Test–retest reliability and feature selection in physiological time series classification

Steinn Gudmundsson[*,a], Thomas Philip Runarsson[a], Sven Sigurdsson[a]

[a]*Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, School of Engineering and Natural Science, University of Iceland, Reykjavik, Iceland.*

Abstract

Feature test-retest reliability is proposed as a useful criterion for the selection/exclusion of features in time series classification tasks. Three sets of physiological time series are examined, EEG and ECG recordings together with measurements of neck movement. Comparisons of reliability estimates from test-retest studies with measures of feature importance from classification tasks suggest that low reliability can be used to exclude irrelevant features prior to classifier training. By removing features with low reliability an unnecessary degradation of the classifier accuracy may be avoided.

Key words: Time series, Classification, Test-retest reliability, Feature selection

1. Introduction

Time series are encountered in a wide variety of scientific and engineering disciplines. They come in many different forms and can possess several different properties such as having deterministic and/or stochastic components, being long or short, single or multivariate, having high or low signal to noise ratio, being stationary or non-stationary. It is often of interest to classify the objects generating the time series into two or more pre-specified classes. Practical applications include

[*]Corresponding author.

Email addresses: steinng@hi.is (Steinn Gudmundsson), tpr@hi.is (Thomas Philip Runarsson), sven@hi.is (Sven Sigurdsson)

- Predicting whether a patient is likely to survive a heart attack following an admission to an intensive care unit. The prediction is (in part) based on electrocardiogram recordings.
- Predicting whether a patient entering a memory clinic is likely to suffer from Alzheimer's disease using electroencephalogram recordings.
- Differentiating between actual and feign whiplash injuries using measurements of neck movement.

Classifiers for time series are often based on *features*, numerical values, calculated from the time series (a task referred to as *feature extraction*) [1, 2, 3]. The feature values are subsequently used as inputs to an off-the-shelf classifier such as a neural network, linear discriminant or a decision tree [4].

A general problem in classifier design is to decide which features should be used as inputs to the classifier (*feature selection*). The features must capture those aspects of the time series which are relevant for discriminating between the classes. While the requirement that features relevant to the classification task must be included is obvious, it may be less obvious that avoiding irrelevant features is also important. Irrelevant features increase the dimensionality of the problem and degrade classifier accuracy in general [5, 6]. The performance decline depends on the type of classifier used, e.g. nearest neighbor classifiers are sensitive to noise features while decision tree ensembles such as Random forests are fairly robust (though not immune, c.f. [7] p. 596). Correlated features are common and pose a similar problem, with different classifiers affected to varying degree. Dozens or even hundreds of features have been proposed at one time or another for many types of physiological signals. After perusing the specialist literature one usually ends up with a lot of features, too many for comfort.

Feature selection algorithms [6, 8] automate the search for useful features. While these methods are often helpful, they can be difficult to apply. Many are computationally expensive (which may or may not be relevant) and with small data sets they are prone to overfit, resulting in a poorly performing classifier [9]. This is especially true when parameter tuning has to be done in addition to feature selection. In many cases, there is also the need to obtain performance estimates. A direct approach utilizing e.g. nested cross-validation, three levels deep, is likely to be unstable unless a large amount of data is available.

It is reasonable to assume that test-retest reliability (*stability*) is an important trait for features, i.e. when the same object is measured repeatedly under the same conditions, the resulting feature values should not vary "too much". The test-retest reliability of many physiological parameters can be found in the literature, e.g. heart rate variability and respiration rate [10]; quantitative EEG features [11]; measures of trunk accelerations in gait analysis [12] and fMRI data [13]. An alternative is to assess the stability of features using external data set(s), partly or fully independent of the training data (sections 5.1 – 5.3).

This study investigates the use of stability as a criterion for feature selection/exclusion using three sets of physiological data, electroencephalogram (EEG) recordings, electrocardiogram (ECG) recordings and measurements of the movement control of the neck. A feature importance measure is obtained separately using an off-the-shelf feature selection algorithm. Plots of stability versus importance show that low stability correlates with low importance. We therefore suggest excluding features with low stability as an initial step in classifier design to get rid of irrelevant features. When the choice of features is not obvious, it is common to include more features rather than less and hope that the classifier does not suffer too much from the inclusion of irrelevant or redundant features. A recent monograph on the subject argues that including too many features is preferred to inadvertently discarding relevant features [6]. Thus it is useful to have some procedure for identifying inherently inadequate features, keeping also in mind the savings in computation, possible reduction in both the stress experienced by the subjects and the cost of data collection.

2. Feature extraction

This section gives a brief description of feature extraction and provides an overview of feature selection methods.

2.1. Feature extraction

The traditional way of constructing classifiers for time series is to extract a set of features from the time series, the choice of features being dictated by problem specifics. The original time series are thus replaced by a set of points in Euclidean space which are then used as inputs to vectorial pattern classifiers.

Some information is lost in the feature extraction process since the complete time evolution of a (complex) system can seldom be summarized by a few numbers. The loss is acceptable as long as the features capture properties relevant for discrimination between different classes. Replacing the original time series with features is therefore a form of dimensionality reduction and can also serve as noise suppression.

If a reasonable model of the data generating process is available (e.g. for the electrical activity of the heart), the parameters of the model, obtained by fitting the model to each training example, are an obvious choice of features. Often no realistic model is available. The reason may be lack of effort in constructing such a model because of time or cost constraints, or because the complexity of the system is simply too high (e.g. the brain). In these cases, it may still be possible to capture the essence of the system using relatively simple features extracted from the time series such as

- Autoregressive model coefficients [1].
- Coefficients based on Fourier [14] or wavelet transforms [15].
- Parameters derived from nonlinear systems theory such as the correlation dimension and Lyapunov exponents [16].
- Parameters based on entropy or complexity of the time series [17].

Note that these features may also apply in the situation where a model is available.

2.2. Feature selection

Feature selection methods attempt to identify the features most relevant for classification and can be broadly categorized as either subset selection methods or ranking methods [18]. The former type returns a subset of the original set of feature which are deemed to be the most important for classification. Ranking methods sort the features according to their usefulness in the classification task. The classifier designer uses the ranking to select the final feature subset, often in an ad-hoc manner [19].

