Beat-to-beat ECG Features for Time Resolution Improvements in Stress Detection

Dustin Axman[1,4,†], Joana S. Paiva[2,3,†], Fernando de La Torre[1], Joao P. S. Cunha[2,4]

***Abstract*— In stress sensing, *Window-derived Heart Rate Variability (W-HRV)* methods are by far the most heavily used feature extraction methods. However, these *W-HRV* methods come with a variety of tradeoffs that motivate the development of alternative methods in stress sensing. We compare our method of using *HeartBeat Morphology (HBM)* features for stress sensing to the traditional *W-HRV* method for feature extraction. In order to adequately evaluate these methods we conduct a Trier Social Stress Test (TSST) to elicit stress in a group of 13 firefighters while recording their ECG, actigraphy, and psychological self-assessment measures. We utilize the data from this experiment to analyze both feature extraction methods in terms of computational complexity, detection resolution performance, and event localization performance. Our results show that each method has an ideal niche for its use in stress sensing. *HBM* features tend to be more effective in an online, stress detection context. *W-HRV* shows to be more suitable for offline post processing to determine the exact localization of the stress event.**

I. Introduction

Recent years have seen a surge in the popularity and convenience of devices that collect physiological data [1]. This has led to numerous efforts to use this data for a wide breadth of pertinent classification tasks such as the detection of cardiac arrhythmia, stress, sleep stages, drug use, and emotion [1], [2]. Success in these classification tasks would have enormously broad and beneficial applications in many areas of public health including: preventing car accidents, increasing worker efficiency, mitigating health problems, monitoring drug use more effectively and improving Human-Computer Interaction [2].

Firefighting is one of the careers upon which stress has the largest negative impact [3]. Firefighters are consistently exposed to stressful and fatiguing situations, giving them a higher risk of coronary diseases which account for a large percentage of deaths among these professionals [3]. This makes them prime candidates for stress sensing experiments. In this way, simultaneously analyzing subject's perceived stress levels and physiological signals such as electrocardiogram (ECG) in firefighters, is the first step towards a general stress sensing solution, applicable to all contexts [2], [4]. Since acute stress events induce physiological responses by our cardiovascular and neuroendocrine systems, ECG-derived features both in time and frequency domains have been widely used for stress monitoring and are highly correlated with subject's stress and arousal state changes [4], [5]. Indeed, numerous authors have been exploring the use of ECG features in a human affect context. Most of these groups focus primarily on affect detection using *Window-derived Heart Rate Variability (W-HRV)* features [6], [7]. However, this latter method has drawbacks. While the use of a window allows for a wide range of features to be used, including spectral features, these windows are usually 80 to 300 seconds which deteriorates the temporal resolution and increases the computational complexity of the detector in which such windows are used.

Recently, we have shown that specific *HeartBeat Morphology (HBM)* features based on temporal distances between ECG fiducial points are able to differentiate "stressful" from "non stressful" events in Firefighters (FFs), using a laboratory protocol [4] composed by a stress inducer task - the Trier Social Stress Test (TSST [8]). Considering the drawbacks associated to *W-HRV* features, we decided to compare performance outcomes using *W-HRV* versus *HBM* features. Based on knowledge of the *HBM* extraction process, we hypothesized that the use of this method as opposed to the *W-HRV* method could mitigate some of the drawbacks of *W-HRV* outlined above. In this paper we therefore examined the extent to which these *HBM* features are useful in stress event sensing, by conducting the same laboratory protocol used in our past study [4] among 13 firefighters. In order to evaluate these two methods, we utilized automatic algorithms for stress event detection based on *Machine Learning* techniques. We evaluated not only accuracy, but also time resolution and computational rapidity of each method. Such metrics could be of high importance, since stressful events have been linked to abnormalities in cardiovascular functions (e.g. arrhythmias, cardiomyopathy, etc) in healthy and non-healthy persons [4], making prompt stress detection very desirable.

