The North Sea Bicycle Race ECG Project: Time-Domain Analysis

Submitted: 6th December 2017; accepted 5th March 2018

Dominika Długosz, Trygve Eftestøl, Aleksandra Królak, Tomasz Wiktorski, Stein Ørn

DOI: 10.14313/JAMRIS_1-2018/3

Abstract:
Analysis of electrocardiogram and heart rate provides useful information about health condition of a patient. The North Sea Bicycle Race is an annual cycling competition in Norway. Examination of ECG recordings collected from participants of this race may allow defining and evaluating the relationship between physical endurance exercises and heart electrophysiology. Parameters reflecting potentially alarming deviations are to be identified in this study. This paper presents results of a time-domain analysis of ECG data collected in 2014, implementing K-Means clustering. A double stage analysis strategy, aimed at producing hierarchical clusters, is proposed. The first phase allows rough separation of data. Second stage is applied to reveal internal structure of the majority clusters. In both steps, discrepancies driving the separation could stem from three sources. Firstly, they could be signs of abnormalities in electrical activity of the heart. Secondly, they may allow discriminating between natural groups of participants – according to sex, age, physical fitness. Finally, some deviations could result from faults in data extraction, therefore serving in evaluation of the parameters. The clusters were defined predominantly by combinations of features: heartbeat signals correlation, P-wave shape, and RR intervals; none of the features alone was discriminative for all the clusters.

Keywords: *ECG, principal Component Analysis, silhouette analysis, clustering*

1. Introduction

The North Sea Race (Nordsjørittet) is an international cycling competition organized annually in Rogaland, western Norway, between cities: Egersund and Sandness. It is open to a wide spectrum of competitors, from amateurs to professionals. In 2014, ECG data was collected from over a thousand participants, on three days: the day of the race (14.06.2014) as well as the day before and after the race. The data set was collected as part of the North Sea Race Endurance Exercise Study (NEEDED). Continuation of this project with extended set of recorded data is planned for years 2017–2019. Additionally, long-term effects are to be studied for 20 years, until 2034.

Analysis of electrocardiogram (ECG) is a valuable tool in monitoring and diagnosis of patients for various cardiac conditions. The procedure of automatic ECG signal analysis can be performed in time domain or frequency domain and is usually divided into two steps: feature extraction and classifier designation [1]. There are various methods for feature extraction that are reported in the literature. The aspects of Principal Component Analysis (PCA) related to ECG signal processing are discussed in [2], application of customized wavelet transform (WT) in ECG discriminant analysis is described in [3], while the use of Hilbert transform for feature extraction from ECG signal was examined in [4]. Comparison of support vector machine (SVM) algorithm and artificial neural network approach (ANN) for classification of arrhythmias in ECG signal is presented in [5]. Deep learning method for active classification of electrocardiogram signals was applied in the research described in [6], while the clustering method for QRS complexes classification was applied in [7].

Measurement of ECG and heart rate (HR) during daily activity is a potential tool for early diagnosis of cardiac diseases and may also provide individualized guidance to exercise and physical training. The aim of this project is to identify ECG and HR parameters useful for differentiating normal and abnormal patterns during prolonged, high intensity endurance exercise. In this part of the study, concerning data from 2014, the objective consisted of three elements. First of all, it aimed at creating ECG processing algorithms which would found a base for future analysis. Particular focus was put on time-domain approaches. Secondly, influence of a major physical effort on electrical activity of the heart was studied. Finally, by means of data clustering algorithms, the project aimed at developing methods to detect possible individuals with ECG parameters significantly different than for most of the participants.

2. The Dataset and Software

The database consisted of 3158 ECG recordings, each of duration of 10 s, stored in .mat files. Since this project aimed at comparison of data obtained from all three collection time points, it was decided to reject participants for whom some of the recordings were missing. As a result, 996 complete sets of three recordings were obtained.

The collection had to be further reduced owing to errors raised in a few cases on the stage of ECG segmentation. After removing these, further analysis was conducted for 989 participants (2967 ECG recordings).

Processing and analysis of the data was conducted using Python programming language, with particular use of packages: BioSPPy [8], SciPy [9], and scikit-learn [10], [11].

3. Data Pre-processing and Feature Extraction

The dataset provided 8-channel ECG recordings, containing signals from leads I, II, and six precordial leads (V1 to V6). In this project, however, only lead-I signal was analyzed. After the channel of interest was extracted, it was subjected to pre-processing and measurements, described in detail in the following sections of this paper. The procedure aimed at visualization of changes in the ECG signal over the three days and extraction of features relevant for comparison of data obtained from different participants.

3.1. Data Pre-processing

In the initial stage of processing, the lead-I ECG signal was subjected to filtering to suppress high-frequency noise and remove baseline drift. This was done by application of a bandpass-type Finite Impulse Response (FIR) filter with cutoff frequencies of 3 and 45 Hz. The filtered signal was used to detect locations of R-peaks, which was done by Engelse-Zeelenberg approach modified by Lourenco *et al.* [12]. As a proofreading, for singular cases in which this method failed to reliably identify the peaks (less than 3 of them found in a ten-seconds recording), the detection was repeated utilizing the method of Christov [13]. The identified R-peaks were used as reference during extraction of heartbeat templates, defined in a time window of 0.3 s before and 0.4 s after the spike. For both procedures, algorithms implemented in the BioSPPy package [8] were used.

The pre-processing stage was finalized by averaging of the heartbeat templates extracted from a single recording to improve signal-to-noise ratio [14]. Additionally, parameters referring to the heart rate (mean duration and standard deviation of R-to-R intervals) were derived.

3.2. Heartbeat Template Measurements

Some of the features used in the further processing stage were defined on the basis of characteristic intervals and amplitudes of waveforms present in a standard lead-I ECG signal. In order to measure those, methods for searching key points (peaks of P, Q, R, S, and T waves, as well as onsets and endpoints of some of them) in the heartbeat templates were developed.