The algorithms can further be classified into *embedded*, *filter* and *wrapper* methods. Embedded methods [6] perform feature selection as part of the training process, examples include decision tree algorithms such as CART [20] and Random forests [21]. Variants of linear support vector machine classifiers

(SVM) such as [22] seek decision rules which are sparse in the set of features and can therefore be regarded as embedded methods. Filter methods [23, 6] are executed prior to the training of a classifier and are independent of the classification algorithm used. A simple filter method for two-class problems ranks features based on the magnitude of the Pearson correlation coefficient of individual inputs with the labels. The top-ranking features are then used to train the final classifier. Since the features are considered independently, this method does not reveal groups of features which have discriminative power when considered together. Methods based on higher order statistics such as mutual information have been proposed to alleviate this shortcoming. Filter methods are usually fast. While some filter methods utilize label information, others do not, e.g. methods based on principal component analysis. In wrapper methods [24], the desired classifier is used as a part of the feature selection process. For a given subset of features, a classifier is trained and an estimate of its accuracy is used as a performance measure for the feature set. Exhaustive enumeration of all subsets is prohibitive, except when the number of features is very small (say 20 – 25). If the feature selection criteria is monotone, a branch and bound approach may be used to speed up the search [25]. An alternative is to employ global search strategies such as genetic algorithms to optimize the performance measure [26]. Heuristic methods such as forward selection or backward elimination are often employed to iterate through the subsets [6]. In forward selection, a classifier is trained using a single feature only. The feature which gives the best accuracy is retained. Each of the remaining features is tested together with the first feature and the one which gives the highest accuracy is added. The process is repeated until the improvement in accuracy becomes negligible. Backward selection starts with all the features. At each stage, the feature which gives the smallest decrease in accuracy upon removal is deleted. The process is carried out until the drop in performance becomes significant. Sequential floating forward search (SFFS) [27] elaborates on the simple forward selection method and has frequently been found to work well in practice. Recursive feature elimination [18] is a wrapper-type method for two-class problems. A linear SVM is trained repeatedly, each time removing the feature with the smallest contribution to the decision rule. The process continues until a single feature is left. The result is an importance ranking of all the features. When the amount of training data is relatively small, care must be taken to avoid overfitting in the final classifier since the same data set may be used for feature selection, parameter tuning and performance estimation. Wrapper

methods can suffer from high computational complexity since the classifier has to be trained and evaluated on each feature subset. On the other hand, the accuracy of the final classifier is often better than for filter methods [24].

3. Feature stability

In the absence of systematic changes, repeated measurements of the same feature should not vary too much. It is therefore reasonable to assume that features useful for discriminating between groups of subjects possess high test-retest reliability. In many applications, e.g. those dealing with physiological measurements, it is possible to assess the stability of the features using unlabeled data which may be fully or partly independent of the training data, or simply by perusal of the relevant literature.

The general idea is to exclude features with low stability and retain those with high stability. High stability in itself is clearly not a sufficient criterion, since trivial features which are highly stable but useless for classification can easily be constructed, e.g. a feature which always takes the value of zero.

Stability is estimated on the basis of difference in measurements either *between* recording sessions or *within* the same recording session. The time between recording sessions in this study ranges from several days (ECG data) up to three weeks (neck data). Within-session stability tends to be higher than between-sessions stability. Several measures of test-retest reliability have been proposed. When there are only two recording sessions available, correlation coefficients such as the Pearson product-moment coefficient, Spearman's rho or Kendall's tau are often used. When more than two sessions are available intra-class correlation coefficients based on a one way analysis of variance model such as ICC(1) [28] can be used, provided that feature values are normally distributed. The ICC is defined as follows

$$ICC = \frac{MS_{\text{between}} - MS_{\text{within}}}{MS_{\text{between}} + (k-1)MS_{\text{within}}}$$

where MS_{between} is the mean square error between subjects, MS_{within} is the mean square error within subjects and k is the number of sessions. The ICC becomes one when there is perfect agreement between sessions and zero when the between subjects error equals the within subjects error. In rare cases the ICC can become negative, i.e. when the within subject error exceeds the between subjects error. Note that heterogeneity in the subject

group leads to increased values of the between subjects error thus inflating the ICC.

If the normality assumptions of the ANOVA model are violated and transforms towards normality are not possible, the non-parametric Kendall's coefficient of concordance (W) [29] can be used instead of the ICC. Kendall's W takes values in the range $0 \leq W \leq 1$, where a higher value indicates stronger agreement between trials. Kendall's W can be expressed as $\bar{\rho} = (kW - 1)/(k - 1)$ where $\bar{\rho}$ is the average of Spearman's rank order correlation for all pairs of sessions [29].

Selection of cut-off values for stability values is currently a qualitative affair. A plot of the stability values sorted in ascending order may reveal a region of low stability (c.f. EEG and neck movement data sets below) which could then be used to define a threshold. It may be possible to associate "low" stability with absolute values of the test-retest measure, e.g. for the Pearson correlation coefficient, values below 0.3 might be considered "low". Heuristic rules such as discarding the bottom 25% or 33% percentiles can also be used.

4. Feature importance

The Random forests (RF) classifier [21] was used to obtain the estimates of feature importance. The method was selected since it is freely available and often works well in practice.

Random forests build many decision trees by repeatedly drawing bootstrap samples of the training data. A decision tree with binary splits is built from each bootstrap sample using a variant of the CART algorithm. Approximately 2/3 of the training samples are included in the bootstrap sample, the remaining 1/3 are termed out-of-bag samples and are used to obtain an error rate estimate for the final classifier. During the tree-building process each tree is grown as large as possible and instead of searching over all the features for the best split at each node, the search is performed over a random subset of the features. The size of the subset is equal to the square root of the number of features in this study. Classification is performed by sending the example to be classified down all the trees in the forest. Each tree votes for a class and the class label is determined by majority vote. The out-of-bag error rate estimate is obtaining by sending each example down the trees for which it was out-of-bag and the label predicted via majority vote. The error rate estimate is the fraction of times the prediction differs from the true label.

Feature importance (a.k.a. variable importance) is obtained by considering the out-of-bag examples again. After classifying the out-of-bag examples in the usual way, the values for feature j in the out-of-bag samples are randomly permuted and sent down the corresponding trees. The votes for each class are counted and subtracted from the number of votes obtained prior to permuting the feature values. The average of this value over all the trees in the forest is the feature importance score.