II. Methods

A. Description of the sample population dataset

A population sample of 13 FFs with a high age variability (3 female, 10 male; age: 31 ± 11 years) from a Portuguese Firefighter unit agreed to participate in this study - see table I. Participants with a history of cardiovascular disease and/or prescription cardiovascular-related drug use were not included in this experiment. This study was approved by the

*This work has been financed by the FCT (Portuguese Foundation for Science and Technology) within the project VR2Market CMUP-ERI/FIA/0031/2013 and PhD Grant PD/BD/135023/2017. It is also funded by the project NanoSTIMA, North Portugal Regional Operational Program (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

† These two authors contribute equally to the work.
[1] Electrical and Computer Engineering, Carnegie Mellon University
[2] INESC TEC - INESC Technology and Science, Porto, Portugal
[3] Astronomy and Physics Department, Sciences Faculty, Porto, Portugal
[4] Faculty of Engineering, University of Porto, Portugal

ISBN 978-0-9928626-7-1 © EURASIP 2017

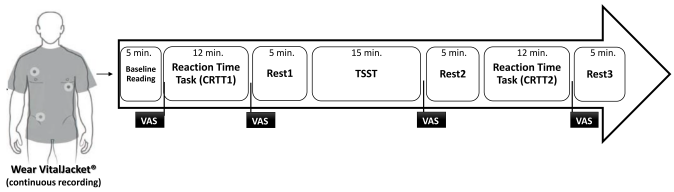

Fig. 1: Diagram of the protocol. VAS - Visual Analogue Scales. TSST - Trier Social Stress Test.

University of Porto Ethics Committee and all the participants signed the corresponding informed consent.

B. Description of the laboratory protocol

The applied laboratory protocol (figure 1) was conducted in a previous study by our laboratory and proved to be a suitable protocol to induce acute stress in FFs [4]. ECG signals were continuously acquired throughout the duration of the experiment ($\approx$ 1 hour) using the VitalJacket® [9] (VJ) at 500 Hz from a single lead. The VJ is a wearable biomonitoring platform (in form of a t-shirt) able to collect ECG signals in a real-time manner, without affecting daily activities of users. It also contains a 3-axis Accelerometer system, allowing ECG signals correction for actigraphy profiles.

The laboratory protocol performed by volunteers was composed of 3 main tasks during which they were comfortably sat in a chair. For evaluating the impact of stress in cognitive performance, a 2-choice reaction time task (CRTT) [10], was conducted. Following this, the Trier Social Stress Test (TSST) [8], a *gold-standard* psychological stress assessment procedure, was applied. After subjects were exposed to the stress condition, they performed again the simple CRTT (CRTT2) described above. Visual Analog Scales (VAS) [11] were used for stress psychological self assessment after each main task (after CRTT1, after TSST and after CRTT2).

C. ECG processing and Features Extraction

Since the primary method for feature extraction in the literature uses a *W-HRV* approach [6], [7], where features are extracted from a fixed-length time interval with each sample representing a different shift of this interval, we compared the accuracy achieved in the proposed classification problem using *W-HRV* versus *HBM*. In this latter approach, each heartbeat waveform is treated as a separate sample. In order to compare the two methods, we created a separate set of labeled samples for each method - see table II.

HBM Features Extraction:

ECG heartbeats acquired during the different stages of the protocol were considered as samples with the temporal metrics extracted from each of these heartbeats as the features that characterize each respective sample of the dataset. ECG heartbeats (dataset samples) were therefore labeled as belonging to a "stressful" or "non-stressful" event according to the protocol stage in which they were acquired. Only the heartbeats that were collected in the TSST portion of the experiment were labeled as in the positive class, as per [4], with all others labeled as being in the the negative class.