The location of R-peak in the heartbeat signal was fixed, resulting from the beat extraction procedure. P wave top was defined as a maximum before the occurrence of the R-peak, excluding 0.05 s directly preceding the latter. A similar, but mirror-reflected procedure was applied for determining the top of the T wave. The Q and S points were found as local minima within a fixed, short time window before and after the R-peak respectively. The S wave endpoint, needed mainly for the purpose of ST elevation measurements, was defined as a point where the positive slope after S falls below 90% of its value at S.

Search for onsets and endpoints of P and T waves was performed following the idea described by Laguna et al. [15]. In a specific time window preceding or following the wave peak of interest (for an onset or an endpoint of the wave respectively), a point with a maximal slope is found. Moving further away from the peak, the algorithm searches for a point at which the slope attains a value of the slope specified by a threshold. The threshold is defined as a percentage (e.g. 2%) of the maximum slope value before or after the peak. In absence of such a point, a point with minimal slope within the time window (taken from the maximal slope point) is marked as the onset or endpoint. The values of thresholds and time window durations were adjusted empirically.

Exemplary results of the ECG key point search are presented in Fig. 1. Each subplot presents an averaged heartbeat template for the respective day of measurements for the same participant. The found ECG points are marked as red dots.

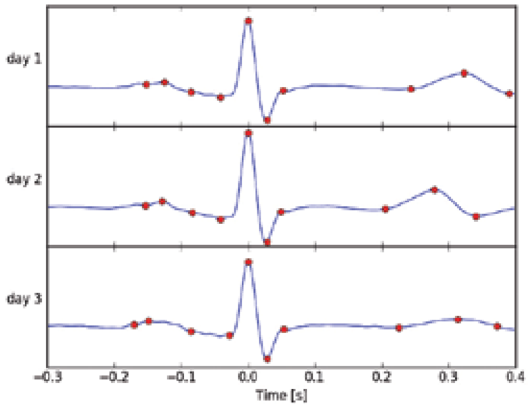

Fig. 1. ECG key points detection – exemplary results

The points were used to measure intervals and amplitudes of ECG signals. For estimation of amplitudes, the level of Q was regarded as the baseline. ST elevation was defined as difference in amplitude between the endpoint of the S wave and the onset of the T wave.

3.3. Morphological Comparison of Heartbeat Templates

Another set of parameters was derived from comparison of morphology of the extracted heartbeats, either a full set of beats from one signal or a set of three averaged beats from the three days (for a given participant). To exclude correlation changes stemming from changing heart rate between the days (which influences durations i.a. of ST interval), processing in this part was conducted only on parts of the heartbeats corresponding to QRS complexes, whose shape did not exhibit any heart-rate dependency.

A basic measure to compare the heartbeats is Pearson r coefficient, also referred to as Pearson product-moment correlation coefficient. Its value was computed for every pair of heartbeats within the analyzed set, creating a matrix of beat-to-beat correlation. To ensure that exclusively the shape of the beats is compared, with no influence of residual baseline drift, the coefficient was calculated using first differences of the signals. From the correlation matrix, the mean value was used as a feature for the further analysis.

Another aspect in beat contour analysis is the idea of morphological classification. It was developed at the University of Glasgow, as a part of their 12-lead ECG analysis algorithm [16], [17]. Following this approach, QRS complexes from the first day were iteratively compared taking into account their morphology (Pearson r coefficient). Similar peaks were grouped into a class; if similarity threshold was exceeded, a new class was created. Beats within each class were averaged to serve as templates for comparison with signals from the second and third day. Beats from these two days were assigned to this of the first-day classes to which they were the most similar. In case Pearson coefficient for a beat and each of the classes templates was falling below a specific threshold, the beat was considered an outlier. Percentage of such morphological outliers for the given participant was another feature derived in this field.

3.4. Features Definition

ECG features were derived from the measurements using the above described approaches. Ten features aimed at comparing the data obtained from the three days were defined as described below. Abbreviations of the feature names, provided in the parentheses, are used later in figures presented in the results section.

- Shape coefficient of P wave, defined as ratio of height of the wave to its width; the used features expressed change in this value from day 1 to day 2 or 3 (P_shape_12 and P_shape_13 respectively).
- Difference in duration of QT interval on day two or three with respect to day 1 (QT_12 and QT_13 respectively).
- Difference in duration of RR interval on day two or three with respect to day 1 (RR_12 and RR_13 respectively).
- Change (difference) in mean correlation of heartbeat templates from the second or third recording with respect to correlation in the first day (correlation_12 and correlation_13 respectively).
- Maximal ST elevation (max_ST_elev) – maximum from values measured on the three days. It was decided to choose the maximum coming from any of the days since the ST elevation itself, not necessarily its change from day to day, should be regarded as an alarming ECG feature. [18]
- Percentage of morphological outliers (morph_outliers) – percentage of beats from days no. 2 and 3 not matching to any beat class defined in day 1 for the given participant (expressed with relation to total number of beats from the three days), as defined in the previous section.

Features based on differences between days are defined by subtracting value on day 2 or 3 from value on day 1. Therefore, positive values of these features indicate a decrease with respect to day 1 (shortening of intervals or decline in correlation).

4. Feature Set Analysis

Analysis of the derived set of features was performed predominantly by unsupervised clustering. Since it was noticed that clustering on the entire dataset tends to yield one or more larger clusters, containing majority of the points, and a few ‘far outliers’ – points significantly separated from the majority group, it was decided to develop a two-stage procedure. After first-attempt analysis and clustering, the outliers clusters (containing less than 10% of the total number of observations) are removed and the analysis is repeated to reveal structure of the majority clusters.

Each of the two stages consists of two main elements: principal component analysis (PCA) and K-means clustering combined with silhouette analysis, described in the following sections of this paper.