It has recently been pointed out that the method used in RF shows preference for both categorical features with many categories and also for correlated features [30]). While categorical variables are not relevant in this study, correlation is an issue.

5. Case studies

Three case studies are presented below. The data involves EEG, Neck movement and ECG recordings. In each of the application examples a labeled data set is used to obtain an importance measure for each feature. Feature stability is estimated from an independent data set in the EEG case. In the neck movement example, stability is estimated from the training data together with measurements recorded in a separate session. For the ECG application, stability is estimated by combining a part of the training data with an external set of measurements. Label information is ignored during stability estimation. Once both stability and importance values are available for each feature, a plot with stability versus importance is generated.

5.1. EEG

Electroencephalograms are real time measurements of brain electrical activity. Most EEG recordings are non-invasive and are carried out by placing 20 – 24 electrodes on the scalp. The amplitude range is between 1 μV and 1 mV with frequencies between 0.1 - 30 Hz. The weak signal is often corrupted by noise.

In the clinical setting, EEGs are mainly used for diagnosis of epilepsy and sleep disorders. To a lesser extent, EEGs are used to monitor anesthesia levels in surgery, to detect the onset of cerebral hypoxia during carotid endarterectomy and in diagnosing dementias such as Alzheimer's.

The variability observed in EEG recordings can be attributed to changes in vigilance (e.g. due to drug effects or drowsiness) and the randomness that is inherent in the EEG. The former can be accounted for to some extent by

carefully controlling experimental conditions but the latter is unavoidable. The EEG variability is reflected to a different extent in different features.

In [11], the stability of several quantitative EEG features was estimated from repeated measurements of a group of elderly subjects (10 sessions over a 2-month period). The features were derived from power spectral, entropy and complexity analysis of the EEG. A brief description of the features is given in appendix A. The purpose of the study was to obtain a general screening tool for features which were later to be included in EEG classifier systems. The working hypothesis was that features with low test-retest reliability were to be avoided in general. The intraclass correlation coefficient ICC(1) was used to quantify stability after carrying out appropriate transformation towards normality.

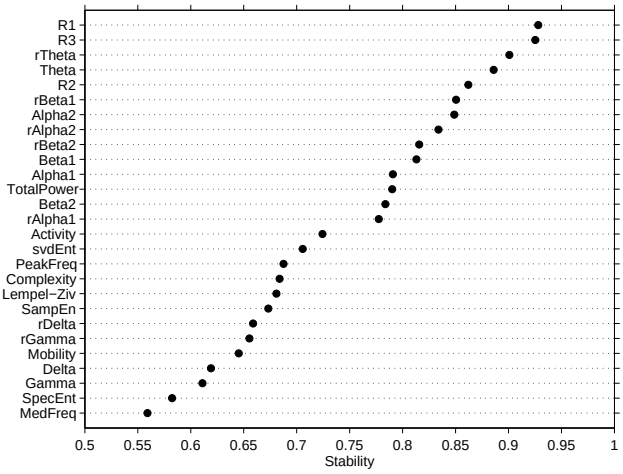

Figure 1: Stability of EEG features, averaged over electrode positions.

Figure 1 shows the stability of the EEG features averaged over electrode positions. The gap between 0.72 and 0.78 suggests a cut-off point at 0.75 to split the features into groups of high and low stability.

The second data set comes from an independent study on the use of EEG in diagnosing Alzheimer's disease (AD). While the clinical use of EEG in AD diagnosis is currently limited, multiple studies have suggested that EEG may play an important role in early diagnosis, especially when coupled with other imaging techniques such as CT, SPECT, PET and MRI (for a review see [31]). Combining EEG with statistical classifiers seems especially promising [32]. The data set consists of EEG recordings from 74 subjects diagnosed with

mild to severe Alzheimer's dementia and 74 healthy age matched subjects. This study was approved by the National Bioethics Committee and written consent was obtained from all the participants or their legal guardians. Features derived from power spectral parameters [33], coherence measures [34] and nonlinear time series analysis [31] have all been suggested for discriminating between healthy subjects and those with AD. The same set of features was computed for both the stability and AD data sets. With 20 channels and 27 features per channel there are 540 features. Many of the features are highly correlated. Each feature *type* (e.g. spectral entropy) is computed at 20 different electrode sites. Sites which are spatially close to each other will give correlated features. Additional correlation is introduced by the average montage used in both studies. Some of the feature types (e.g. R1 and R3) are defined similarly and are therefore expected to be highly correlated. Other features such as mobility, svd entropy, sample entropy and Lempel-Ziv complexity have been shown empirically to be correlated [11].

A Random Forests classifier trained on the AD data was used to obtain feature importance scores. Figure 2 shows stability values for each of the EEG features computed from the stability study versus feature importance values from the AD study, in both cases the values have been averaged over all channels. The highest importance scores are obtained for features with stability above 0.88. Inspection of figure 1 suggested a cut-off point at 0.75 which is conservative in retrospect. The features $\theta, \%\theta$, R_1, R_2 and R_3 all reflect activity in the theta band (3.5 - 7.5 Hz) and the feature importance scores indicate that this particular aspect of the EEG is highly relevant for discrimination between the AD and healthy groups. This is in agreement with previously reported findings on the slowing of EEG. Increased slow wave activity has been reported for some AD patients when compared to healthy controls (for an overview of EEG abnormalities in AD see [31]). This type of activity would be directly reflected in the δ and $\%\delta$ features but both receive low importance scores. This might be explained by the fact that both features are very sensitive to artifacts such as eye movement which frequently contaminate EEG recordings. The sensitivity to artifacts is reflected in their low stability values (figure 1).