A set of *nine* features was extracted from each heartbeat waveform. The features used in this approach were based on temporal distances between fiducial points Q, R, S and T and were extracted using a ECG morphology-based patent pending [12] processing scheme adopted in our previous study [4] - see figure 2. R points were the first fiducials to be located, using the widely known Pan Tompkins algorithm [13]. Considering that existing literature shows that the best method for detecting ECG fiducial points is based on low order polynomial filtering [14], the remaining fiducials - Q, S and T - were located after applying a second order Butterworth low-pass filter with a cut off frequency of 10 Hz to the raw signal. Fiducial points were discovered based on previously established physiological time intervals [15]. The Q points were identified by computing the signal derivative considering a time window of 0.10 seconds before each R point. The last peak within this time window was marked as point Q for each heartbeat. Point S was located by applying a similar method, also based on signal derivatives. The first temporal mark at which the derivative changed from negative to positive values, 0.05 seconds after the R point, was assigned as the point S. For locating the peak of the T wave, it was determined the last temporal index where the derivative of the signal changed from positive to negative values, within a time window of 0.05 to 0.40 seconds after each QRS complex, for each heartbeat.

QR, RT, ST and QRS segments were calculated as depicted in figure 3. RR intervals were defined as the interval between two consecutive R points. The index of the beginning of QT was computed as the last point where

TABLE I: Dataset characterization. HB - heartbeats.

Number of Participants (N)	13
ECG Sampling Rate (Hz)	500
Total Length ECG acquired* (min)	1042
Average Length ECG per subject acquired (min)	80 ± 39
Total Number of HB analyzed*	26313
Total Number of "stressful" HB analyzed*	5550
Total Number of "non-stressful" HB analyzed*	20763

*across subjects

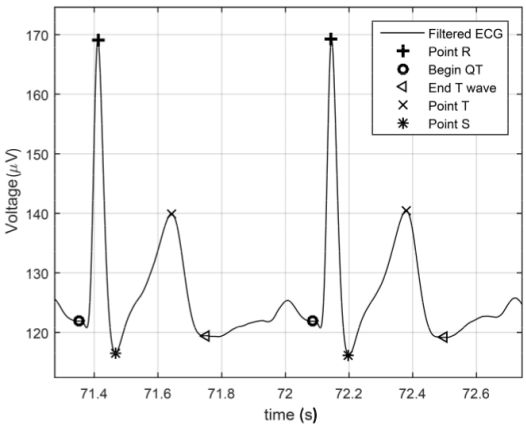

Fig. 2: Portion of ECG from Subject 2, with fiducial points Q, R, S and T; and points that contributed for extracting QT and ST intervals marked.

ISBN 978-0-9928626-7-1 © EURASIP 2017

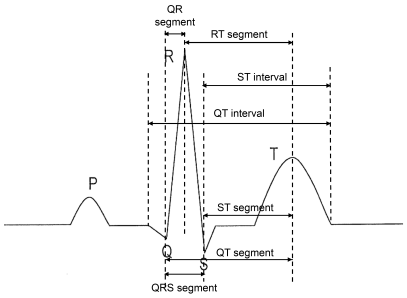

Fig. 3: Sketch of a heartbeat waveform illustrating six of the nine temporal intervals used in the classification task: QR segment; RT segment; ST interval; QT interval; ST segment and QRS segment.

the derivative changed from positive to negative in a time window of 0.03 seconds before each Q point. The end of the T wave was computed as the index corresponding to the last point at which the signal derivative changed from negative to positive values within 0.15 seconds after the T peak. ST_C and QT_C intervals were also included. These are the ST and QT intervals corrected for the interference of heart rate for each heartbeat, using the *Bazett Formula* [16]:

$$QT_C = \frac{QT}{\sqrt{RR}}, ST_C = \frac{ST}{\sqrt{RR}} \quad (1)$$

Through this process the QR, RT, ST, RR and QRS segments, as well as the ST, QT, ST_C and QT_C intervals were calculated for each heartbeat. The QT segment was also initially considered but revealed to be highly correlated to several other features, it was excluded in the classification task so as to reduce multicollinearity, leaving the remaining 9 features used in the *HBM* method - please see table II. Noisy HBs were removed after computing all the nine temporal distance measures, by identifying the HBs which did not satisfy the following conditions [15]:

$$QR \leq 0.075s \quad and \quad 0.200s < QT_C < 0.360s \quad (2)$$

W-HRV Features Extraction:

A fixed window of length 80 seconds was chosen through 10-fold cross validation over our training set with a generic Random Forest [17] classifier. An overlap of 80% was chosen empirically so as to give the same time resolution as the averaged *HBM* features, as described in subsection II-D. Each window was labeled as belonging to a stress event if the majority of the heartbeat waveforms contained in this window were labeled as belonging to a stress event. In accordance with the literature [6], [7], the Lomb-Scargle Periodogram was calculated on the RR-intervals determined with Pan Tompkins algorithm [13] and 6 spectral features were extracted based on the power in each of several bands, described in Table II [6], [7].

In addition to spectral features, we also extracted 5 time-based HRV features (described in Table II) [6], [7].

A total of 11 *W-HRV* features were therefore extracted from each window, contrasting with the 9 features in the *HBM* method (Table II).

D. Classification Task

We trained and tested on the same person in our protocol to evaluate the effectiveness of each method with regard to within subject (rather than between subject) event sensing. This was done by dividing the samples in each subject into 5 equally sized, random groups and using a leave-one-out testing scheme. This was done 5 times such that in the end, every sample in each subject's time series had a score associated with it. These scores were then *un-permuted* so as to rearrange them back into the temporally sequential order in which they had been collected. This gave us a vector of scores for every sample in each subject's time series, in order. Using this score vector and the ground truth vector, we compared *HBM* to *W-HRV* in 5 different standard metrics: Accuracy, Precision, Recall, F1 Score, and the Area Under the Receiver Operating Characteristic curve (AUROC).

Several classifiers were compared for use in evaluating the effectiveness of *HBM* features versus *W-HRV* features. Among these models were: Linear Support Vector Machines (SVM), Kernel Support Vector Machines (K-SVM), K-NN (K-Nearest Neighbor) and Random Forest [18]. We used 5-fold cross validation grid search to find the best parameters for each subject for each model. Number of models in Random Forest was chosen by grid search from 2 to 70 in increments of 2. SVM C parameter grid search was from 10^{-4} to 10^{4} over factors of 10. The K-SVM *sigma* grid search was from 10^{-4} to 10^{4} over factors of 10. K-NN K grid search was from 1 to 20 by increments of 2.

After this was done, the model with the highest average cross validation fold F1-Score for each subject was used for the remainder of our experimentation and evaluation for that respective subject. In every case, the model with the highest performance on the validation set was a Random Forest Classifier, differing in the number of trees used depending on the subject and the method. The number of predictors sampled from on each tree split was the square root of the number of total predictors [17].

III. Results and Discussion

We compared HBM with W-HRV in three main areas:

A. Computational Complexity

The *HBM* features require only a single pass through the ECG signal to detect fiducial points and perform the elementary operations necessary to derive the associated features, then using a linear moving average filter (II-D), making the entire *HBM* method *O(n)* in computation.

For each new shift of the *W-HRV*, 16 seconds are removed from the end of the old window and the next 16 seconds of the time series are added onto the front of the new window. The Lomb Periodogram is derived for the entire new window. The Lomb Periodogram is *O(nlog(n))* [19]. The remaining HRV features are *O(n)* making the entire *W-HRV* method *O(nlog(n)+n)* (O(nlog(n))). This difference in computational complexity suggests different niches in stress event sensing where *HBM* features or *W-HRV* features shine. *W-HRV* features may not be as applicable to online sensing or wearable technology given the need for streamlined computation in these areas. Instead, *W-HRV* may be more suited for offline analysis of stress events.

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE II: Enumeration of the features used in the classification task for each type.