4.1. Principal Component Analysis

Principal Component Analysis is a statistical operation aimed at reduction of dimensionality of the clustering data. It performs mapping of the observation matrix on a new orthogonal space, whose axes are referred to as principal components (PCs). The orientation of the new space is chosen such that the first principal component is aligned with the direction of the highest possible variance in the data; the same applies then to each consecutive principal component, with the assumption, that the new PC is orthogonal to all the previously defined ones. Consequently, each PC explains smaller portion of the dataset variance, expressed as eigenvalue of each component. It is then possible to reduce the dimensionality by discarding the less meaningful principal components and retaining only the first few, which in total stand for majority (e.g. 80%) of the data variance. [19] PCA is frequently applied prior to K-means clustering. It allows not only reducing computational effort by decreasing number of dimensions to be analyzed, but also suppressing the effect of possible correlation between the original features (which is referred to as whitening [20]). Furthermore, by investigation of eigenvectors of the components it is possible to evaluate contribution of each of the original features to the principal components, hence defining their statistical significance.

PCA was applied to the set of features on both main stages of the analysis after data normalization. Six principal components, explaining about 80% of the data variance, were retained. The number of the PCs was chosen such that a balance was reached between dimensionality reduction and the retained portion of the variance. The data mapped on the PC space was passed to clustering and silhouette analysis.

4.2. Clustering with Silhouette Analysis

Since no prior assumptions on the structure of the data were made, and the K-means clustering requires specified number of clusters as an input, silhouette analysis was launched on the dataset to determine the best number of clusters. Silhouette analysis allows validating consistency of computed clusters by comparing cohesion of each sample (describing how well it belongs to a cluster it was assigned to) and its separation from other clusters. The resulting silhouette score is expressed as a fraction between –1 and 1. A high score represents good sample classification, whereas negative values indicate that the sample might have been assigned to an improper cluster. Average silhouette score of all the samples allows assessing general consistency and validity of the clustering. [21]

In this project, silhouette analysis on the PC-transformed data was performed for numbers of clusters (computed by K-means algorithm) ranging from 2 to 7. Average silhouette scores were compared and the number of clusters corresponding to the highest score (the best cluster separation) was chosen for further analysis.

K-means clustering with the chosen number of clusters was applied to the dataset mapped to the reduced principal components space. The result was presented and analyzed graphically both in the PC and the original feature space.

4.3. Feature Set Analysis Framework

In this section, methods of the feature set analysis are summarized and detailed sequence of operations on the dataset is presented.

(a) The 10-dimensional set of features is first subjected to normalization. (b) PCA is performed to map the set to a reduced, 6-dimensional space. (c) Number of clusters is chosen by the silhouette analysis. (d) K-means clustering is applied to the PC-transformed dataset. (e) The results of clustering are presented in both feature spaces. Additionally, eigenvectors and eigenvalues are visualized to analyze statistical significance of the original features. (f) If any of the clusters contains less than 10% of all observations, the corresponding samples are removed from the original observations matrix (with all the 10 features retained). (g) Steps a-e are repeated for the corrected observations set.

5. Results and Discussion

Results of the first-stage clustering analysis are presented in Fig. 2 and Fig. 3, and for the second stage – in Fig. 4 and Fig. 5.

Fig. 6 depicts outcome of PCA. As it can be seen in the figures, 2D presentation of the results provides only a limited view and it is necessary to look at different combinations of the dimensions to observe separation between clusters.

The results of clustering in the PC space and original feature space are presented using scatter plot of observations in two of the feature space dimensions (as shown in Fig. 2 to Fig. 5). It should be noted that the features have been normalized, therefore the exact displayed values should not be taken into account. The results of PCA are shown as bar plots of the eigenvectors of the components (Fig. 6). Starting from the top, the subplots refer to consecutive principal components. Statistical significance of the latter, defined as portion of the dataset variance they explain, is added to the vertical label of each subplot (marked as ExpVar). Heights of bars in the subplots correspond to contribution of the original, normalized features (whose names are listed at the bottom of the plots) to the principal components.

As shown in Fig. 2, the first stage of the analysis produced expected unbalanced results: the majority (more than 90%) of observations was assigned to a single cluster (labeled as 1), while the remaining two clusters are much smaller. As presented in Fig. 2a, cluster 2 is well separated from the other two with respect to the fourth and fifth principal component, which are defined predominantly by percentage of morphological outliers and ST elevation (Fig. 6a). Indeed, this separation is explained predominantly by the first of them – as presented in Fig. 3a, cluster 2 is composed of the observations with relatively high values for morphological outliers percentage, while for most observations the values are equal or close to 0. On the other hand, cluster 0 in this projection is overlapped partially with both clusters 1 and 2. However, it is clearly separated when observed from principal components 1 and 3, both of which exhibit

