

Evaluating and Comparing Performance of Feature Combinations of Heart Rate Variability Measures for Cardiac Rhythm Classification

Alan Jovic¹, Nikola Bogunovic²

¹Corresponding author: Alan Jovic, M.Sc. Computer Science
Department of Electronics, Microelectronics, Computer and Intelligent Systems,
Faculty of Electrical Engineering and Computing, University of Zagreb,
Unska 3, HR-10000 Zagreb, Croatia
Telephone: ++385 1 6129 548
Fax number: ++385 1 6129 653
E-mail address: alan.jovic@fer.hr

²Nikola Bogunovic, Prof., Computer Science
Department of Electronics, Microelectronics, Computer and Intelligent Systems,
Faculty of Electrical Engineering and Computing, University of Zagreb,
Unska 3, HR-10000 Zagreb, Croatia
E-mail address: nikola.bogunovic@fer.hr

Abstract

Automatic classification of cardiac arrhythmias using heart rate variability (HRV) analysis has been an important research topic in recent years. Explorations reveal that various HRV feature combinations can provide highly accurate models for some rhythm disorders. However, the proposed feature combinations lack a direct and carefully designed comparison. The goal of this work is to assess the various HRV feature combinations in classification of cardiac arrhythmias. In this setting, a total of 56 known HRV features are grouped in eight feature combinations. We evaluate and compare the combinations on a difficult problem of automatic classification between nine types of cardiac rhythms using three classification algorithms: support vector machines, AdaBoosted C4.5, and random forest. The effect of analyzed segment length on classification accuracy is also examined. The results demonstrate that there are three combinations that stand out the most, with total classification accuracy of roughly 85% on time segments of 20 seconds duration. A simple combination of time domain features is shown to be comparable to the more informed combinations, with only 1–4% worse results on average than the three best ones. Random forest and AdaBoosted C4.5 are shown to be comparably accurate, while support vector machines was less accurate (4–5%) on this problem. We conclude that the nonlinear features exhibit only a minor influence on the overall accuracy in discerning different arrhythmias. The analysis also shows that reasonably accurate arrhythmia classification lies in the range of 10 to 40 seconds, with a peak at 20 seconds, and a significant drop after 40 seconds.

Keywords: heart rate variability; linear features; nonlinear features; random forest; support vector machines; boosted C4.5

1. Introduction

Heart rate variability (HRV) analysis examines fluctuations in the lengths of a sequence of cardiac interbeat (RR) intervals. The time-series of interbeat intervals is usually obtained from electrocardiogram (ECG) recordings using high-accuracy R peak detection algorithms [1,2]. HRV analysis allows researchers to assess how the heart rhythm fluctuations can be employed in detecting presence of cardiovascular diseases [3]. Cardiac dysfunction is often manifested by systematic changes in the variability of the RR interval sequence relative to that of a normal rhythm [4], with less variability being associated with old age and the presence or prognosis of various cardiac diseases [5,6]. Many types of irregular ECG patterns can be successfully detected by HRV analysis alone, whereas some patterns (e.g. left and right bundle branch block, nodal rhythm) are known to be practically indiscernible from normal rhythm [7].

HRV analysis using linear time and frequency domain features was pursued in the 1980s and the first half of the 1990s, when it was shown that these features have high descriptive, classification, and predictive capabilities [6]. Linear features characterize time and frequency domain properties of the RR interval series. The problem of nonstationarities and apparent nonlinear behavior of the cardiac rhythm not accurately describable by linear features motivated the researchers to explore other approaches for more accurate modeling of cardiac disorders. Time-frequency features [8,9] and nonlinear features [4,10] were developed and applied. Nonlinear methods were designed to fully describe the observed complex patterns of rhythm disorders, including fractal [11], multifractal [12], and chaotic properties [13].

In the 2000s, the research on HRV analysis continued in two directions. A number of researchers developed novel nonlinear measures for more accurate cardiac rhythm classification [14,15], while others examined various feature combinations for automatic classification and prediction of cardiac arrhythmias [7,16–18]. This second group of authors relies on a well-established assumption that nonlinear features should be used together with the linear ones for the best possible classification results. Nevertheless, the question of actual quantification of the cumulative effect on the accuracy of most of the employed nonlinear features has, to our knowledge, remained open.

Official guidelines for HRV analysis from the European Society of Cardiology date back to 1996 [6]. Thus they do not cover recently developed time-frequency and nonlinear methods. Probably

due to this lack of a common framework for HRV analysis, most authors have problems in comparing their proposed feature combination with others [7,17,19]. The comparison is usually based only on the evaluation measures of final results. Sometimes, the authors do not specify clearly the input data [20] or the selection procedure used for input data segments [17,18]. This leads to an observation that each author uses his own best feature combination for specific classification problem. However, to our knowledge, none of the authors reported how it was established that the proposed combination is better than some other possible combination on the same dataset, with the same analysis parameters. Also, most authors focus their research on classification of two or more rhythm patterns that can be, in theory, rather easily discerned [17,18,21].

An evaluation of the limits of employing HRV analysis to a larger number of similar arrhythmic patterns has not been previously performed. The classification of cardiac arrhythmias based on the combinations of ECG morphological and HRV features is commonly used in literature in order to obtain the best results [22–24]. However, in situations where a full ECG is not available or where morphological features are not reliable enough due to noise, HRV features based on cardiac rhythm are the only ones that can be used for arrhythmia detection. Research in this area in recent years [14,17–19,25] strongly suggests that not all is known about the potential of HRV analysis for automatic classification of cardiac arrhythmias.

This work focuses on answering some of the addressed open questions in HRV analysis. The main objective is to evaluate and compare classification performance of various HRV feature combinations on the same dataset in order to discover which combinations are the most suitable for automatic classification of cardiac arrhythmias. For that purpose, we use three well-known classification algorithms, namely support vector machines (SVM) [26], AdaBoosted C4.5 decision tree [27], and random forest (RF) [28]. In addition, we compare the classification algorithms themselves on this particular problem in order to establish which of the algorithms is the most accurate, with acceptable model construction times. Finally, we perform the analysis on different segment lengths in order to establish the most suitable period for accurate detection of arrhythmias.

We focus on the automatic classification of nine types of non life-threatening cardiac rhythm patterns. These include: normal sinus rhythm (NSR), premature ventricular contractions (PVC), premature atrial contractions (PAC), atrial fibrillation (AF), paced rhythm (P), ventricular

tachycardia (VT), Mobitz II 2nd degree heart block (BII), ventricular bigeminy (VBI), and ventricular trigeminy (VTR). For descriptions of these disorders we refer the reader to the works of authors [29,30]. Discovering these patterns with high accuracy based on HRV information alone can lead to better prevention of more serious cardiac disorders. Perfect discerning of the rhythms based solely on the cardiac rhythm information is not possible due to the similarities between some of them (i.e. PAC and PVC; PVC, VBI, and VTR). Therefore, we examine only the limits of automatic classification.