To test the effect of selecting features on the basis of their stability, four different RF classifiers were trained on subsets of features which were chosen so that one of the sets had only features with low stability, one had only features with high stability and two were somewhere in-between. The first subset corresponds to the features with the 7 lowest stability values (the first

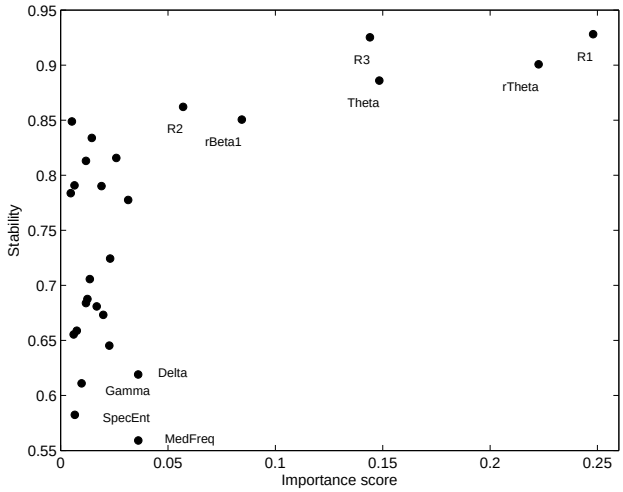

Figure 2: Stability versus feature importance for the EEG features, averaged over electrode positions. The importance scores are derived from the AD dataset and the stability values from a separate dataset. Low stability correlates with low importance.

quartile). The second subset corresponds to the 14 lowest stability values (values below the median). The third subset contains the features with the 13 highest stability values and the final subset corresponds to the 7 largest stability values (the fourth quartile). The performance of the corresponding classifiers was estimated as follows. Two-thirds of the data were randomly selected and used to train a RF classifier and the remaining third of the data (test set) was used to obtain performance estimates. The procedure was repeated 100 times and the average and standard deviations reported. The error rates in table 1 show that classifiers built with low stability features have higher error rates compared to classifiers trained on stable features. The reason for the relatively small difference in error rates (9%) between the best and the worst classifiers may be that all the features are somewhat relevant for AD classification. This does not come as a surprise since the feature set was intentionally biased towards AD. In fact most of the features have at some point been suggested as biomarkers for AD.

The Sequential floating forward search method of [27] was used to determine the optimal feature set[1]. The algorithm was applied to the whole data set since the objective was not to evaluate the performance of the result-

[1]The SFFS implementation in PRTools [35] was used with default parameter settings.

Stability values	Error rate	Sensitivity	Specificity
Lowest 7	29.0 (5.5)	71.8 (8.8)	70.6 (9.2)
Lowest 14	26.7 (5.3)	70.1 (8.4)	77.2 (9.7)
Highest 13	22.0 (4.7)	76.3 (7.6)	79.8 (8.2)
Highest 7	20.2 (4.4)	78.3 (7.6)	81.3 (8.4)
All	21.5 (5.4)	76.6 (8.1)	80.9 (8.0)

Table 1: Performance estimates (%) and their corresponding standard deviations of Random Forests classifiers trained on different subsets of features of the AD data. The terms highest and lowest refer to the ranking of stability values.

ing classifier but rather to investigate whether the algorithm would suggest features with low stability. The resulting feature set consisted of R1, R3, rTheta, rBeta1 and rAlpha2 which all have fairly high stability.

5.2. Neck movement

A new method for detecting fraudulent whiplash claims based on measurements of neck movement was proposed in [3]. The movement of the head is measured by an electromagnetic tracking system. Specialized software is used to convert the measurements to coordinates of an on-screen cursor which is controlled by the user. Another cursor on the screen (called the Fly) traces out predetermined movement patterns of varying difficulty (easy, medium and hard). The subjects are asked to follow the computer controlled cursor as closely as possible with their own. Only the cursors are visible, not their trajectories, which makes prediction of movement difficult. The deviation between the measured and actual trajectory is quantified and used as input to an ensemble of support vector machine classifiers (see appendix B for a description of the features). The ensemble was trained on a group of 34 subjects with chronic whiplash disorder together with a group of 31 healthy subjects which had been instructed to feign whiplash injury. The healthy group consisted of 16 women and 15 men, ages 16 - 67 years (mean 38, SD 17). The whiplash group consisted of 28 women and 6 men, ages 21 - 56 years (mean 41, SD 9).

Each subject was measured on two separate occasions, 2 – 3 weeks apart and multiple trials were carried out on each visit, three trials for each of the easy, medium and hard trajectories. Data from both sessions was used to assess the stability while data from the first week only was used to train the

classifier, i.e. the training set is a subset of the data set used to estimate stability. This is in contrast to the EEG case above where feature importance and stability were estimated from separate data sets.

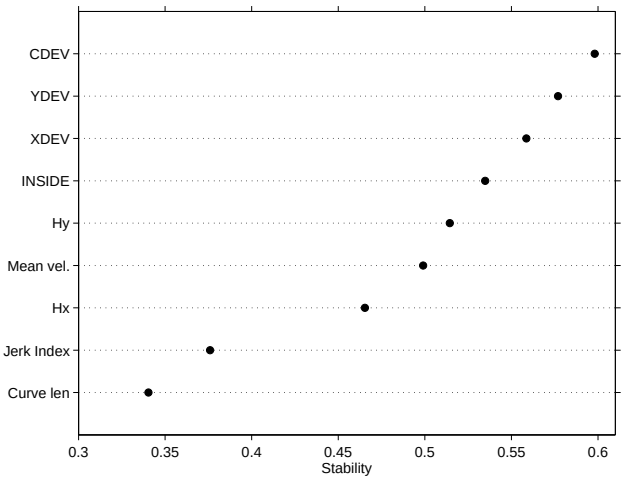

Figure 3: Stability of the neck movement features.

Figure 3 shows the stability of the nine features described in [3], obtained using the trajectory of medium difficulty, the first trial in each week and Kendall's tau rank correlation coefficient. Note that absolute stability values cannot be compared with the EEG study since different measures of stability are involved. There are at least two factors confounding the reliability estimates. First there are some subjects who got better at the tracking task between the sessions and secondly there are some whiplash subjects who do considerably worse in the later session, e.g. because of tiredness. Two features stand out in terms of low reliability, the jerk index, which attempts to quantify the smoothness of movement, and curve length. These two features are therefore expected to be of limited use in classification.

A Random forests classifier was trained on the healthy and whiplash groups to obtain feature importance scores. Figure 4 shows the stability values versus feature importance. The results confirm that both the jerk index and curve length features are not useful for discriminating between the two groups. This had previously been established in [3] by a scatter plot of the features using the class label as a grouping variable.