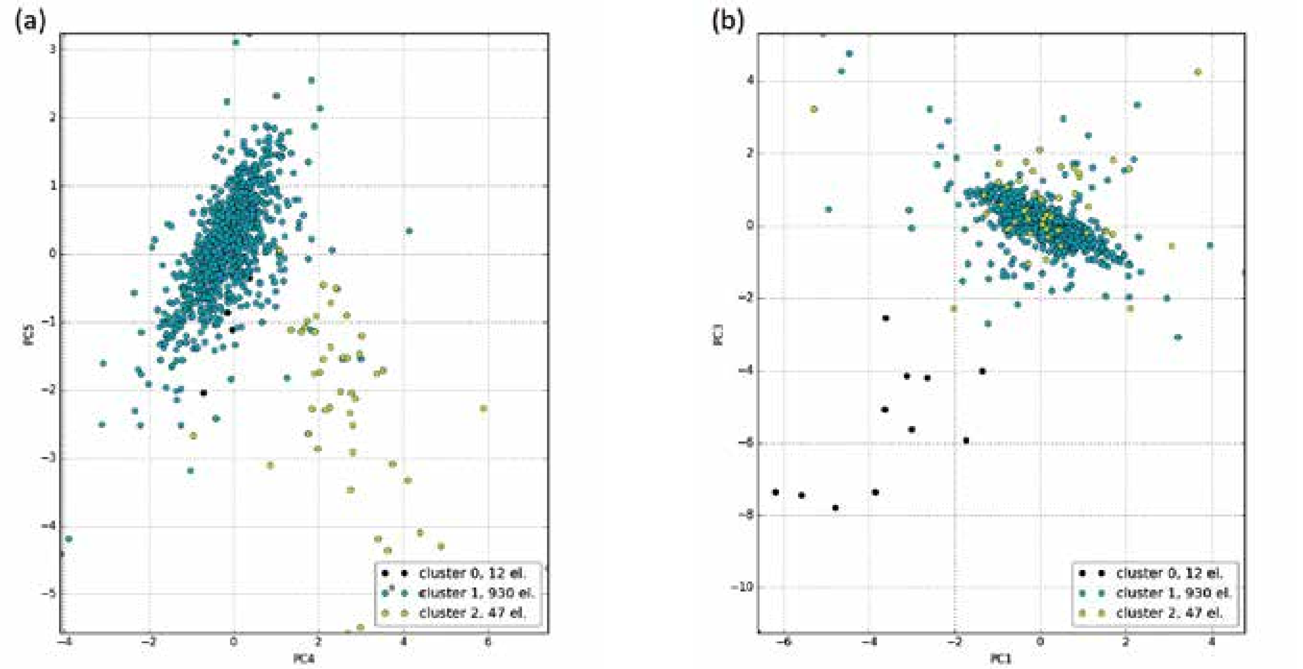

Fig. 2. Result of clustering on the full dataset, in the principal component space; (a) projection on principal components 4 and 5; (b) projection on principal components 2 and 3

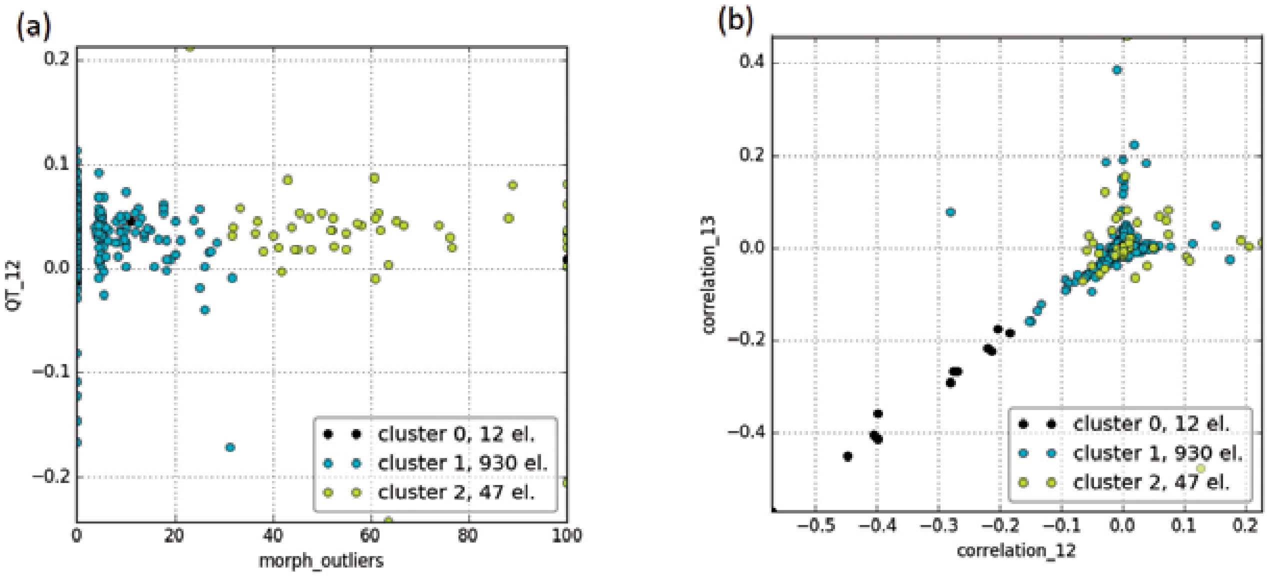

Fig. 3. Result of clustering on the full dataset, in the original feature space; (a) projection QT interval difference (days 1 and 2) and percentage of morphological outliers; (b) projection on correlation difference-related features

