orcid.org/0000-0002-9992-3221
Luca Faes https://orcid.org/0000-0002-3271-5348

References

Aggrawal R and Pal S 2020 Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease *SN Comput. Sci.* **1** 344

Ahmad G N, Ullah S, Algethami A, Fatima H and Akhter S M H 2022 Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection *IEEE Access* **10** 23808–28

Awasthi K, Nanda P and Suma K 2020 Performance analysis of machine learning techniques for classification of stress levels using ppg signals *2020 IEEE Int. Conf. on Electronics, Computing and Communication Technologies (CONECCT)* (IEEE) pp 1–6

Bari V, Girardengo G, Marchi A, De Maria B, Brink P A, Crotti L, Schwartz P J and Porta A 2015 A refined multiscale self-entropy approach for the assessment of cardiac control complexity: application to long qt syndrome type 1 patients *Entropy* **17** 7768–85

Benditt D G, Ermis C and Lü F 2004 Chapter 88 - head-up tilt table testing *Cardiac Electrophysiology* 4th edn, ed D P Zipes and J Jalife (W.B. Saunders) pp 812–22

Bishop C M and Nasrabadi N M 2006 *Pattern Recognition and Machine Learning* vol 4 (Springer)

Castaldo R, Montesinos L, Melillo P, James C and Pecchia L 2019 Ultra-short term hrv features as surrogates of short term HRV: a case study on mental stress detection in real life *BMC Med. Inf. Decis. Making* **19** 1–13

Chan G S, Middleton P M, Celler B G, Wang L and Lovell N H 2007 Change in pulse transit time and pre-ejection period during head-up tilt-induced progressive central hypovolaemia *J. Clin. Monit. Comput.* **21** 283–93

Chandrashekar G and Sahin F 2014 A survey on feature selection methods *Comput. Electr. Eng.* **40** 16–28

Dalmeida K M and Masala G L 2021 Hrv features as viable physiological markers for stress detection using wearable devices *Sensors* **21** 2873

Ding C and Peng H 2005 Minimum redundancy feature selection from microarray gene expression data *J. Bioinf. Comput. Biol.* **3** 185–205

Faes L, Masè M, Nollo G, Chon K H and Florian J P 2013 Measuring postural-related changes of spontaneous baroreflex sensitivity after repeated long-duration diving: frequency domain approaches *Auto. Neurosci.* **178** 96–102

Faes L, Porta A and Nollo G 2015 Information decomposition in bivariate systems: theory and application to cardiorespiratory dynamics *Entropy* **17** 277–303

Faes L, Porta A, Nollo G and Javorka M 2016 Information decomposition in multivariate systems: definitions, implementation and application to cardiovascular networks *Entropy* **19** 5

Gerrig R J and Zimbardo P G 2002 *American Psychological Association: Glossary of Psychological Terms* (Pearson Education, Education, Incorporated (COR))

Giannakakis G, Marias K and Tsiknakis M 2019 A stress recognition system using hrv parameters and machine learning techniques *2019 8th Int. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos(ACIIW)* (IEEE) pp 269–72

Goldberger A L, Peng C-K and Lipsitz L A 2002 What is physiologic complexity and how does it change with aging and disease? *Neurobiol. Aging* **23** 23–26

Hernando D, Roca S, Sancho J, Alesanco A and Bailón R 2018 Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects *Sensors* **18** 2619

Huang S, Li J, Zhang P and Zhang W 2018 Detection of mental fatigue state with wearable ECG devices *Int. J. Med. Inf.* **119** 39–46

Iovino M, Javorka M, Faes L and Pernice R 2023 Comparison of machine learning approaches for physiological states classification using heart rate and pulse rate variability indices *Proc. 8th National Congress of Bioengineering* (PàTron Editore) pp 679–82

Iovino M, Lazic I, Loncar-Turukalo T, Javorka M, Pernice R and Faes L 2023 Classification of physiological states through machine learning algorithms applied to ultra-short-term heart rate and pulse rate variability indices on a single-feature basis *Mediterranean Conf. on Medical and Biological Engineering and Computing* (Springer) pp 114–24

Izzah N, Sutarto A P and Hariyadi M 2022 Machine learning models for the cognitive stress detection using heart rate variability signals *J. Tek. Ind.* **24** 83–94

Javorka M, Krohova J, Czippelova B, Turianikova Z, Lazarova Z, Javorka K and Faes L 2017 Basic cardiovascular variability signals: mutual directed interactions explored in the information domain *Physiol. Meas.* **38** 877

Jiménez-Limas M A, Ramírez-Fuentes C A, Tovar-Corona B and Garay-Jiménez L I 2018 Feature selection for stress level classification into a physiologycal signals set *2018 15th Int. Conf. on Electrical Engineering, Computing Science and Automatic Control (CCE)* pp 1–5

Lim H W, Hau Y W, Lim C W and Othman M A 2016 Artificial intelligence classification methods of atrial fibrillation with implementation technology *Comput. Assis. Surgery* **21** 154–61

McDuff D, Gontarek S and Picard R 2014 Remote measurement of cognitive stress via heart rate variability *2014 36th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (IEEE) pp 2957–60

Mejía-Mejía E, May J M, Torres R and Kyriacou P A 2020 Pulse rate variability in cardiovascular health: a review on its applications and relationship with heart rate variability *Physiol. Meas.* **41** 07TR01