HeartBeat Morphology (HBM) Features [4], [12]		Windowed Heart Rate Variability (W-HRV) Features [6], [7]
1. RR segment 2. QR segment 3. RT segment 4. ST interval 5. ST_C interval 6. QT interval	**Frequency Domain**	1. Spectral power in [0-0.015] Hz 2. Spectral power in [0.015-0.025] Hz band 3. Spectral power in [0.025-0.050] Hz band 4. Spectral power in [0.050-0.120] Hz band 5. Spectral power in [0.120-0.300] Hz band 6. Spectral power in [0.300-0.400] Hz band
7. QT_C interval 8. ST segment 9. QRS segment	**Time Domain**	7. AVNN (average of NN-intervals) 8. SDNN (standard deviation of NN-intervals) 9. rMSSD (square root of the mean squared difference of successive NN intervals) 10. pNN50 (number of pairs of successive NN intervals that differ by more than 50 ms) 11. RMS of the mean of the square of NN intervals

B. Stress Localization

Stress Localization is the process of determining the exact temporal bounds of a stress event. The nature of stress localization inherently gives equal importance to all samples within the stress event. With this in mind, in order to evaluate each method in this area, we used the metrics derived in II-D. The mean over all 13 subjects, for each method, of each of these metrics is shown in Table III.

TABLE III: Average test scores for each method.

	HBM	W-HRV
Accuracy	0.90	0.90
Precision	0.87	0.82
Recall	0.56	0.69
F1 Score	0.64	0.74
AUROC	0.95	0.93

The accuracy for each method is roughly equivalent, which means that the total number of samples correctly classified by each model was almost exactly the same. However, accuracy can be misleading in data like ours, where there is some degree of imbalance between the positive and negative classes.

By evaluating all the performance measures of table III, *W-HRV* shows slight benefits in localization overall. We can see that, while the *HBM* method shows slight improvements in robustness (AUROC) and Precision, the *W-HRV* method has a higher F1 score, indicating that it provides slightly better localization information overall with regard to the stress event. The *W-HRV* method appears to be more effective for post processing the data offline when the goal is to determine the exact beginning, end, and duration of the stress event.

C. Stress Detection

Unlike, Stress Localization, Stress Detection does not aim to determine exact bounds on the support of the event. Instead, it attempts to determine as soon as possible when the event begins, in an online fashion. We use several visualizations of this detection to evaluate how well each method detects the stress event in each subject. Detection in these visualizations was done by iterating sequentially through the score vector (II-D). At each point in the time series our detector triggers if the score of any subsequence seen so far is past a certain score threshold. In this way, for a given threshold value, we can generate the time at which detection of the event would occur in a real world online scenario and compare each method in that way.

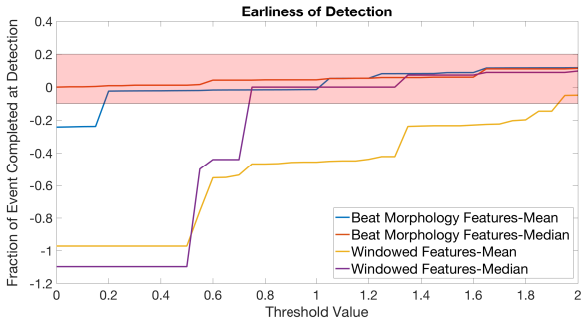

Fig. 4: Mean and median over all subjects of the time of detection as a percentage of the event duration, against the threshold used for detection. The "feasible region" is highlighted at pink.

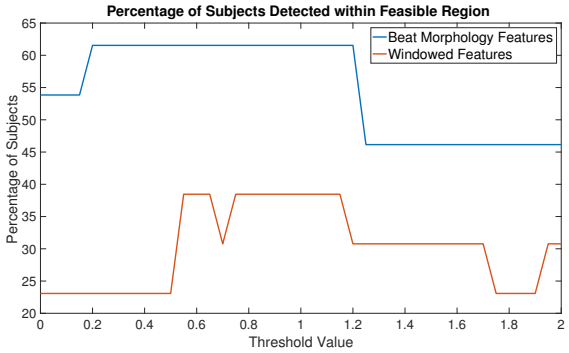

Fig. 5: Comparison between the HBM and the W-HRV features methods in terms of percentage of the subjects, for which we detected the start of the event sometime between -10% and 20% of the event duration.