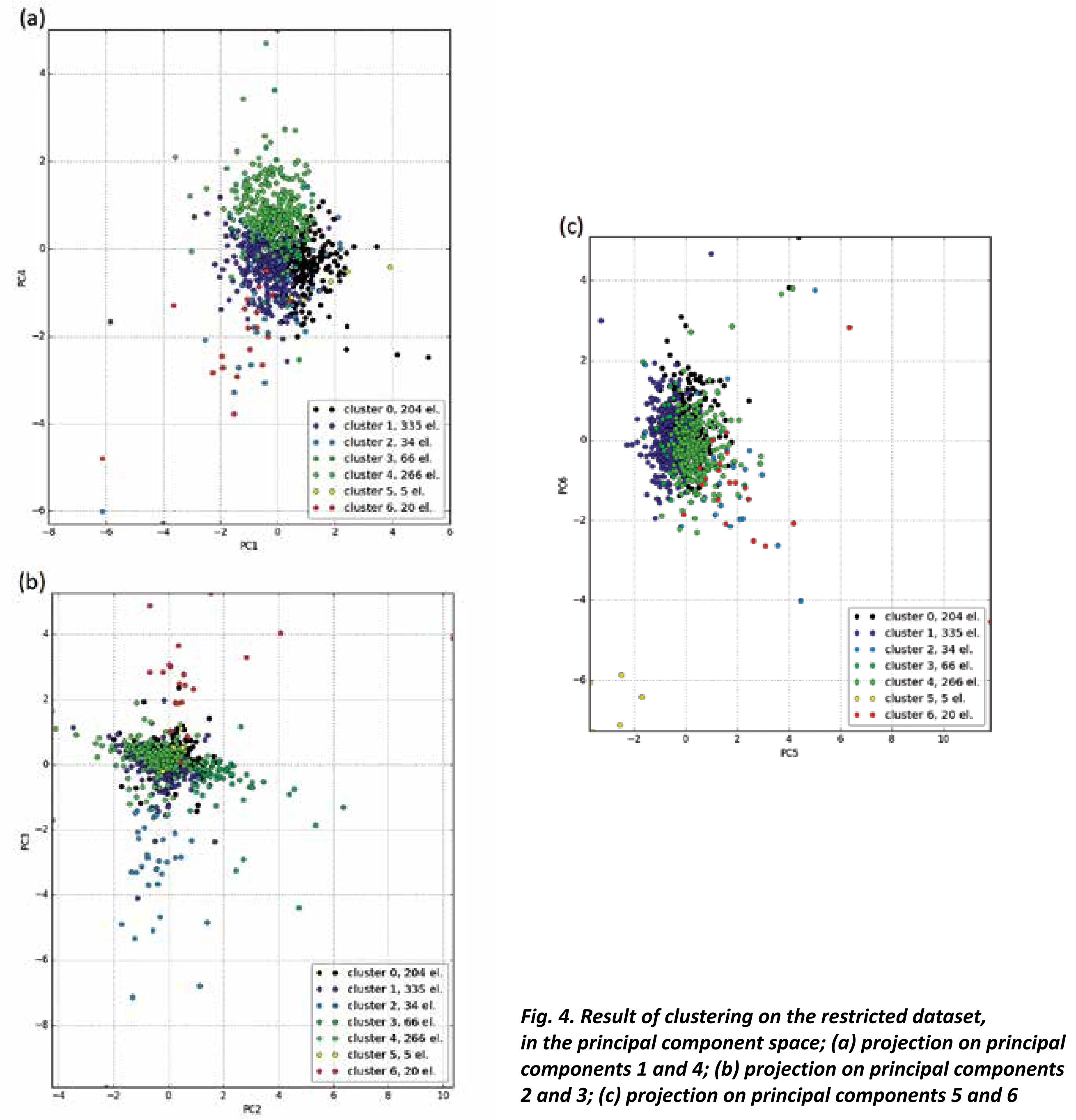

Fig. 4. Result of clustering on the restricted dataset, in the principal component space; (a) projection on principal components 1 and 4; (b) projection on principal components 2 and 3; (c) projection on principal components 5 and 6

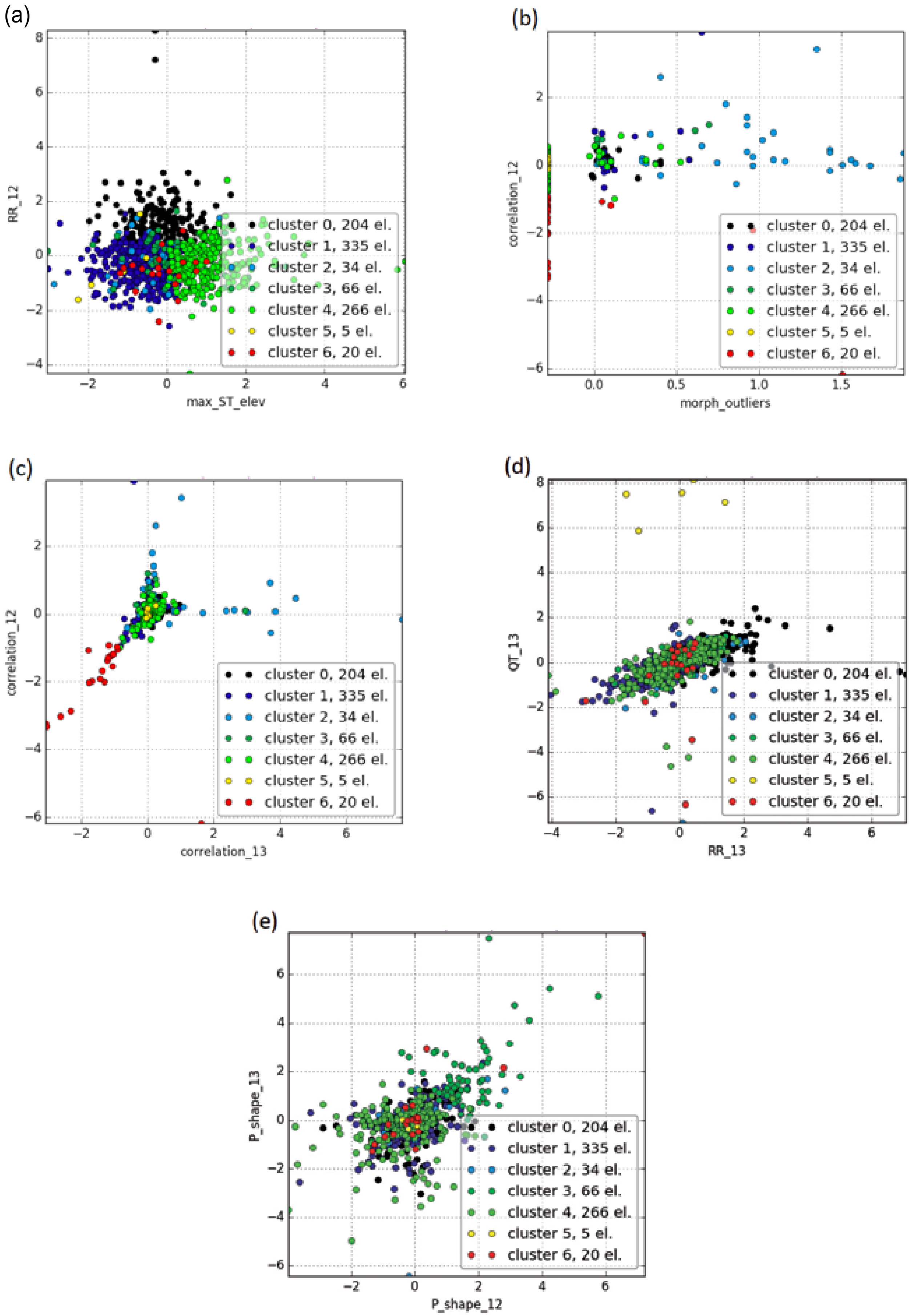

Fig. 5. Result of clustering on the restricted da taset, in the original feature space; (a) projection on RR interval difference (days 1 and 2) and maximal ST elevation; (b) projection on correlation difference between days 1 and 2 and percentage of morphological outliers; (c) projection on the correlation difference-related features; (d) projection on differences in QT and RR intervals between days 1 and 3; (e) projection on the P-shape-related features

high dependence on correlation-related features (see Fig. 3b). Furthermore, analysis of projection of the dataset onto these two features leads to interesting observations. Majority of the points are concentrated around the (0,0) point, indicating little change in intra-recording heartbeat templates correlation from the first to the second or the third day. In some cases, the correlation is reduced with respect to the first day. However, for some participants, the correlation was considerably increased by approximately the same portion on both the second and the third day. The latter group constitutes cluster 0. Hence, for participants in this cluster, correlation on the second and third day was on comparable level, relatively high compared to day 1. This is typically not accompanied by increased percentage of morphological outliers since this feature always uses day 1 as a reference.