The final classifier ensemble used only two features, CDEV and HY. The features XDEV, YDEV and INSIDE were not included since they were

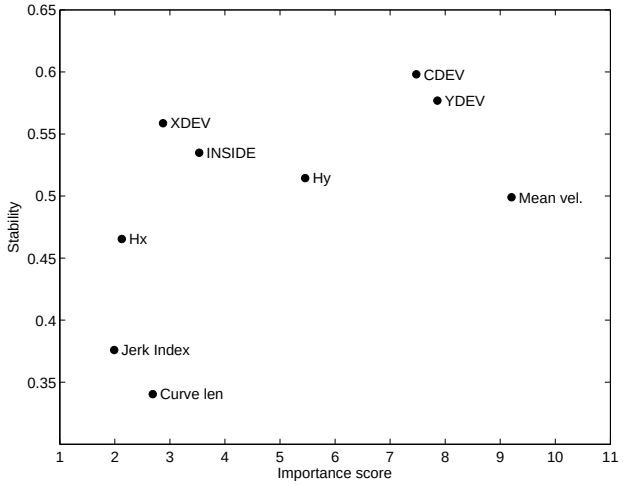

Figure 4: Stability versus feature importance for the neck movement features. Low stability correlates with low importance.

strongly correlated with CDEV and provided little additional information.

The SVM ensemble classifier from [3] was trained on four different feature subsets chosen on the basis of stability values in the same way as in section 5.1. The results are shown in table 2. Using the two features with the lowest stability, the jerk index and curve length, gives an inaccurate classifier. The good performance of the classifier trained on the four lowest features can probably be attributed to the mean velocity feature which was found useful for separating the two groups in [3].

Stability values	Error rate	Sensitivity	Specificity
Lowest 2	32.7 (1.2)	43.1 (4.3)	89.4 (2.4)
Lowest 4	16.4 (1.5)	80.0 (3.0)	86.8 (2.9)
Highest 5	16.4 (0.4)	84.7 (2.4)	82.5 (0.7)
Highest 2	14.2 (0.8)	89.2 (1.9)	82.7 (1.4)
All	16.3 (1.6)	85.4 (2.3)	82.1 (3.0)

Table 2: Performance estimates (%) and their corresponding standard deviations of an SVM ensemble trained on different subsets of features of the neck movement data. The terms highest and lowest refer to the ranking of stability values. The figures are averages from 100 cross-validation runs, randomly shuffling the training examples each time.

The SFFS algorithm was then applied to the data. The feature set re-

turned by SFFS consists of the MEANVEL, INSIDE, HY and HX features, i.e. the algorithm omits both JERKINDEX and CURVELEN which have the lowest stability.

5.3. ECG

Electrocardiograms are recordings of the electrical activity of the heart. They are used to diagnose a wide variety of heart conditions such as arrhythmias, electrolyte imbalance, sleep apnea, enlargement of the heart, and coronary artery disease.

Myocardial infarction (MI) is the necrosis of parts of the heart muscle, commonly known as a heart attack. A single (brief) ECG recording is not sufficient for accurate diagnosis of myocardial infarction, a patient with acute MI may have normal ECG, and the diagnosis of MI is normally based on several tests such as blood tests, chest X-rays and ECGs.

A classifier to detect MI from healthy ECG is described in [36]. It uses a single feature based on the entropy of heart rate variability and achieves an accuracy of 70%. A neural network classifier based on Hermite expansions of 12-lead ECG had an accuracy of 84% in [2]. By including the results of extensive laboratory tests and other clinical data with the ECG an accuracy of 97.5% is reported in [37].

The PTB database [38, 39] available on PhysioNet [40] contains 549 records from 290 subjects, both volunteers and patients suffering from various types of heart disease. Each subject is represented by one to seven records which enables the estimation of feature stability.

Of these 290 subjects, 148 had MI and 52 were healthy. The remaining subjects had other diseases and were excluded from this study. Of the 200 MI and healthy subjects, 89 MI subjects and 14 healthy subjects had multiple recordings. Two recordings from each of the 103 subjects were used to estimate stability after extracting several simple ECG features thought to be relevant to the classification task (see appendix C for details). The average time between sessions was 4.4 days.

Figure 5 shows the stability of the ECG features measured by Kendall's tau rank correlation coefficient. Apart from three fairly small gaps, the decline is gradual and selecting a cut-off point from the graph is difficult. When the stability values are averaged over the ECG leads, there is a fairly clear separation into regions of high, medium and low stability (figure 6). The duration of the PR, QRS and PQ intervals stand out in terms of high stability while all the features derived from the R-R intervals, except meanRR, have

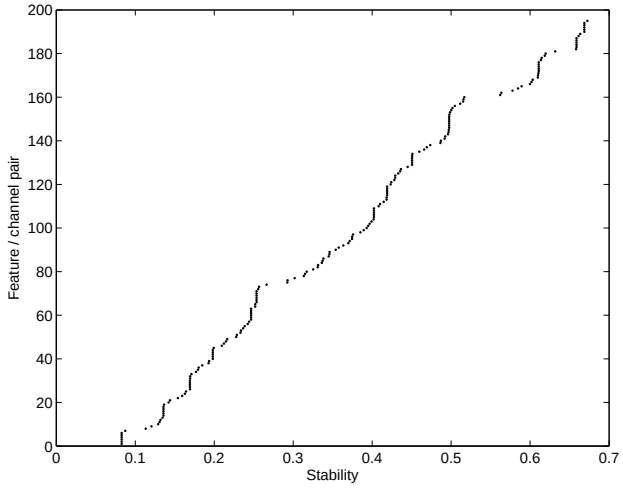

Figure 5: Stability of the ECG features (all features and leads).

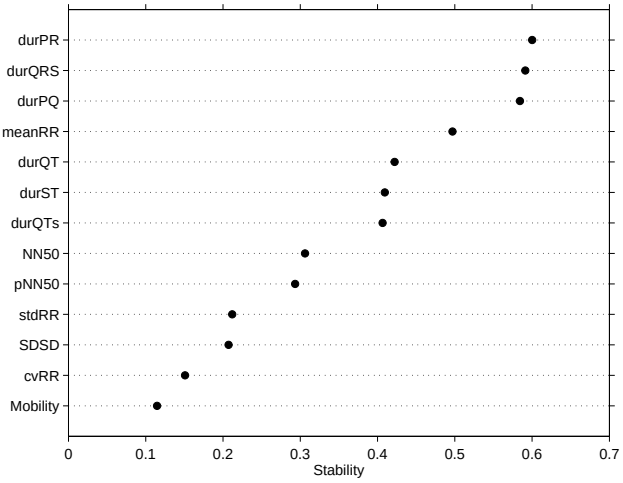

Figure 6: Stability of the ECG features averaged over leads.

low stability. The features NN50, pNN50, stdRR, SDSD, cvRR and Mobility all measure variation in instantaneous heart rate. The relatively low stability of these features is probably because the heart rate fluctuates considerably under normal conditions. Reduction in variability can be a sign of coronary disease such as MI (c.f. [41]).