As we can see in Figure 4, the *HBM* method on average achieves much more timely and accurate detection than the *W-HRV* method for many different threshold values. The highlighted, application-specific "feasible region" shows a time period during which detection must occur for the event detection to be considered a success. For our purposes, we chose -10 to 20% of the event duration from the true start of the event. We can see that especially for lower threshold values, the *W-HRV* method has many false positives that occur far before the event begins. In comparison, the *HBM* method shows to not suffer from this issue to such a great degree, and is far more robust to different threshold choices. The median is also plotted to mitigate the effects from outliers on the visualization. While outliers affect both methods heavily, the *HBM* method is more greatly impacted by outliers, achieving near perfect results in it's median over subjects. We also found that the HBM method was consistently able to achieve detection within the "feasible region" for a larger number of subjects than W-HRV method (Figure 5). The number of subjects for whom detection was

ISBN 978-0-9928626-7-1 © EURASIP 2017

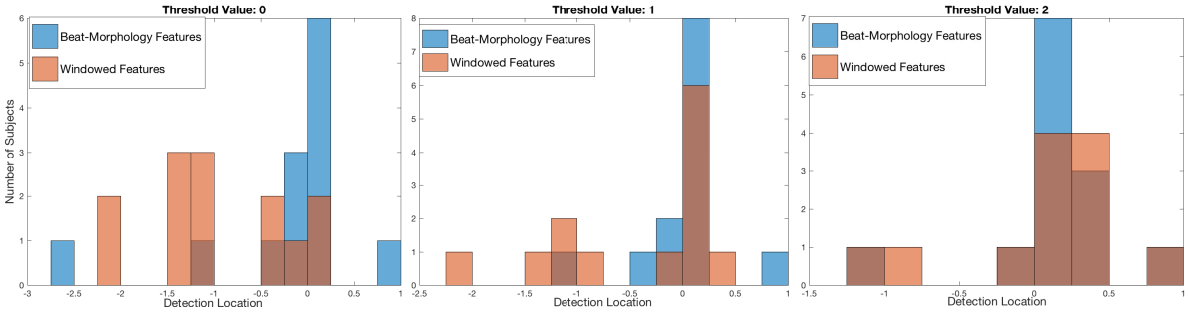

Fig. 6: Histograms showing the number of subjects for whom our detector triggered in each given time range. Three histograms are shown above, each one using a detector set to trigger at 0, 1, and 2 respectively.

achieved in each time period is also shown in Figure 6. It is clear that both methods differ wildly in their distributions, especially for low threshold values. A look at the graph for threshold value of 0 validates our earlier interpretation that the *W-HRV* method seems to have a high false positive early triggering rate whereas the *HBM* method does not seem to suffer from this shortcoming. It is also important to note that these false positives occur far too early for them to be interpreted as a "prediction" of the upcoming stress event.

IV. Conclusions

It is evident from this study that *W-HRV* and *HBM* have specific niches in stress event sensing. *W-HRV* methods have slightly higher "accuracy" (F1 Score), yet they require more computation and do not achieve very good detection results. In comparison, *HBM* methods require far less computation ($(O(n))$ computational complexity) and show excellent results in the area of detection. This makes the *HBM* method a perfect candidate for use in online processing and detection, while *W-HRV* methods are possibly more suitable for offline post-processing of the time series data.

The two methods become more similar, and seemingly more accurate, for higher threshold values. Keep in mind that this does not validate the practice blindly choosing higher thresholds for all applications. In general, lower thresholds yield earlier detection. Therefore, each application should weigh the benefits of *accurate* detection with *early* detection. To provide a fair comparison, we used a roughly equivalent number of features for both and we did not partake in extensive testing of different complex features derivable from the *HBM* features. In the future, we hope to incorporate features to account for temporal dependencies in the *HBM*, while still maintaining the linear time complexity that the current method enjoys. This may also allow us to eliminate the noise-smoothing 16-beat moving average and increase time resolution. Although we do not consider data leakage to have been a prominent problem because of the small number of features, in the future we intend to take measures to further mitigate possible data leakage from time series data. For example, conducting stress-evoking experiments that provide more than one time series from each subject, or the use of domain adaptation methods, would eliminate the need for training and testing within the same time series.