This paper is organized as follows. In section 2, we describe the input data records, explicate the feature combinations and provide a brief explanation of the employed classification methods. Some details regarding the feature calculation procedure, classification algorithms' parameters and evaluation methods are given in section 3. In section 4 we provide the results of the analysis. In section 5 we discuss the results. Section 6 concludes the paper.

2. Material and methods

2.1 Cardiac records

Cardiac rhythm records have been collected from the PhysioBank database collection of biological signals [31]. For the analysis of rhythm patterns, we decided to use the highly-referenced MIT-BIH Arrhythmia Database [32]. The database consists of 48 patient records, each of them approximately 30 minutes in duration. Roughly 109,000 cardiac beats are recorded in total. The database contains relevant information on: times of R peak occurrence, beat type annotations, rhythm annotations, and other information. Beat and rhythm annotations enabled us to use supervised learning classification methods. RR intervals' time-series has been extracted from the information on the times of the R peaks, filtering the non-beat type information and recording the rhythm type present in the segment. Nine rhythm types stated in the introduction section have been selected for the analysis.

Feature vectors were extracted from each record in the database. The analyzed segment lengths for linear and nonlinear features included 10, 15, 20, 25, 30, 40, 50, 75, and 100 seconds. Longer segments were avoided because the focus is on early detection of arrhythmias. There was no overlap between the segments in order to avoid data dependencies. A total of 23,736 feature vectors have been extracted. In Table 1, statistics regarding the number of feature vectors for each type of cardiac rhythm pattern and segment length is provided.

--> Table 1

2.2 Feature Combinations

In order to establish which of the various possible feature combinations is best suited for the classification of cardiac rhythm patterns, we needed to implement a large number of HRV features. A total of 56 linear and nonlinear features were implemented in our own framework. In this analysis, we group the features into eight combinations. The list of the feature combinations, with references to implementation details, is provided in Table 2. We consider that a feature is nonlinear if it is not a standard statistical description of RR interval segments in time and frequency domain, such as those given in the HRV guidelines [6]. Next, we explain each feature combination in more detail.

--> Table 2

2.2.1 Linear, time domain

The first combination contains most of the linear, time domain features used in HRV analysis. This includes eight statistical and one geometric feature (HRV triangular index, HTI). The extraction of a number of pNNX features is used (see Table 2), because it is not clear which one of them would be the best for overall disorder diagnostics [6,17,33]. It is reasonable to assume that for distinguishing between several heart rhythms one should take a number of different values for X, expressed in milliseconds. The purpose of the first combination is to assess how much information about cardiac disorders can be obtained based exclusively on a simple, linear time domain analysis.

2.2.2 All linear

Frequency analysis has been used extensively in HRV research. Its main purpose is to quantify the relationship between the influences of two branches of autonomic nervous system on the cardiac rhythm [34]. Power spectral density (PSD) estimate is calculated in two frequency bands (LF = [0.04-0.15] Hz, HF = [0.15-0.4] Hz) [6]. For the PSD estimate, we used parametric, autoregressive model of order = 16, which was calculated using Burg algorithm [35]. The second combination uses both nine time domain and four frequency domain features. Frequency features: LF, HF, LF/HF ratio, and total PSD were included for segments of 25 s or longer, because the calculation for shorter segments is unreliable.

2.2.3 Linear + nonlinear variations

This feature combination considers all linear features and includes three other features: Poincaré plot SD1/SD2 ratio [36], Fano factor [4], and Allan factor [4]. SD1/SD2 ratio is used to quantify the relation between short and long-term HRV. Fano and Allan factors are nonlinear statistical features that are used to count the variation of the number of RR intervals by dividing the analyzed segment into m sub-segments of equal length. The third feature combination is used to encompass variations on shorter and longer time scales.

2.2.4 Linear + nonlinear entropy

The fourth feature combination examines the influence of nonlinear entropy measures proposed by several authors on the overall accuracy of automatic classification. We consider the most often used entropy measures in HRV analysis literature, including approximate entropy (ApEn) [37], sample entropy on multiple time scales (multiscale SampEn) [38], and spectral entropy (SpEn) [39]. Also included is the maximum value of ApEn [40].

2.2.5 Combination #5

The fifth feature combination was proposed by Asl et al. [17]. It is a selection of 14 linear time domain, frequency domain, nonlinear variations, entropy and chaos features.

2.2.6 Combination #6

The feature combination proposed by Yaghouby et al. [18] is considered. It consists of only nine features of various types.

2.2.7 Combination #7

The seventh feature combination, consisting of 23 features of various types was recently proposed by Jovic and Bogunovic [19] and is considered here.

2.2.8 All features

The eighth feature combination contains all features extracted by any of the previous seven feature combinations, with a total of 56 features. The purpose of this combination is to compare the accuracy of a large-scale feature combination with the more limited feature sets.

2.3 Classification methods

Classification algorithms employed in the domain of ECG and HRV analysis include various types of artificial neural networks (ANN) [18,39], SVM [14,17], and decision trees [23,41].

In this work, we compare three data mining classification methods. The methods include SVM, AdaBoosted C4.5, and RF. A brief description of the used methods follows.

2.3.1 Support vector machines

SVM is a kernel based machine learning family of methods that are used to accurately classify both linearly separable and linearly inseparable data [26]. The basic idea when the data is not linearly separable is to transform them to a higher dimensional space by using a transformation kernel function. In this new space the samples can usually be classified with higher accuracy. Many types of kernel functions have been developed, with the most used ones being polynomial and radial-basis. The main advantage of using SVM over ANN is that SVM always finds a global minimum, which is not the case with the neural network that can be stuck in a sub-optimal solution. Therefore, in our work, we use SVM instead of ANN. Sequential minimal optimization algorithm (SMO) algorithm resolves quadratic programming optimization problem that arises when determining the maximum margin hyperplane of the support vector machines classifier [42]. This is a standard SVM learning algorithm. Because SMO is a binary classification algorithm, for multiclass classification purposes required in this work it is adapted such that it performs $n*(n-1)/2$ binary classifications. The algorithm is parametric and deterministic. By varying the parameters such as cost and kernel type one obtains different models, but the models do not vary on the same dataset for independent executions when using the same parameters.

2.3.2 AdaBoosted C4.5

C4.5 is the landmark decision tree algorithm developed by Quinlan [41]. It is a divide-and-conquer algorithm that splits attributes at tree nodes using the information-based gain ratio criterion. The trees are first grown to full length and pruned afterwards in order to avoid data overfitting. AdaBoost is a known algorithm for boosting weak classifiers developed by Freund et al. [43]. Although initially used to boost the accuracy of a series of weak learning decision trees called decision stumps, AdaBoost can be applied as a meta-classifier to any known classifier. Its main purpose is to reduce the classifier's variance error, although it also reduces the bias error to a degree [44]. Structure bias and output variance errors are characteristic to the majority of decision tree learning methods, including C4.5. Hence, the rationale for using AdaBoost to improve C4.5 is sound, as it was also demonstrated on a number of real-world datasets by other authors [27,45]. AdaBoosted C4.5 is a parametric and deterministic algorithm.