To construct the RF classifier and obtain feature importance, 52 MI and 52 healthy controls were used. The remaining MI cases were discarded in

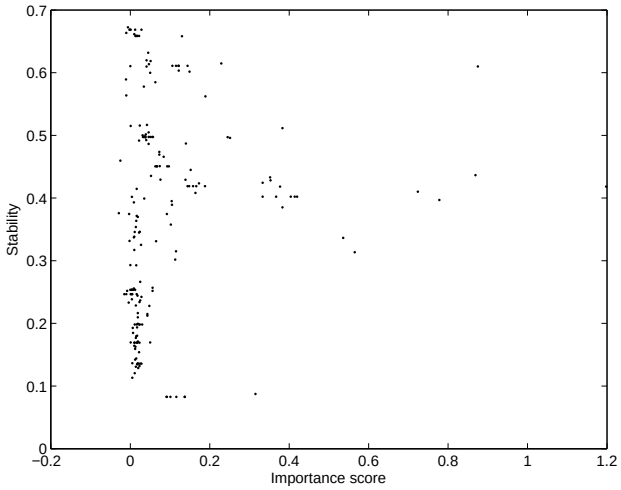

Figure 7: Stability versus feature importance for the ECG features (all features and leads).

order to get a balanced training set. Figure 7 shows the relationship between stability and feature importance for all feature/lead pair and figure 8 show the corresponding averages over the 15 leads. Low stability again correlates with low importance. For the EEG and Neck data sets the most important features were also highly stable. The ECG features durPQ and durPR show that this is not true in general, i.e. high stability does not imply high importance. Here the "best" features have stability in the range 0.4 – 0.6.

The performance estimates in table 3 were obtained with the procedure described in section 5.1. The estimates show that classifiers built with the low stability features have higher error rates compared to the classifiers which exclude them. And also, using only the most stable features may be counter-productive.

The SFFS algorithm was then applied to the data. Of the 17 features returned by SFFS, three had farily low stability values (0.19 – 0.30) while the remaining features had stability above 0.3 (compare to figure 7).

6. Discussion

In all three case studies low stability was correlated with low feature importance suggesting that stability can be used to pre-screen features prior to training a classifier, excluding features with low reliability. On the other hand, the ECG data set demonstrates that high stability does not imply that

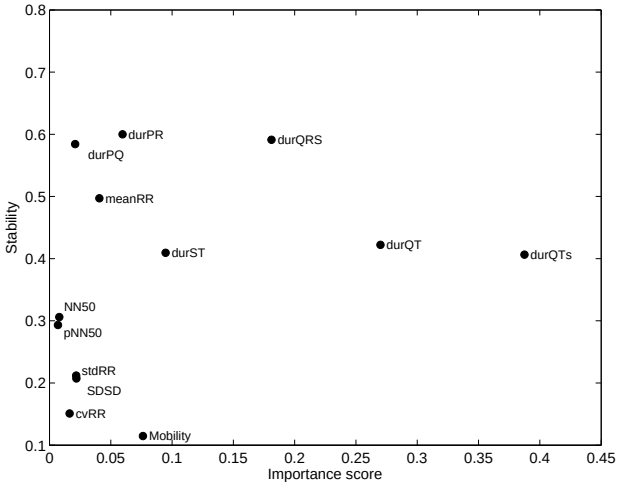

Figure 8: Stability versus feature importance for the ECG features averaged over leads.

Stability values	Error rate	Sensitivity	Specificity
Lowest 49	37.7 (7.7)	59.5 (12.8)	66.1 (13.3)
Lowest 98	32.7 (7.8)	64.3 (12.9)	71.7 (12.9)
Highest 97	25.5 (6.2)	69.5 (11.4)	80.1 (8.4)
Highest 49	29.7 (7.3)	69.0 (11.9)	72.9 (12.9)
All	24.9 (6.6)	72.4 (11.1)	78.3 (9.7)

Table 3: Performance estimates (%) and their corresponding standard deviations of Random Forests classifiers trained on different subsets of features of the ECG data. The terms highest and lowest refer to the ranking of stability values.

the corresponding features are useful in classification.

To investigate whether the observed results were merely an artifact of using the Random Forests feature selection procedure, the conceptually different recursive feature elimination algorithm [18] was used to obtain importance scores[2]. When the importance scores were plotted against stability values (not shown), the results were qualitatively very similar to those described in sections 5.1 – 5.3.

After excluding features with low stability, classifier design proceeds as

[2]The C parameter in the SVM was set to 0.1 in all cases.

usual, e.g. by first applying a traditional feature selection algorithm, followed by training and evaluation of the classifier. The removal of (presumably) irrelevant features is expected to lead to a classifier with increased accuracy. That excluding features of low stability is beneficial for classifier accuracy was further confirmed by the results of applying the SFFS algorithm to each of the three data sets. In most cases, SFFS returned moderately or highly stable features.

In some cases the training data can be used to assess stability, in other cases a separate data set is used. It may even not be necessary to carry out an independent stability study since test-retest reliability data for many physiological parameters is readily available in the specialist literature.

Based on the ECG and Neck movement studies, a cut-off value of 0.4 may be suitable when the test-retest interval is $0.5 - 2$ weeks and Kendall's tau is used to assess stability. In the EEG case, a cut-off value of 0.75 was found to be conservative (low) for the ICC stability measure which was derived from 10 recording sessions performed over a 2-month period.

Apart from biomedical tasks, classifiers for other types of data might benefit from the stability criteria. One such example is feature-based face recognition. Databases containing facial images often contain multiple images of the same person, making stability estimation of image features possible.

Future work includes developing quantitative approaches for selecting the cut-off values for stability. Such methods can be based on hypothesis testing or confidence intervals for the test-retest measures. Confidence intervals for the ICC can be computed using the procedure in [28]. A significance test for Kendall's W is given in [29] and confidence intervals for the rank order correlations can be computed using bootstrap approaches [42]. Correlation amongst feature values and the issue of multiple hypothesis testing would have to be taken into account in order to develop formal criteria.