References

[1] J. Hogan and B. Baucom, "Behavioral, affective, and physiological monitoring," *Computer-Assisted and Web-Based Innovations in Psychology, Special Education, and Health*, p. 1, 2016.

[2] J. Cunha, "PHealth and wearable technologies: A permanent challenge," *Stud. Health Technol. Inform*, vol. 177, pp. 185–195, 2012.

[3] L. Rosenstock and J. Olsen, "Firefighting and death from cardiovascular causes," *The New England Journal of Medicine*, 2007.

[4] J. Paiva, S. Rodrigues, and J. Cunha, "Changes in ST, QT and RR ECG intervals during acute stress in firefighters: A pilot study," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 3378–3381.

[5] A. Bhide, R. Durgaprasad, L. Kasala *et al.*, "Electrocardiographic changes during acute mental stress," *International Journal of Medical Science and Public Health (Online First)*, vol. 5, no. 5, 2016.

[6] A. Camm, M. Malik, J. Bigger, G. Breithardt, S. Cerutti, R. Cohen, P. Coumel, E. Fallen, H. Kennedy, R. Kleiger *et al.*, "Heart rate variability. standards of measurement, physiological interpretation, and clinical use," *European heart journal*, vol. 17, no. 3, pp. 354–381, 1996.

[7] A. Voss, R. Schroeder, A. Heitmann, A. Peters, and S. Perz, "Short-term heart rate variability – influence of gender and age in healthy subjects," *PloS one*, vol. 10, no. 3, p. e0118308, 2015.

[8] M. Birkett, "The Trier Social Stress Test protocol for inducing psychological stress," *Journal of visualized experiments: JoVE*, vol. 19, no. 56, 2011.

[9] J. Cunha, B. Cunha, A. Pereira *et al.*, "Vital-jacket®: A wearable wireless vital signs monitor for patients' mobility in cardiology and sports," in *Pervasive Computing Technologies for Healthcare, 2010 4th International Conference on*. IEEE, 2010, pp. 1–2.

[10] J. Paiva, "Predicting lapses in attention: a study of brain oscillations, neural synchrony and eye measures," *MSc Thesis, University of Coimbra*, pp. 33–36, 2014.

[11] F. Lesage, S. Berjot, and F. Deschamps, "Clinical stress assessment using a visual analogue scale," *Occupational medicine*, no. 62, p. 140, 2012.

[12] J. Cunha and J. Paiva, "Biometric Method and Device for Identifying a Person Through an Electrocardiogram (ECG) Waveform - ref PT109357," 2016, pT109357.

[13] J. Pan and W. Tompkins, "A real-time QRS detection algorithm," *Biomedical Engineering, IEEE Transactions on*, vol. 32, no. 3, pp. 230–236, 1985.

[14] S. Israel, J. Irvine, A. Cheng *et al.*, "ECG to identify individuals," *Pattern recognition*, vol. 38, no. 1, pp. 133–142, 2005.

[15] D. Clifford, "ECG statistics, noise, artifacts, and missing data," *Advanced Methods and Tools for ECG Data Analysis*, vol. 6, pp. 55–99, 2006.

[16] H. Bazett, "An analysis of the time-relations of electrocardiograms," *Heart*, vol. 7, pp. 353–370, 1920.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[18] Y. Anzai, *Pattern recognition and machine learning*. Elsevier, 2012.

[19] P. Stoica and N. Sandgren, "Spectral analysis of irregularly-sampled data: Paralleling the regularly-sampled data approaches," *Digit. Signal Process.*, vol. 16, no. 6, pp. 712–734, Nov. 2006.

ISBN 978-0-9928626-7-1 © EURASIP 2017