2.3.3 Random forest

Random forest (RF) is a state-of-the-art ensemble decision tree learner developed by Breiman [28]. Decision trees choose their splitting attributes (in our case, the extracted HRV features) from a random subset of k attributes at each internal node. The best split is taken among these randomly chosen attributes and the trees are built without pruning, as opposed to C4.5. Random forest avoids overfitting due to two sources of randomness - random attribute subset selection and bootstrap data sampling. Breiman has shown that if one constructs the forest consisting of a high-enough number of random trees, the overall classification error will be minimized and the accuracy will reach a plateau [28]. RF is widely used in various classification problems, especially in domains with larger numbers of attributes and instances, because of its high speed and accuracy [46]. The interpretation of the ensemble's results is, however, problematic due to the inability to extract useful rules out of the forest. RF is a parametric algorithm with respect to the number of trees in the forest. Moreover, it is a stochastic algorithm because of its two sources of randomness. Therefore, the algorithm has to be run several times in order to obtain reliable estimates of mean value and variance of the evaluation measures.

3. Calculation specifics

Feature extraction was performed in our own Java-based framework for HRV analysis, called HRVFrame [47] (previously ECG Chaos Extractor [48]). The only purpose of the framework is to extract HRV features from cardiac rhythm records. Feature vectors are calculated under some predetermined parameters, including the record start time, end time, segment length, option of filtering the non-beat annotations, and others. The framework currently implements more than 30 linear and nonlinear methods, with more than 100 individual features. The 56 features used in this work are only a subset of the total number of implemented features. The framework is still a work in progress, albeit at an advanced stage, and is available free-of-charge for non-commercial scientific purposes.

Classification of the proposed feature combinations was performed in Weka platform for knowledge discovery and data mining, version 3.6.5 [49]. Weka has been chosen as the data mining system in this analysis because of its user-friendliness and efficient implementation of many machine learning methods for small- and medium-size classification tasks.

A pre-requirement that we used for all three classifiers was that their execution time is comparable. Therefore, a squared polynomial SMO algorithm was selected for SVM. We varied the parameters until an optimal setting was found (cost = 20.0, $\epsilon = 1.0 \cdot 10^{-12}$, linear members included, data pre-normalization performed), with respect to the execution time and accuracy. Selecting a higher order polynomial kernel SVM resulted in somewhat more accurate models, however the classification duration increased sharply. AdaBoosted C4.5 was used with the following settings: number of iterations = 30, reweighting of the C4.5 trees, weight threshold for pruning = 100. C4.5 trees used by the meta-classifier were parameterized at the following settings: C4.5 pruning method, minimal number of instances at leaves = 2, confidence factor for pruning = 0.4. The number of trees in the forest used by the RF classifier was varied between 30 and 50, depending on the number of features in the combinations.

The classification was performed using 10-fold cross-validation. For evaluation purposes, we use standard measures in the medical domain, namely total classification accuracy (ACC), sensitivity (SENS), specificity (SPEC), and positive predictive value (PPV) [17,50]. These measures are defined as:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1),$$

$$SENS = \frac{TP}{TP+FN} \quad (2),$$

$$SPEC = \frac{TN}{TN+FP} \quad (3),$$

$$PPV = \frac{TP}{TP+FP} \quad (4),$$

where TP, TN, FP, and FN are the numbers of: true positives, true negatives, false positives, and false negatives, respectively. SENS is the proportion of actual positive examples that are correctly identified as positive. SPEC is, likewise, the proportion of actual negative examples that are correctly identified as negative. PPV is the proportion of actual positive examples that are identified among all positive examples. For multiclass case, these measures can be obtained after the classification is performed by using a confusion matrix, as shown in Fig. 1.

-> Fig. 1

For results comparison between classification algorithms, as well as between different feature combinations, we adopt one-tailed paired student *t*-test, with significance level $\alpha=0.05$.

4. Results

Classification results for the examined feature combinations are presented in Fig. 2–4. The combinations are ranked from left to right in ascending order of the overall model accuracy for all segments.

-> Fig. 2

-> Fig. 3

-> Fig. 4

In Fig. 2, total classification accuracy obtained using RF algorithm is given. Because RF is a stochastic algorithm, its results had to be averaged, which was done over 5 independent executions. Standard deviations of the results are shown in Table 3, with statistically significant results highlighted. Feature combinations had the highest score for the segment length of 20 s, with an exception of combination #4, where the results for several segment lengths were not significantly different. The overall best combination is #8 ($p=0.006$, $p=0.003$, over two second best combinations, #3 and #5, respectively).

-> Table 3

Figure 3 shows total classification accuracy of the models achieved using the AdaBoosted C4.5 algorithm. Three combinations: #3, #5, and #8 stand out the most, with no significant differences between them.

In Fig. 4, the results of the SVM algorithm are given. Four combinations: #3, #4, #5, and #8 have the best results with no significant differences between them.

From Fig. 2–4 it can be observed that the overall classification results for the SVM classifier is much lower (roughly 4–5%) when compared to AdaBoosted C4.5 and RF. In Table 4, we summarize the relation between the results for all three classifiers for each combination. There are nine segment lengths for which we can determine significance of the differences between the classifiers. If there is no statistical difference, the result is counted as a tie. Otherwise, a win is associated to one of the classifiers, and a loss to the other one. Apparently, the results for AdaBoosted C4.5 and RF are even, and both of the classifiers win against SVM.

-> Table 4

In Tables 5–11 we show detailed results with respect to cardiac rhythm patterns for the statistically best models for each feature combination. The models are selected based on the

following procedure: 1) Find the significantly best segment length with respect to classification accuracy for both RF and AdaBoosted C4.5, 2) Choose the model with statistically higher accuracy between the two algorithms. 3) If both models are not significantly different, select one at random.

-> Table 5	-> Table 6
-> Table 7	-> Table 7
-> Table 8	-> Table 8
-> Table 9	-> Table 10
-> Table 11	

Table 5 contains both the model for combinations #1 and #2, because the combinations had the same features for segment length of 20 s.

The best results are obtained for the following rhythm patterns: NSR, PVC, AF, P, and BII. All of them had sensitivity greater than 80%, specificity greater than 93%, and positive predictive value greater than 70%. Somewhat less satisfactory results are obtained for PAC, VBI, and VBI. The results for VT are poor. The highest results are obtained for feature combinations #3, #5, and #8, presented in Tables 6, 8, and 11, respectively.

Most of the feature combinations have the best model for the segment length of 20 s, with the exception of combinations #3 (10 s), and #8 (30 s). High model accuracies can be obtained in the range 10–40 s by all feature combinations, with the accuracies falling after 40 s.