Appendix A. EEG feature description

The EEG recordings were manually scored for artifacts and segments containing visible artifacts were excluded from the feature extraction process. Table 4 lists the features extracted from the EEG data sets. For more details on the treatment of artifacts and feature extraction, see [11].

Feature	Description
Delta	Absolute power in the δ band (0.5 – 3.5 Hz).
Theta	Absolute Power in the θ band (3.5 – 7.5 Hz).
Alpha1	Absolute Power in the α_1 band (7.5 – 9.5 Hz).
Alpha2	Absolute Power in the α_2 band (9.5 – 12.5 Hz).
Beta1	Absolute Power in the β_1 band (12.5 – 17.5 Hz).
Beta2	Absolute Power in the β_2 band (17.5 – 25 Hz).
Gamma	Absolute Power in the γ band (25 – 40 Hz).
rDelta	Relative power in the δ band.
rTheta	Relative power in the θ band.
rAlpha1	Relative power in the α_1 band.
rAlpha2	Relative power in the α_2 band.
rBeta1	Relative power in the β_1 band.
rBeta2	Relative power in the β_2 band.
rGamma	Relative power in the γ band.
TotalPower	Total power in the (0.5 – 40 Hz) band.
PeakFreq	Peak frequency in the (7.5 – 12.5 Hz) band.
MedFreq	Median frequency.
SpecEnt	Spectral entropy.
R_1	Power ratio $R_1 = \theta/(\alpha_1 + \alpha_2 + \beta_1)$.
R_2	Power ratio $R_2 = (\delta + \theta)/(\alpha_1 + \alpha_2 + \beta_1 + \beta_2)$.
R_3	Power ratio $R_3 = \theta/(\alpha_1 + \alpha_2)$.
Activity	$A = a_0$.
Mobility	$M = (a_1/a_0)^{1/2}$.
Complexity	$C = (a_2/a_1 - a_1/a_0)^{1/2}$.
SampEn	Sample entropy.
Lempel-Ziv	Lempel-Ziv complexity.

Table 4: The EEG features used in the study, a_0 is the variance of the signal, a_1 is the variance of the derivative of the signal and a_2 is the variance of the second derivative of the signal.

Appendix B. Features derived from neck movement

Each measurement results in a bi-variate time series, $P(t)$, containing the on-screen x and y coordinates derived from the electromagnetic tracking system. The corresponding actual trajectory is denoted by $Q(t)$. Table 5 lists the features extracted from the neck movement data. The computation

is described in detail in [3].

Feature	Description
CDEV	Average distance between $P(t)$ and $Q(t)$.
XDEV	Average horizontal deviation between $P(t)$ and $Q(t)$.
YDEV	Average vertical deviation between $P(t)$ and $Q(t)$.
INSIDE	Fraction of points which satisfy $\|\|P(t) - Q(t)\|\| <$ constant.
CURVELEN	Length of the measured trajectory.
MEANVEL	Mean velocity of the user controlled cursor.
JERKINDEX	Rate of change of acceleration of the user controlled cursor.
HX	Entropy of x-axis deviations.
HY	Entropy of y-axis deviations.

Table 5: The features derived from the measurements of neck movement, $P(t)$ is the measured trajectory, $Q(t)$ is the actual trajectory.

Appendix C. ECG feature extraction

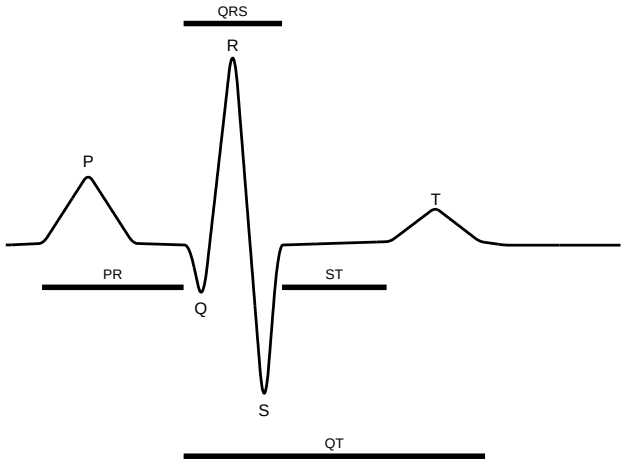

Figure 9: The QRS complex.

Each record consists of 32 seconds of 12 conventional and three Frank leads, sampled at 1000 Hz. The durST, durPR, durPQ, durQT, durQTS and durQRS features listed in table 6 are well known time-domain descriptors

derived from the QRS complex (figure 9). The relatively short recordings exclude the (sensible) use of many "standard" ECG features such as those based on low frequency analysis of the signal and features derived from chaos theory. A heart rate variability time series (HRV) is derived from the duration of adjacent R-R intervals. The HRV series is then used to compute the stdRR, cvRR, NN50, pNN50 and a proxy for HRV entropy. The HRV *mobility* is a simple approximation to Sample Entropy [11]. Mobility is defined as SDSD divded by stdRR. The ECGPUWAVE package [43] on PhysioNet was used to detect the QRS complexes and the locations of the P, QRS and S-T waveforms. With 13 features and 15 channels, the total number of features is 195.

Feature	Description
durST	Duration of the ST interval.
durPR	Duration of the PR interval.
durPQ	Duration of the PQ interval.
durQT	Duration of the QT interval.
durQTs	Duration of the QT interval corrected for heart rate [44].
durQRS	Duration of the QRS interval.
meanRR	Mean of the R-R interval duration.
stdRR	Standard deviation of R-R interval duration.
cvRR	Coefficient of variation of the R-R interval duration.
NN50	The number of consecutive R-R intervals which differ by more than 50 ms.
pNN50	The fraction of consecutive R-R intervals which differ by more than 50 ms.
SDSD	The standard deviation of differences between consecutive R-R intervals.
Mobility	Approximation of Sample entropy.

Table 6: The ECG features used in the study.

Acknowledgments

We thank Gudny Lilja Oddsdottir for providing the neck movement data.

Conflict of interest

None.

References

[1] S. J. Roberts, L. Tarassenko, New method of automated sleep quantification, Medical and Biological Engineering and Computing 30 (5) (1992) 509–517.