Since the clusters 0 and 2 encompassed minor portion of the observations (1.2% and 4.8% respectively), they were excluded from further analysis and the second stage of the procedure was conducted on the points originally assigned to cluster 1, which is presented in Fig. 4 and Fig. 5. Due to a high number of clusters (7), proper visualization of separation in just two dimensions is further obstructed. The three major clusters, labeled as 0, 1, and 4, can be discriminated by looking i.a. at principal components 1 and 4 (Fig. 4a), which are dependent on maximal ST elevation and features related to QT and RR interval (Fig. 5a). However, the separation cannot be clearly visualized in just two dimensions. Possibly, this division is of lesser significance when compared to other clusters distinguished in this set.

Clusters 2, 3, and 6, can be distinguished by projection onto principal components 2 and 3, defined predominantly by features associated with shape of the P wave, correlation, and morphological outliers percentage (Fig. 6b and Fig. 4b). Statistical significance of the latter was slightly lower than in the first stage of the analysis (considering its contribution to the first two principal components); however, it is still one of main components differentiating cluster 2 from others (as shown in Fig. 5b). This is particularly interesting when compared to correlation representation of the clustering result (Fig. 5c). Cluster 2 is constituted by points for which decreased correlation was indeed observed, but predominantly either on day 2 or 3, rarely on both days. On the other hand, cluster 6 exhibits improved correlation on both day 2 and 3. Similarly as in the first stage of the analysis, this does not necessarily entail an increase in the percentage of the morphological outliers.

On the other hand, closer look at the P shape allows to discriminate cluster 3 (as shown in Fig. 5e). For participants belonging to this cluster, P wave was flattened (lower height-to-width ratio) in days 2 and 3 with respect to day 1. The change in shape was more prominent than observed in the other groups.

Finally, cluster 5, which appears to overlap with other clusters in most of dimensions, is in fact distinctly separated with respect to principal components 5 and 6 (Fig. 4c). The fact that it was not reflected in any of the first, more important principal components could be attributed to relatively small size of this cluster (about 0.5% of all observations), which diminishes its impact on the total variance of the dataset. Original features that contribute the most to this component include those related with QT and RR intervals. As presented in Fig. 5d, decrease in duration of QT interval is in general correlated with increase in RR interval. For cluster 5, however, this trend does not apply. Values for RR interval overlap with other clusters, but QT interval on day 3 is shortened to a much higher extent. This effect is not present on day 2. Further detailed investigation of these cases is needed to determine whether the phenomenon is a question of improper key point localization or a sign of potential cardiac issue.

Although the identified clusters are usually not distinctly separated from one another, they are defined by common trends in relation to combinations of certain features. Summary of the clustering procedure and results is presented in Fig. 7 in the Appendix.

6. Conclusions

The NEEDED study focuses on characterization patterns associated with a prolonged endurance exercise. One of its major goals is identification of parameters related to ECG and heart rate which could be used to distinguish between regular and deviated performance of the heart. In this part of the research, several potentially discriminative features were recognized. Further investigation and validation with additional data is needed to verify which of them could serve as criterions in detection of electrocardiophysiological abnormalities.

In the first stage of the analysis, the crucial features were associated predominantly with correlation between the beats. The impact of the correlation-related features was slightly diminished, but still considerable during the second stage of the clustering. The other particularly meaningful features included: P shape, RR interval and QT interval, the latter two exhibiting some correlation. Interestingly, the heart rate (described by RR interval) was not always increased after the race; frequently, the direction of the change was the same on day 2 and 3 with respect to day 1. More valuable information could be extracted by comparison of these trends with additional data, including participant details (sex, age, level of physical activity) as well as data on time interval between finishing the race and collecting the recording of the individual participant.

It should be noted that there was no universal feature or principal component which would provide separation between all the clusters globally. On the other hand, each cluster could be described by a combination of two to four features that made it distinguishable from the other clusters. Determination of features defining the individual clusters was facilitated by analysis of eigenvectors of the principal components. However, PCA is only based on variance prominent in one of the first principal components, what makes it only a candidate for a cluster-determining property. On the other hand, features truly significant for separation are always marked in the principal components' eigenvectors.