We performed a separate analysis that uses the same number of feature vectors for each segment length in order to verify that the 20 s segment is the best choice for arrhythmia classification regardless of the number of feature vectors used for training. This required taking a random stratified subsample (disorder ratios in the subsample did not change) of the feature vectors for shorter time segments in order to have roughly the same number of vectors as the 50 s segments (~1303 vectors). Longer segments (75 s and 100 s) were disregarded because the variations were already shown to be too large. Five random subsamples were analyzed with the RF algorithm, running 5 independent executions of the algorithm for each subsample. We show the results for combination #8 in Fig. 5. Standard deviations in this case are larger than when all the vectors for each length are considered (compare with Table 3). The reason for the larger deviations is because shorter segments now cover only a portion of the database. The 20 s segment length was

again shown to be the best choice for arrhythmia classification ($p < 0.001$ compared to the second best, 40 s segment).

5. Discussion

This section is divided into two parts, the first one addressing the results of the proposed feature combinations and the second one discussing the employed methods.

5.1 Feature combinations

Linear time domain features demonstrated very high classification accuracy. When compared to feature combination #8 using RF algorithm, the results were only 1–4% worse on average, depending on the segment length. For the best segment length of 20 s the results were only approximately 1% worse ($p=0.001$). Obviously, simple linear time domain features have a very high classification potential and are not computationally demanding. Perhaps one of the most useful features in this combination is the mean heart rate. This can be concluded when comparing the results to the only combination that did not have this feature, i.e. combination #7.

In combination #2, frequency domain features were added to time domain features. It appears that the frequency features improve the results of combination #1 to a small extent after adding them at 25 s segment lengths and higher.

Feature combination #3 achieved high accuracy (AdaBoosted C4.5: 87.50%, 10 s, and RF: $85.05\pm 0.14\%$, 20 s) and was comparable to the other two best combinations, #5 and #8. It appears that the inclusion of features measuring both short and long-term variations, which are unrelated to the linear measures [50], improves their result up to a level where it is not necessary to add any other feature to obtain the best models.

Nonlinear entropies examined in combination #4 are shown to have fair results in differentiating between arrhythmias, improving the accuracy of linear features. A probable reason why this combination did not score as high as combinations #3, #5, and #8 is because most nonlinear entropies require at least several hundred intervals for reliable estimates [34]. It may be that the better application of the entropy measures is in discerning healthy patients from those suffering from a heart disorder such as congestive heart failure.

Feature combination #5 proposed by Asl et al. [17] gives high classification results of $ACC=85.24\pm 0.17\%$ with RF for 20 s segments. The feature combination is well balanced and was originally used to distinguish between six types of rhythm patterns: NSR, PVC, AF, ventricular fibrillation, sick sinus syndrome, and BII. The analysis was originally pursued on

segments of equal length of 32 interbeat intervals and achieved high accuracy of 99.16% using SVM classifier with radial-basis kernel function and general discriminative analysis (GDA) based feature reduction. It should be noted that the rhythms patterns analyzed in our work are more challenging than the ones authors considered in their work. Therefore, the classification results achieved on this problem are significantly lower than the ones reported by the authors, regardless of the employed classifier.

Combination #6 [18] contains a small number of features that cover both linear and various aspects of nonlinear analysis. The proposed combination was originally reported achieving complete separation (100% accuracy) of four types of arrhythmias: left bundle branch block, first degree heart block, supraventricular tachycardia, and ventricular trigeminy. The authors performed the classification using multilayer perceptron with GDA based feature reduction. In our work it has been shown that this combination is not as accurate as the combination #5 ($p < 0.001$), having approximately 1% worse results over all segment lengths, with the best result being $ACC = 84.13 \pm 0.15\%$ for RF, 20 s segment. The combination is comparable to the combination #2 in terms of accuracy and other evaluation measures.

Using the combination #7, the authors originally analyzed the classification problem of four types of cardiac disorders, including normal heart rhythm, any arrhythmia, supraventricular arrhythmia, and congestive heart failure. A total of 23 features included linear time domain and a number of nonlinear phase space description features. The nonlinear features were extracted at five different lags between the analyzed RR intervals. This feature combination exhibited 99.6% accuracy on the problem using random forest classifier [19]. Here we have shown that it is the least accurate (AdaBoosted C4.5: $ACC = 81.40\%$, 20 s) among the inspected combinations on this problem, probably because the authors did not consider the mean heart rate. A possible reason why it was not done is because the disorders they classified were not as sensitive to the mean rate as are the arrhythmias inspected in this work.

Although feature combination #8, consisting of all the extracted features, achieved one of the best results (AdaBoosted C4.5: $ACC = 85.63\%$, 30 s, and RF: $85.63 \pm 0.27\%$, 20 s), one should note that the calculation burden of this combination is quite high. Therefore, this combination is probably not suitable for on-line automatic classification. A question arises whether removing many of the possibly redundant features from this combination would result in an even better model. Due to the large number of features in the combinations, this cannot be performed manually, but rather

using some of the known feature selection methods such as various filters and wrappers [51]. This should be a topic for further research.

5.2 Methods

The request for an approximately equal time for model construction and evaluation resulted in a weaker performance of the SVM classifier when compared to RF and AdaBoosted C4.5. We tried to find a better model for SVM that is comparable to the other two methods. We succeeded in finding approximately equally accurate model, however the polynomial was of the sixth degree or higher, with cost parameter over 40, which required an unacceptably long execution time for even the smallest number of feature vectors. One of the possible explanations for poor results of the SVM classifier is the unbalanced dataset, with some of the arrhythmias having less than 5% instances, and NSR having almost 50%. When constructing binary classifiers, the feature vectors ratios may easily go to 30:1, which is known to affect SVM [52]. The other two multiclass classifiers seem to be more robust to this kind of imbalance, at least on this dataset. AdaBoosted C4.5 was comparable to random forest on most of the examined feature combinations both in terms of execution speed and accuracy. For comparison purposes, we also tried running a single non-boosted C4.5 decision tree learner, which resulted in a ten times faster execution, but with much less accurate results (comparable to SVM).

The interpretation of the classification models obtained using the three employed algorithms is difficult. Some authors have devised expert systems based on classification rules, mostly in the form of decision trees, that allow physicians an easier interpretation of the obtained results [23,25]. Other authors have focused solely on the best classification results, disregarding the interpretation issues [14,17,18]. From the physicians perspective, it is important to have a clear understanding of the diagnostics results in order to give accurate diagnosis [53]. However, classification methods that have a clear interpretation, such as C4.5, also tend to have reduced accuracy when compared to classifiers without a clear interpretation such as SVM and RF [19]. The compromise between classification accuracy and interpretability cannot be easily resolved and is a valid topic for further research. In this work, we focused on the limits of accuracy of the models, not on the interpretation.

Although a 10-fold cross-validation was applied as model evaluation procedure, the models for time segments of length 75 s and 100 s usually vary more than for the shorter segments, as

demonstrated in Table 4 for the RF classifier. The variations can be explained as a result of the smaller number of feature vectors for longer lengths. However, a drop in total classification accuracy for longer segments cannot be explained as a result of the reduced number of feature vectors, mainly because of two reasons:

- 1) The shortest time segments of 10–15 s that have the largest number of feature vectors do not have more accurate models than the segments of length 20–40 s.
- 2) All of the segment lengths have enough feature vectors for valid statistical analysis, and all of them contain vectors that cover all the records in MIT-BIH Arrhythmia database.