Panganiban F C and de Leon F A 2021 Stress detection using smartphone extracted photoplethysmography *2021 IEEE Region 10 Symp. (TENSYMP)* (IEEE) pp 1–7

Penttilä J, Helminen A, Jartti T, Kuusela T, Huikuri H V, Tulppo M P, Coffeng R and Scheinin H 2001 Time domain, geometrical and frequency domain analysis of cardiac vagal outflow: effects of various respiratory patterns *Clin. Physiol.* **21** 365–76

Pernice R, Javorka M, Krohova J, Czippelova B, Turianikova Z, Busacca A and Faes L 2019 Comparison of short-term heart rate variability indexes evaluated through electrocardiographic and continuous blood pressure monitoring *Med. Biol. Eng. Comput.* **57** 1247–63

Pernice R, Sparacino L, Nollo G, Stivala S, Busacca A and Faes L 2021 Comparison of frequency domain measures based on spectral decomposition for spontaneous baroreflex sensitivity assessment after acute myocardial infarction *Biomed. Signal Process. Control* **68** 102680

Pinto H, Pernice R, Silva M E, Javorka M, Faes L and Rocha A P 2022 Multiscale partial information decomposition of dynamic processes with short and long-range correlations: theory and application to cardiovascular control *Physiol. Meas.* **43** 085004

Porta A *et al* 2014 Effect of age on complexity and causality of the cardiovascular control: comparison between model-based and model-free approaches *PLoS One* **9** e89463
Porta A, De Maria B, Bari V, Marchi A and Faes L 2016 Are nonlinear model-free conditional entropy approaches for the assessment of cardiac control complexity superior to the linear model-based one? *IEEE Trans. Biomed. Eng.* **64** 1287–96
Porta A, Guzzetti S, Furlan R, Gnecchi-Ruscone T, Montano N and Malliani A 2006 Complexity and nonlinearity in short-term heart period variability: comparison of methods based on local nonlinear prediction *IEEE Trans. Biomed. Eng.* **54** 94–106
Posada-Quintero H F and Bolkhovsky J B 2019 Machine learning models for the identification of cognitive tasks using autonomic reactions from heart rate variability and electrodermal activity *Behav. Sci.* **9** 45
Quintana D S, Elstad M, Kaufmann T, Brandt C L, Haatveit B, Haram M, Nerhus M, Westlye L T and Andreassen O A 2016 Resting-state high-frequency heart rate variability is related to respiratory frequency in individuals with severe mental illness but not healthy controls *Sci. Rep.* **6** 37212
Rani S, Shelyag S, Karmakar C, Zhu Y, Fossion R, Ellis J, Drummond S and Angelova M 2022 Differentiating acute from chronic insomnia with machine learning from actigraphy time series data *Front. Netw. Physiol.* **2** 1036832
Rinella S, Massimino S, Fallica P, Giacobbe A, Donato N, Coco M, Neri G, Parenti R, Perciavalle V and Conoci S 2022 Emotion recognition: photoplethysmography and electrocardiography in comparison *Biosensors* **12** 811
Scardulla F, Cosoli G, Spinsante S, Poli A, Iadarola G, Pernice R, Busacca A, Pasta S, Scalise L and D'Acquisto L 2023 Photoplethysmograhic sensors, potential and limitations: Is it time for regulation? a comprehensive review *Measurement* **218** 113150
Schäfer A and Vagedes J 2013 How accurate is pulse rate variability as an estimate of heart rate variability?: a review on studies comparing photoplethysmographic technology with an electrocardiogram *Int. J. Cardiol.* **166** 15–29
Shaffer F and Ginsberg J P 2017 An overview of heart rate variability metrics and norms *Front. Public Health* **5** 258
Shah S A 2018 Internet of things for sensing: a case study in the healthcare system *Appl. Sci.* **8** 508
Tsunoda K, Chiba A, Yoshida K, Watanabe T and Mizuno O 2017 Predicting changes in cognitive performance using heart rate variability *IEICE Trans. Inf. Syst.* **100** 2411–9
Umair M, Chalabianloo N, Sas C and Ersoy C 2021 HRV and stress: a mixed-methods approach for comparison of wearable heart rate sensors for biofeedback *IEEE Access* **9** 14005–24
Valente M, Javorka M, Porta A, Bari V, Krohova J, Czippelova B, Turianikova Z, Nollo G and Faes L 2018 Univariate and multivariate conditional entropy measures for the characterization of short-term cardiovascular complexity under physiological stress *Physiol. Meas.* **39** 014002
Vetrov D P, Kropotov D A and Ptashko N O 2009 An efficient method for feature selection in linear regression based on an extended akaike's information criterion *Comput. Math. Math. Phys.* **49** 1972–85
Volpes G, Barà C, Busacca A, Stivala S, Javorka M, Faes L and Pernice R 2022 Feasibility of ultra-short-term analysis of heart rate and systolic arterial pressure variability at rest and during stress via time-domain and entropy-based measures *Sensors* **22** 9149
Wang C and Guo J 2019 A data-driven framework for learners' cognitive load detection using ECG-PPG physiological feature fusion and xgboost classification *Proc. Comput. Sci.* **147** 338–48
Xiong W, Faes L and Ivanov P C 2017 Entropy measures, entropy estimators and their performance in quantifying complex dynamics: effects of artifacts, nonstationarity and long-range correlations *Phys. Rev.* E **95** 062114
Ying X 2019 An overview of overfitting and its solutions *J. Phys.: Conf. Ser.* **1168** 022022