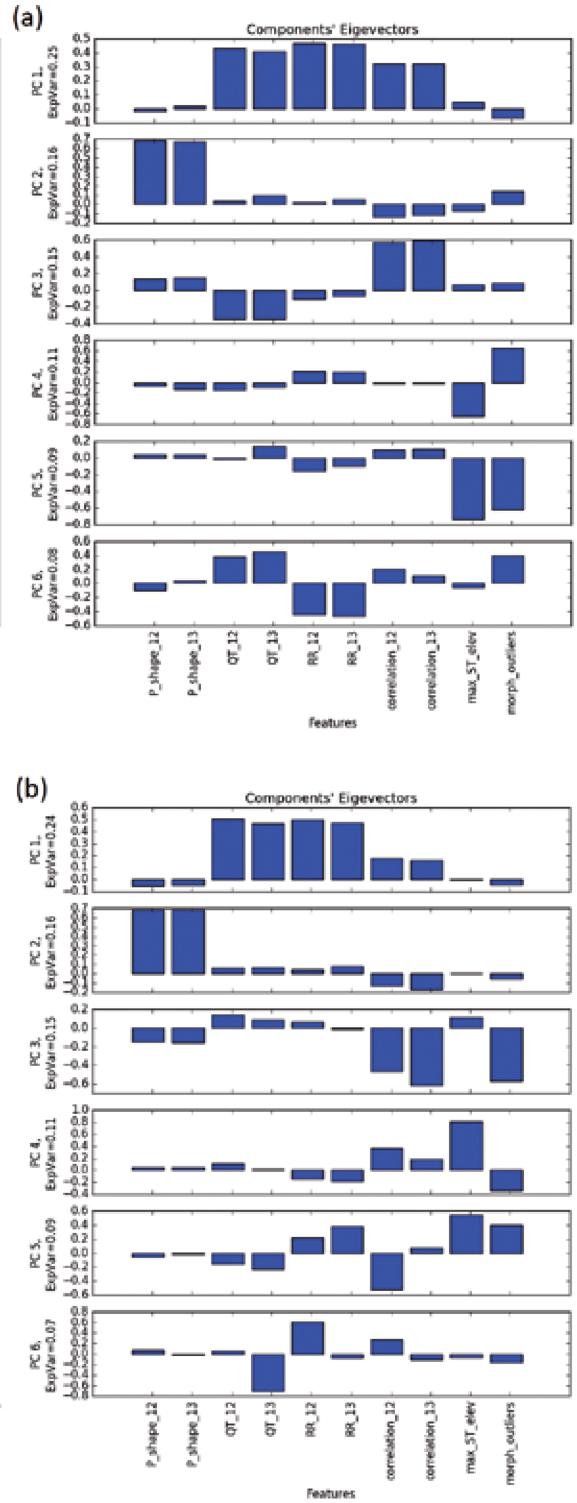

Fig. 6. Principal component analysis results: eigenvectors and explained variance portions of the six components; results of the first (a) and second (b) stage of the analysis

The presented method produces hierarchical structure of clusters from the dataset. This allows two-level investigation of the data structure and separate investigation of huge discrepancies and more subtle trends in the dataset. Furthermore, the hierarchy scheme is also followed in analysis of features having particular impact on the dataset partitioning. Combined with additional data, it could be used in differentiation between natural, physiological groups among the population and early detection of certain cardiac abnormalities.

It should be noted that the produced model of the analyzed dataset well suits the expected structure of the test population. Participants with deviating ECG parameters constitute a minority. Most of observations fall into the normal ranges or exhibit only slight alterations of different types, reflecting physiological phenomena with ontogenetic variability.

Future works include a fusion of time-domain and frequency-domain analysis of the collected ECG data. Furthermore, the dataset will be supplemented with additional information, including i.a. patients' age, gender, the race completion time, and indication of cardiovascular system condition. This will allow to verify the results concerning significance of the ECG features derived and investigated in this paper. What is more, the supplementary data will enable introducing supervised learning methods to the analysis. The algorithm will be trained to eventually gain the ability of differentiating between natural groups of participants and reporting possible cases of alarming ECG parameters. Additional analysis will be launched for a set of competitors participating in more than one edition of the race to study long-term influence of endurance effort on cardiac physiology in professionals and amateurs.

AUTHOR

Dominika Długosz – Łódź University of Technology, Institute of Electronics, ul. Wólczańska 211/215, 90-924 Łódź, Poland, e-mail: 195887@edu.p.lodz.pl
Trygve Eftestøl* – University of Stavanger, Faculty of Science and Technology, Department of Electrical and Computer Engineering, 4036 Stavanger, Norway, e-mail: trygve.eftestol@uis.no
Aleksandra Królak* – Łódź University of Technology, Institute of Electronics, ul. Wólczańska 211/215, 90-924 Łódź, Poland, e-mail: aleksandra.krolak@p.lodz.pl
Tomasz Wiktorski – University of Stavanger, Faculty of Science and Technology, Department of Electrical and Computer Engineering, 4036 Stavanger, Norway, e-mail: tomasz.wiktorski@uis.no
Stein Ørn – University of Stavanger, Faculty of Science and Technology, Department of Electrical and Computer Engineering, 4036 Stavanger, Norway, e-mail: stein.orn@uis.no

*Corresponding author

REFERENCES

[1] X. Dong, C. Wang, W. Si, "ECG beat classification via deterministic learning", *Neurocomputing*, vol. 240, May 2017, 1–12. DOI: 10.1016/j.neucom.2017.02.056.

[2] F. Castells, P. Laguna, L. Sornmo, A. Bollmann, J. Roig, "Principal component analysis in ECG signal processing", *EURASIP J. Adv. Signal Process.*, 2007. DOI: 10.1155/2007/74580.

[3] A. Daamouche, L. Hamami, N. Alajlan, F. Melgani, "A wavelet optimization approach for ECG signal classification", *Biomed. Signal Process. Control*, vol. 7, 342–349, Jul. 2012. DOI: 10.1016/j.bspc.2011.07.001.

[4] D. Benitez, P. Gaydecki, A. Zaidi, A. P. Fitzpatrick, "The use of the Hilbert transform in ECG signal analysis", *Comput. Biol. Med.*, vol. 31, no. 5, 399–406, 2001. DOI: 10.1016/S0010-4825(01)00009-9.

[5] M. Moavenian, H. Khorrami, "A qualitative comparison of Artificial Neural Networks and Support Vector Machines in ECG arrhythmias classification", *EXPERT Syst. Appl.*, vol. 37, no. 4, Apr. 2010, 3088–3093. DOI: 10.1016/j.eswa.2009.09.021.

[6] M. M. A. Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, R. Yager, "Deep learning approach for active classification of electrocardiogram signals", *Inf. Sci.*, vol. 345, Jun. 2016, 340–354. DOI: 10.1016/j.ins.2016.01.082.

[7] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo, "Clustering ECG complexes using Hermite functions and self-organizing maps", *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, 838–848, Jul. 2000. DOI: 10.1109/10.846677.

[8] "biosppy.signals — BioSPPy 0.2.2 documentation" [Online] Available: http://biosppy.readthedocs.io/en/stable/biosppy.signals.html#biosppy-signals-ecg. [Accessed: 16-Jul-2016].