Performing arrhythmia classification analysis when using the same number of feature vectors for each segment length has some limitations that we are aware of. Taking a random subsample of the feature vectors for lengths shorter than 50 s means disregarding or taking few vectors from some of the cardiac records, if one wishes to obtain a stratified subsample with respect to the disorder classes. Segments of 50 s length thus cover the entire database, while shorter segments cover only a portion of the database, which may produce negative bias towards the results for shorter segments. This bias is more pronounced in our case because the number of records in the database having a specific rhythm pattern is limited, particularly for non-NSR patterns. More records would be welcome in order to have more reliable results for this kind of analysis. It should also be considered that the choice of the optimal period for arrhythmia discrimination is dependent on the type of arrhythmia examined. As we have restricted us to non-life threatening arrhythmias that had enough data for valid conclusions, it is likely that other analyses might find other segment lengths produce more accurate result because of a different choice of analyzed arrhythmias.

The reason why some of the disorders present in the MIT-BIH Arrhythmia Database are not considered in the study is either because they have too few episodes in the records for valid statistical conclusions (e.g. atrial bigeminy, idioventricular rhythm), or because the disorders do not differ significantly from some of the examined rhythm patterns. For instance, bundle branch blocks can only be discerned from normal sinus rhythm using morphological data from several ECG trails [26].

If one examines the SENS, SPEC, and PPV for the analyzed disorders, it is clear that some of them are more easily detected (NSR, PVC, P, AF, BII), while others are more difficult (PAC, VBI, VTR, VT). Detecting VT in particular is quite hard due to the fact that the episodes are of

varying duration, that the start of the episode is similar to PAC or PVC, and that during the episodes the rhythm is hard to discern from sinus tachycardia, which is considered to be NSR in this analysis. If VT and VTR feature vectors are removed from the analysis, the classifiers yield a more accurate model (RF: ACC=89.18±0.17%, 20 s, combination #8), see Table 12. Further eliminating PAC and VBI (because of the similarity with PVC) provides an even more accurate model for five rhythm patterns (NSR, PVC, AF, P, and BII) (RF: ACC=94.44±0.13%, 20 s, combination #8), see Table 13. It should be noted that SENS, SPEC, and PPV have improved significantly only for those rhythm patterns that were similar to the ones excluded. In particular, PVC is now much easier to discern than before.

-> Table 12

-> Table 13

The last reported results are still far from the ones reported by the authors Asl et al. [17] and Yaghouby et al. [18], which are 99.16% and 100%, respectively. The main reason for the less accurate models is because we did not perform any pre-selection of the segments for classifier learning. The whole MIT-BIH Arrhythmia Database has been used “as is”, while the other authors performed manual selection of feature vectors. In real-world applications, one deals with much noise and interference. Thus, the pre-selecting feature vectors for learning might not reflect accurately the difficult classification tasks found in practice.

6. Conclusions

Three feature combinations: #3, #5, and #8 stand out the most. One of the combinations, #5 has been previously proposed by other authors for automatic classification of cardiac rhythms based on HRV analysis. The application of feature combination #8 is unjustified from the computational efficiency point of view. For automatic classification of cardiac rhythms based on HRV features alone we can recommend any of the two feature combinations: #3 or #5. It was found that the contribution of nonlinear measures to the overall classification accuracy is quite small on this problem (~1–2%).

RF and AdaBoosted C4.5 were the most accurate classifiers with no significant differences between them, and SVM was the least accurate on this problem. Both RF and AdaBoosted C4.5 can be recommended for future analyses due to their accuracy and speed.

We have shown that the best classification results for the nine cardiac rhythm patterns are obtained for segments of length 20 seconds, with reasonably accurate results in the range of 10–40 seconds, and falling after 40 s.

An analysis of the application of feature combinations examined in this work to the earliest classification of different cardiac disorders (e.g. congestive heart failure, sudden cardiac death) should be performed. The analysis performed in this work did not include finding the optimal feature combination from the 56 analyzed features. This task should be performed using various feature selection methods and is a topic for further research.

Conflicts of interest

The authors would like to state that they had no conflict of interest in performing this study.

References

- [1] J. Pan, W.J. Tompkins, Real time QRS detector algorithm, *IEEE Trans. Biomed. Eng.* 32 (1985) 230–236.
- [2] M. Adnane, Z. Jiang, S. Choi, Development of QRS detection algorithm designed for wearable cardiorespiratory system, *Comp. Meth. Prog. Biomed.* 93 (200) 20–31.
- [3] R. Kitney, D. Linkens, A. Selman, A. McDonald, The interaction between heart rate and respiration: part II – nonlinear analysis based on computer modeling, *Automedica* 4 (1982) 141–153.
- [4] M.C. Teich, S.B. Lowen, B.M. Jost, K. Vibe-Rheymer, C. Heneghan, Heart-Rate Variability: Measures and Models, in: M. Akay (Ed.), *Nonlinear Biomedical Signal Processing*, vol. II, *Dynamic Analysis and Modeling*, IEEE Press, New York, 2001, pp. 159–213.
- [5] N. Iyengar, C.K. Peng, A.L. Goldberger, L.A. Lipsitz, Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics, *Am J Physiol* 271 (1996) 1078–1084.
- [6] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, Heart rate variability guidelines: Standards of measurement, physiological interpretation, and clinical use, *Eur Heart J* 17 (1996) 354–381.
- [7] M.G. Tsipouras, D.I. Fotiadis, Automatic arrhythmia detection based on time and time–frequency analysis of heart rate variability, *Comp. Meth. Prog. Biomed.* 74(2) (2004) 95–108.
- [8] S.W. Chen, A wavelet-based heart rate variability analysis for the study of nonsustained ventricular tachycardia, *IEEE Trans. Biomed. Eng.* 49(7) (2002) 736–742.
- [9] O. Faust, R.U. Acharya, S.M. Krishnan, L.C. Min, Analysis of cardiac signals using spatial filling index and time-frequency domain, *BioMedical Engineering OnLine* 3 (2004) 30.
- [10] D. Cysarz, H. Bettermann, P. van Leeuwen, Entropies of short binary sequences in heart period dynamics, *Am J Physiol Heart Circ Physiol* 278 (2000) H2163–H2172.
- [11] A.L. Goldberger, Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside, *Lancet* 11 (1996) 1312–1314.
- [12] P.C. Ivanov, L.A. Amaral, A.L. Goldberger, S. Havlin, M.G. Rosenblum, Z.R. Struzik, H.E. Stanley, Multifractality in human heartbeat dynamics, *Nature* 399 (1999) 461–465.

- [13] C. Braun, P. Kowallik, A. Freking, D. Haderler, K.D. Kniffki, M. Meesmann, Demonstration of Nonlinear Components in Heart Rate Variability of Healthy Persons, *Am J Physiol Heart Circ Physiol* 275(5) (1998) 1577–1584.
- [14] K.C. Chua, V. Chandran, R.U. Acharya, L.C. Min, Cardiac Health Diagnosis Using Higher Order Spectra and Support Vector Machine, *Open Med Inform J* 3 (2009) 1–8.
- [15] C.K. Karmakar, A.H. Khandoker, J. Gubbi, M. Palaniswani, Complex Correlation Measure: a novel descriptor for Poincare plot, *BioMedical Engineering OnLine* 8 (2009) 17.
- [16] G. Kudaiberdieva, B. Görennek, B. Timuralp, Heart rate variability as a predictor of sudden cardiac death, *Anadolu Kardiyol Derg* 7 (2007) 68–70.
- [17] B.M. Asl, S.K. Setarehdan, M. Mohebbi, Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal, *Artif. Intell. Med.* 44(1) (2008) 51–64.
- [18] F. Yaghouby, A. Ayatollahi, R. Soleimani, Classification of Cardiac Abnormalities Using Reduced Features of Heart Rate Variability Signal, *World Applied Sciences Journal* 6(11) (2009) 1547–1554.
- [19] A. Jovic, N. Bogunovic, Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features, *Artif. Intell. Med.* 51 (2011) 175–186.
- [20] R.U. Acharya, P.S. Bhat, S.S. Iyengar, A. Raod, S. Dua, Classification of heart rate data using artificial neural network and fuzzy equivalence relation, *Pattern Recogn.* 36 (2003) 61–68.
- [21] F. Yaghouby, A. Ayatollahi, R. Bahramali, M. Yaghouby, A.H. Alavi, Towards automatic detection of atrial fibrillation: A hybrid computational approach, *Comput. Biol. Med.* 40(11-12) (2010) 919-930.
- [22] M. H. Kadbi, J. Hashemi, H. R. Mohseni, A. Maghsoudi, Classification of ECG arrhythmias based on statistical and time-frequency features, In: S. G. Fabri, (Ed.), *Proc IET 3rd Int Conf on Advances in Medical, Signal and Information Processing MEDSIP 2006*, July 17-19, 2006, Glasgow, Scotland, UK, Curran Associates, Red Hook NY, USA, 2006, pp. 1–4.
- [23] T.P. Exarchos, M.G. Tsipouras, C.P. Exarchos, C. Papaloukas, D.I. Fotiadis, L.K. Michalis, A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree, *Artif. Intell. Med.* 40(3) (2007) 187–200.

- [24] J. Kim, H.S. Shin, K. Shin, M. Lee, Robust algorithm for arrhythmia classification in ECG using extreme learning machine, *BioMedical Engineering OnLine* 8 (2009) 31.
- [25] L. Pecchia, P. Melillo, M. Sansone, M. Bracale, Discrimination power of short-term heart rate variability measures for CHF assessment, *IEEE Trans. Inf. Technol. Biomed.* 15(1) (2011) 40–46.
- [26] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.
- [27] J.R. Quinlan, Bagging, Boosting, and C4.5, in: *AAAI, Proc 13th National Conf Artif Intell* , August 4-8, 1996, Portland, OR, USA, MIT Press, Cambridge, MA, USA, 1996, pp. 725–730.
- [28] L. Breiman, Random forests, *Mach Learn* 45 (2001) 5–32.
- [29] T.B. Garcia, N.E. Holtz, *12-Lead ECG: The Art of Interpretation*, Jones and Bartlett Publishers, Sudbury, MA, USA, 2001.
- [30] G. Karraz, G. Magenes, Automatic classification of heartbeats using neural network classifier based on a Bayesian framework, in: A. Laine, K.H. Chon (Eds.), *Conf Proc IEEE Eng Med Biol Soc 1*, August 30-Sept. 3, 2006, New York, IEEE Press, New York, 2006, pp. 4016–4019.
- [31] PhysioBank: Physiologic signal archives for biomedical research, <http://www.physionet.org>, accessed on 2011-09-17.
- [32] G.B. Moody, R.G. Mark, The impact of the MIT-BIH Arrhythmia Database, *IEEE Eng. Med. Biol. Mag.* 20(3) (2001) 45–50.
- [33] T.P. Hutchinson, Statistics and graphs for heart rate variability: pNN50 or pNN20? *Physiol. Meas.* 24(3) (2003) N9–14.
- [34] R.U. Acharya, K.P. Joseph, N. Kannathal, C.M. Lim, J.S. Suri, Heart rate variability: a review, *Med. Bio. Eng. Comput.* 44 (2006) 1031–1051.
- [35] S. de Waele, P.M.T. Broersen, The Burg Algorithm for Segments, *IEEE Trans. Signal Process.* 48(10) (2000) 2876–2880.
- [36] A. Kitlas, E. Oczeretko, M. Kowalewski, M. Borowska, M. Urban, Nonlinear dynamics methods in the analysis of the heart rate variability, *Roczniki Akademii Medycznej w Białymstoku, Annales Academiae Medicae Bialostocensis* 50(Suppl. 2) (2005) 46–47.
- [37] S.M. Pincus, A.L. Goldberger, Physiological time-series analysis: what does regularity quantify? *Am J Physiol* 266 (Heart Circ Physiol 35) (1994) H1643–H1656.

- [38] M. Costa, A.L. Goldberger, C.K. Peng, Multiscale entropy analysis of biological signals, *Phys Rev E* 71 (2005) 021906.
- [39] Y. Özbay, R. Ceylan, B. Karlik, A fuzzy clustering neural network architecture for classification of ECG arrhythmias, *Comput. Biol. Med.* 36 (2006) 376–388.
- [40] K.H. Chon, C.G. Scully, S. Lu, Approximate Entropy for All Signals, *IEEE Eng. Med. Biol. Mag.* 28(6) (2009) 18–23.
- [41] J.R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann Publishers, San Francisco CA, USA, 1993.
- [42] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schölkopf, C.J. Burges, A.J. Smola (Eds.), *Advances in kernel methods—support vector learning*, MIT Press, Cambridge, MA, USA, 1998, pp. 185–208.
- [43] Y. Freund, R.E. Schapire, A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting, in: P.M.B. Vitányi (Ed.), *Proc 2nd Eur Conf Comput Learn Theory EuroCOLT '95*, March 13-15, 1995, Barcelona, Spain, Springer-Verlag London, UK, 1995, pp. 23–37.
- [44] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, Cambridge, MA, USA, 2004.
- [45] S. Pramanik, U.N. Chowdhury, B.K. Pramanik, N. Huda, A Comparative Study of Bagging, Boosting and C4.5: The Recent Improvements in Decision Tree Learning Algorithm, *Asian Journal of Information Technology* 9(6) (2010) 300–306.
- [46] A. Bureau, J. Dupuis, B. Hayward, K. Falls, P. Van Eerdewegh, Mapping complex traits using Random Forests, *BMC Genet.* 4(Suppl. 1) (2003) S64.
- [47] A. Jovic, N. Bogunovic, HRVFrame: Java-Based Framework for Feature Extraction from Cardiac Rhythm, *Lect. Notes Artif. Intell.* 6747 (2011) 96–100.
- [48] A. Jovic, N. Bogunovic, Feature Set Extension for Heart Rate Variability Analysis by Using Non-linear, Statistical and Geometric Measures, in: V. Luzar-Stiffler, I. Jarec, Z. Bekic (Eds.), *Proc 31st Int Conf Information Technology Interfaces ITI 2009*, June 22-25, 2009, Cavtat/Dubrovnik, Croatia, SRCE, Zagreb, Croatia, 2009, pp. 35–40.
- [49] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, *The WEKA Data Mining Software: An Update*, *SIGKDD Explor. Newsl* 11(1) (2009) 10–18.
- [50] M.P. Tulppo, T.H. Makikallio, T.E. Takala, T. Seppanen, H.V. Huikuri, Quantitative beat-to-beat analysis of heart rate dynamics during exercise, *Am J Physiol* 71 (1996) H244–252.