[9] "Documentation — SciPy.org" [Online]. Available: https://www.scipy.org/docs.html. [Accessed: 29-Apr-2017].

[10] "scikit-learn: machine learning in Python — scikit-learn 0.18.1 documentation" [Online]. Available: http://scikit-learn.org/stable/. [Accessed: 29-Apr-2017].

[11] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python", *J. Mach. Learn. Res.*, vol. 12, Oct. 2011, 2825–2830. DOI: 10.1016/j.patcog.2011.04.006.

[12] A. Lourenço, H. Silva, P. Leite, R. Lourenco, A. Fred, "Real Time Electrocardiogram Segmentation for Finger based ECG Biometrics (PDF) – Semantic Scholar". [Online]. Available: https://www.semanticscholar.org/paper/Real-Time-Electrocardiogram-Segmentation-for-Louren%C3%A7o-Silva/358eee4f2080303f1ad0c7df866b98fb89222d8d/pdf. [Accessed: 13-Aug-2016].

[13] I. I. Christov, "Real time electrocardiogram QRS detection using combined adaptive threshold", *Biomed. Eng. OnLine*, vol. 3, 2004, p. 28. DOI: 10.1186/1475-925X-3-28.

[14] A. Gautam, Y. D. Lee, W. Y. Chung, "ECG Signal De-noising with Signal Averaging and Filtering Algorithm". In: *Third International Conference on Convergence and Hybrid Information Technology*, 2008, vol. 1, 409–415. DOI: 10.1109/ICCIT.2008.393.

[15] P. Laguna, R. Jané, P. Caminal, "Automatic detection of wave boundaries in multilead ECG signals: Validation with the CSE database", *Comput. Biomed. Res.*, vol. 27, no. 1, 1994, 45–60. DOI: 10.1006/cbmr.1994.1006.

[16] P. W. Macfarlane, B. Devine, E. Clark, "The university of Glasgow (Uni-G) ECG analysis program", Computers in Cardiology, 2005, Lyon, 2005, 451–454. DOI: 10.1109/CIC.2005.1588134.

[17] "Glasgow 12-lead Analysis Program – Physician's Guide", Physio Control. [Online]. Available: https://docs.google.com/viewerng/viewer?url=http://www.physio-control.com/uploadedFiles/learning/clinical-topics/Glasgow_PhysiciansGuide.pdf. [Accessed: 21-Jul-2016].

[18] K. Wang, R. W. Asinger, H. J. Marriott, "ST-segment elevation in conditions other than acute myocardial infarction", *N. Engl. J. Med.*, vol. 349, no. 22, 2003, 2128–2135. DOI: 10.1056/NEJMra022580.

[19] U. Demšar, P. Harris, C. Brunsdon, A. S. Fotheringham, and S. McLoone, "Principal Component Analysis on Spatial Data: An Overview", Ann. Assoc. Am. Geogr., vol. 103, no. 1, 106–128, Jan. 2013. DOI: 10.1080/00045608.2012.689236.

[20] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative Decorrelation for Clustering and Classification", in *Computer Vision* – ECCV 2012, 2012, 459–472. DOI: 10.1007/978-3-642-33765-9_33

[21] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *J. Comput. Appl. Math.*, vol. 20, 53–65, 1987. DOI: 10.1016/0377-0427(87)90125-7.

7. Appendix

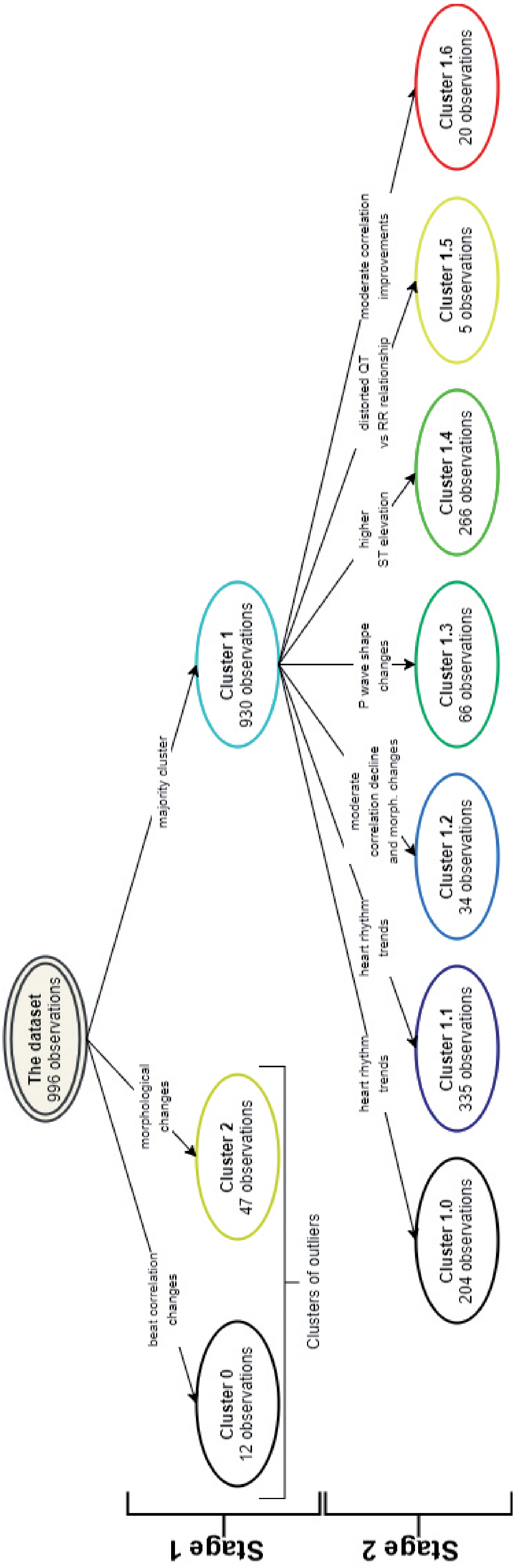

Fig. 7. An overview of the hierarchical cluster analysis results with feature combinations defining each of the clusters