[51] M.A. Hall, G. Holmes, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, *IEEE Trans. Knowl. Data Eng.* 15(3) (2003) 1437–1447.

[52] G. Wu, E. Chang, Class-Boundary Alignment for Imbalanced Dataset Learning, in: T. Fawcett, N. Mishra (Eds.), *Proc ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, August 21-24, 2003, Washington DC, AAAI Press, 2003, pp. 49–56.

[53] K. J. Cios, G. W. Moore, Uniqueness of medical data mining, *Artif. Intell. Med.* 26(1-2) (2002) 1–24.

Figure captions

Fig. 1. Determining TP, TN, FP, and FN from a confusion matrix in a multiclass case. Classes a, b, c, and d represent cardiac rhythm patterns in our case. The numbers in the matrix represent classified feature vectors.

Fig. 2. Classification results for all the examined feature combinations using the RF algorithm.

Fig. 3. Classification results for all the examined feature combinations using the AdaBoosted C4.5 algorithm.

Fig. 4. Classification results for all the examined feature combinations using the SVM algorithm.

Fig. 5. Classification results for feature combination #8 using the RF algorithm. The same number of feature vectors for each segment length is obtained by performing random stratified resampling 5 times, each subsample analyzed with 5 independent executions of the algorithm.

Tables

Table 1. The number of feature vectors for rhythm patterns per segment length, whole MIT-BIH Arrhythmia Database.

Segment length	Total	NSR	PAC	PVC	VT	VBI	VTR	P	AF	BII
100	644	203	64	80	45	79	44	59	60	10
75	858	295	77	110	49	96	54	80	85	12
50	1,303	495	107	175	51	127	69	125	138	16
40	1,628	661	126	218	50	142	72	152	186	21
30	2,185	940	156	285	56	176	90	210	248	24
25	2,636	1,181	176	347	60	197	91	254	301	29
20	3,296	1,556	194	427	59	218	103	318	382	39
15	4,466	2,223	239	574	60	252	118	427	526	47
10	6,720	3,606	291	808	67	319	135	646	778	70
Total / portion, %	23,736	11,160 / 47.02	1,430 / 6.02	3,024 / 12.74	497 / 2.09	1,606 / 6.76	776 / 3.27	2,271 / 9.57	2,704 / 11.39	268 / 1.13

Table 2. Feature combinations.

Combination number	Features in combination	Number of features	Description	Comment
1	Mean [6], SDNN [6], RMSSD [6], SDSD [17], pNN5 [17], pNN10 [17], pNN20 [33], pNN50 [6,33], HTI [6]	9	Linear, time domain	
2	Linear, time domain + (total PSD, LF, HF, LF/HF) [6]	13	All linear	Autoregressive model, Burg algorithm, order = 16
3	All linear, SD1/SD2 [36], Fano factor [4], Allan factor [4]	16	Linear + nonlinear variations	
4	All linear, Approximate entropy (ApEn1-ApEn4) [37], Maximum approximate entropy (MaxApEn) [40], Multiscale sample entropy (SampEn1-SampEn15) [38], SpEn [39]	34	Linear + nonlinear entropy	Autoregressive model, Burg algorithm, order = 16 for SpEn
5	(Mean, RMSSD, SDSD, pNN5, pNN10, pNN50, LF/HF, SD1/SD2, ApEn3, SpEn, LLE, DFA, STA1, STA2) [17]	14	Author combination	$m=2$ for ApEn
6	(Mean, SDNN, pNN50, HTI, LF/HF, SD1/SD2, D ₂ , LLE, SpEn) [18]	9	Author combination	reconstruction dimension $d=10$ for D ₂
7	(SFI, D ₂ , CTM, ApEn1-ApEn4, SDNN, pNN20, RMSSD, HTI) [19]	23	Author combination	Lag $T=\{1,2,5,10, \text{ and } 20\}$ for D ₂ , SFI, CTM; reconstruction dimension $d=10$ for D ₂ , SFI, CTM; $m=2$ for ApEn
8	All linear, SD1/SD2, Fano factor, Allan factor, ApEn1-ApEn4, MaxApEn, SampEn1-SampEn15, SpEn, D ₂ , SFI, CTM, LLE, DFA, STA1, STA2	56	All features	Lag $T=\{1,2,5,10, \text{ and } 20\}$ for D ₂ , SFI, CTM; reconstruction dimension $d=10$ for D ₂ , SFI, CTM; $m=2$ for ApEn

Table 3. Results for 5 independent executions of the RF algorithm for each feature combination. Statistically significant best results (one-tailed paired t -test, $\alpha=0.05$) are highlighted.

	Total classification accuracy, mean \pm standard deviation, %							
	Comb #1	Comb #2	Comb #3	Comb #4	Comb #5	Comb #6	Comb #7	Comb #8
10	83.46 \pm 0.15	83.46 \pm 0.15	84.36 \pm 0.25	82.75 \pm 0.14	83.46 \pm 0.21	82.44 \pm 0.24	76.77 \pm 0.22	84.38 \pm 0.10
15	83.84 \pm 0.13	83.84 \pm 0.13	84.12 \pm 0.29	83.07 \pm 0.22	83.88 \pm 0.12	83.18 \pm 0.20	79.93 \pm 0.19	84.79 \pm 0.31
20	84.61 \pm 0.30	84.61 \pm 0.30	85.05 \pm 0.14	83.49 \pm 0.27	85.24 \pm 0.17	84.13 \pm 0.15	80.95 \pm 0.20	85.63 \pm 0.27
25	82.61 \pm 0.30	83.26 \pm 0.24	84.03 \pm 0.42	83.62 \pm 0.10	84.40 \pm 0.29	83.16 \pm 0.26	79.49 \pm 0.38	84.90 \pm 0.07
30	82.02 \pm 0.29	83.18 \pm 0.12	84.44 \pm 0.09	83.49 \pm 0.26	84.13 \pm 0.31	83.50 \pm 0.25	78.65 \pm 0.20	84.60 \pm 0.02
40	82.11 \pm 0.23	83.66 \pm 0.28	83.98 \pm 0.40	83.70 \pm 0.35	84.28 \pm 0.27	82.80 \pm 0.56	77.86 \pm 0.14	84.71 \pm 0.58
50	79.40 \pm 0.16	80.88 \pm 0.42	81.66 \pm 0.18	81.21 \pm 0.33	82.41 \pm 0.64	80.87 \pm 0.24	76.85 \pm 0.42	82.85 \pm 0.62
75	77.29 \pm 0.59	77.72 \pm 0.63	78.98 \pm 1.17	79.51 \pm 0.70	79.84 \pm 0.40	78.11 \pm 0.59	75.96 \pm 0.65	81.96 \pm 0.34
100	76.15 \pm 0.85	77.61 \pm 0.88	78.88 \pm 0.77	78.38 \pm 0.89	78.48 \pm 0.29	78.01 \pm 0.65	75.50 \pm 1.05	80.93 \pm 0.58

Table 4. Win/loss/tie scores when comparing three classifiers for each feature combination. The scores are based on one-tailed paired t -test with $\alpha=0.05$.

	Win/loss/tie		
	RF vs. AdaBoosted C4.5	RF vs. SVM	AdaBoosted C4.5 vs. SVM
Comb #1	7/2/0	9/0/0	9/0/0
Comb #2	5/2/2	8/0/1	9/0/0
Comb #3	2/4/3	8/0/1	9/0/0
Comb #4	0/5/4	8/0/1	9/0/0
Comb #5	2/5/2	8/0/1	9/0/0
Comb #6	6/2/1	9/0/0	9/0/0
Comb #7	2/6/1	8/1/0	9/0/0
Comb #8	3/4/2	9/0/0	9/0/0
Total:	28/28/16	67/1/4	72/0/0

Table 5. Combination #1 and #2, segment length = 20 s, RF, ACC=84.61 \pm 0.30%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
Normal	94.70 \pm 0.12	96.14 \pm 0.26	95.67 \pm 0.26
PAC	67.84 \pm 2.39	98.44 \pm 0.11	73.44 \pm 1.54
PVC	80.74 \pm 0.44	95.08 \pm 0.22	70.92 \pm 0.84
VT	6.46 \pm 2.22	99.66 \pm 0.05	26.32 \pm 8.46
VBI	61.00 \pm 1.48	98.08 \pm 0.04	69.24 \pm 0.80
VTR	37.68 \pm 2.48	98.90 \pm 0.14	53.12 \pm 2.82
P	91.14 \pm 0.68	98.50 \pm 0.07	86.58 \pm 0.79
AF	88.48 \pm 1.36	96.48 \pm 0.15	76.90 \pm 0.87
BII	90.26 \pm 4.22	100.00 \pm 0.00	97.78 \pm 1.25

Table 6. Combination #3, segment length = 10 s, AdaBoosted C4.5, ACC=87.50%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	93.81	93.49	94.38
PAC	62.52	99.02	74.04
PVC	79.70	96.58	76.01
VT	14.89	99.67	34.53
VBI	74.03	98.71	74.23
VTR	53.34	99.30	61.01
P	86.67	98.10	83.19
AF	93.11	98.12	86.65
BII	98.64	100.00	98.60

Table 7. Combination #4, segment length = 20 s, AdaBoosted C4.5, ACC=84.77%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	94.31	96.28	95.82
PAC	68.03	98.64	75.04
PVC	80.78	95.13	74.98
VT	16.89	99.50	40.03
VBI	87.21	96.78	77.97
VTR	45.64	99.19	66.20
P	91.18	98.42	86.11
AF	87.19	96.81	78.02
BII	92.30	100.00	97.29

Table 8. Combination #5, segment length = 20 s, RF, ACC=85.24±0.17%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	94.78±0.13	96.22±0.37	95.76±0.42
PAC	68.34±2.07	98.34±0.13	71.90±2.17
PVC	81.60±1.48	94.96±0.31	70.74±1.03
VT	10.54±2.22	99.80±0.07	48.22±9.67
VBI	63.84±1.84	98.08±0.08	70.06±1.46
VTR	43.10±3.66	99.14±0.05	61.94±1.30
P	90.62±1.23	98.62±0.04	87.34±0.45
AF	89.78±0.87	96.92±0.08	79.32±0.36
BII	85.66±2.29	99.98±0.04	98.28±2.53

Table 9. Combination #6, segment length = 20 s, RF, ACC=84.13±0.15%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	94.68±0.15	95.30±0.21	94.74±0.25
PAC	62.06±1.30	98.32±0.22	69.92±2.30
PVC	80.80±0.44	95.62±0.16	73.34±0.63
VT	12.92±0.93	99.60±0.00	37.26±1.89
VBI	62.56±1.29	98.02±0.13	69.18±1.79
VTR	34.96±1.83	98.90±0.16	50.86±3.82
P	89.88±0.91	98.22±0.04	84.26±0.38
AF	87.54±0.78	96.58±0.13	77.08±0.59
BII	87.21±0.00	99.91±0.00	92.90±1.37

Table 10. Combination #7, segment length = 20 s, AdaBoosted C4.5, ACC=81.40%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	93.30	94.29	93.56
PAC	57.70	98.14	65.45
PVC	76.77	94.36	67.12
VT	21.97	99.47	43.33
VBI	60.61	97.82	65.70
VTR	31.07	98.63	42.10
P	80.24	97.71	78.72
AF	84.82	96.78	77.89
BII	89.70	99.92	92.09

Table 11. Combination #8, segment length = 30 s, AdaBoosted C4.5, ACC=85.63%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	96.10	96.71	95.74
PAC	73.66	98.52	79.32
PVC	81.44	95.64	73.41
VT	17.97	99.35	45.54
VBI	72.37	97.83	73.82
VTR	54.42	98.92	68.10
P	94.31	98.91	90.37
AF	87.53	97.08	79.52
BII	83.32	99.94	90.89

Table 12. Combination #8, rhythms VT and VTR excluded, segment length = 20 s, RF, ACC=89.18±0.17%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	95.60±0.16	95.94±0.23	95.90±0.24
PAC	65.76±2.30	98.84±0.11	79.16±1.66
PVC	84.80±0.78	95.54±0.23	75.06±0.99
VBI	64.68±0.32	98.90±0.12	81.62±1.57
P	92.28±1.19	98.48±0.11	87.42±0.65
AF	91.84±0.26	98.28±0.08	88.12±0.61
BII	83.10±1.37	100.00±0.00	100.00±0.00

Table 13. Combination #8, rhythms VT, PAC, VBI, and VTR excluded, segment length = 20 s, RF, ACC=94.44±0.13%

Rhythm pattern	Sensitivity, %	Specificity, %	Positive predictive value, %
NSR	95.66±0.24	96.14±0.15	97.06±0.11
PVC	90.50±0.46	97.62±0.11	87.68±0.51
P	95.24±0.13	98.60±0.16	89.90±0.89
AF	94.40±0.42	99.22±0.08	95.32±0.43
BII	83.60±1.37	100.00±0.00	100.00±0.00

		Actual class			
		a	b	c	d
Predicted class	a	55	15	20	5
	b	12	45	4	10
	c	0	5	120	15
	d	8	12	10	12

	- TP
	- TN
	- FP
	- FN