[2] H. Haraldsson, L. Edenbrandt, M. Ohlsson, Detecting acute myocardial infarction in the 12-lead ECG using Hermite expansions and neural networks, Artificial Intelligence in Medicine 32 (2) (2004) 127–136.

[3] S. Gudmundsson, G. L. Oddsdottir, T. P. Runarsson, S. Sigurdsson, E. Kristjansson, Detecting fraudulent whiplash claims by support vector machines, Biomedical Signal Processing and Control (2010) in–press.

[4] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification (2nd Edition), Wiley-Interscience, 2000.

[5] G. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: Proceedings of the Eleventh International Conference on Machine Learning, Vol. 129, 1994, pp. 121–129.

[6] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh (Eds.), Feature Extraction Foundations and Applications, Springer, 2006.

[7] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction, Springer, 2009.

[8] A. Jain, R. Mao, Statistical pattern recognition: A review, IEEE Transactions on pattern analysis and machine intelligence 22 (1) (2000) 4–37.

[9] R. Kohavi, D. Sommerfield, Feature subset selection using the wrapper method: Overfitting and dynamic search space topology, in: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 1995, pp. 192–197.

[10] A. Guijt, J. Sluiter, M. Frings-Dresen, Test-retest reliability of heart rate variability and respiration rate at rest and during light physical activity in normal subjects, Archives of medical research 38 (1) (2007) 113–120.

[11] S. Gudmundsson, T. P. Runarsson, S. Sigurdsson, G. Eiriksdottir, K. Johnsen, Reliability of quantitative EEG features, Clinical Neurophysiology 118 (10) (2007) 2162–2171.

[12] M. Henriksen, H. Lund, R. Moe-Nilssen, H. Bliddal, B. Danneskiod-Samsře, Test-retest reliability of trunk accelerometric gait analysis, Gait & Posture 19 (3) (2004) 288–297.

[13] R. Maitra, S. Roys, R. Gullapalli, Test-retest reliability estimation of functional MRI data, Magnetic Resonance in Medicine 48 (1) (2002) 62–70.

[14] S. Gudmundsson, T. P. Runarsson, S. Sigurdsson, Automatic sleep staging using support vector machines with posterior probability estimates, in: Proceedings of 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, Vol. 2, 2005, pp. 366–372.

[15] K. Sternickel, Automatic pattern recognition in ECG time series, Computer methods and programs in biomedicine 68 (2) (2002) 109–115.

[16] J. Fell, J. Röschke, K. Mann, C. Schäffner, Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures, Electroencephalography and clinical Neurophysiology 98 (5) (1996) 401–410.

[17] N. Kannathal, M. Choo, U. Acharya, P. Sadasivan, Entropies for detection of epilepsy in EEG, Computer methods and programs in biomedicine 80 (3) (2005) 187–194.

[18] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (1-3).

[19] E. Tuv, A. Borisov, K. Torkkola, Ensemble-based variable selection using independent probes, in: I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh

(Eds.), Feature Extraction: Foundations and Applications, Springer-Verlag, 2006, pp. 131–145.

[20] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman & Hall, Monterey, CA, 1984.

[21] L. Breiman, A. Cutler, Random forests, Software and documentation downloaded Jan 2009 from `http://www.stat.berkeley.edu/ breiman/RandomForests`.

[22] T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.-R. Müller, K. Obermayer, R. Williamson, Classification on proximity data with LP-machines, in: Ninth International Conference on Artificial Neural Networks, IEE, London, 1999, pp. 304–309.

[23] K. Kira, L. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: Proceedings AAAI-92, MIT Press, Cambridge, MA, 1992, pp. 129–134.

[24] R. Kohavi, G. John, Wrappers for feature subset selection, Artificial intelligence 97 (1-2) (1997) 273–324.

[25] P. Narendra, K. Fukunaga, A branch and bound algorithm for feature subset selection, IEEE Transactions on Computers 100 (26) (1977) 917–922.

[26] W. Siedlecki, J. Sklansky, A note on genetic algorithms for large-scale feature selection, Pattern Recognition Letters 10 (5) (1989) 335–347.

[27] P. Pudil, J. Novovicová, J. Kittler, Floating search methods in feature selection, Pattern recognition letters 15 (11) (1994) 1119–1125.

[28] K. McGraph, S. Wong, Forming inferences about some intraclass correlation coefficients, Psychological Methods 1 (1) (1996) 30–46.

[29] R. Meddis, Statistics Using Ranks - A Unified Approach, Basil Blackwell, Oxford, 1984.

[30] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for Random forests, BMC Bioinformatics 9 (2008) 307.

[31] J. Jeong, EEG dynamics in patients with Alzheimer's disease, Clinical Neurophysiology 115 (7) (2004) 1490–1505.

[32] C. Lehmann, T. Koenig, V. Jelic, L. Prichep, R. John, L. Wahlund, Y. Dodge, T. Dierks, Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG), Journal of neuroscience methods 161 (2) (2007) 342–350.

[33] K. Bennys, G. Rondouin, C. Vergnes, J. Touchon, Diagnostic value of quantitative EEG in Alzheimer's disease, Neurophysiologie Clinique 31 (2001) 153–160.

[34] M. Brunovsky, M. Matousek, A. Edman, K. Cervena, V. Krajca, Objective assessment of the degree of dementia by means of EEG, Neuropsychobiology 48 (2003) 19–26.

[35] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D. Tax, S. Verzakov, PRTools 4.1, A Matlab Toolbox for Pattern Recognition, Software and documentation downloaded May 2010 from `http://prtools.org`.

[36] S. Lau, J. Haueisen, E. G. Schukat-Talamazzini, A. Voss, M. Goernig, U. Leder, H.-R. Figulla, Low HRV entropy is strongly associated with myocardial infarction, Biomed Tech (Berl) 51 (2006) 186–189.

[37] D. Conforti, R. Guido, Kernel-based support vector machine classifiers for early detection of myocardial infarction, Optimization Methods and Software 20 (2) (2005) 401–413.

[38] R. Bousseljot, D. Kreiseler, A. Schnabel, Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das Internet, Biomedizinische Technik 40 (1) (1995) S317–S318.

[39] H. Koch, R. Bousseljot, D. Kreiseler, L. Schmitz, The PTB diagnostic ECG database, Data downloaded Oct 2010 from `http://physionet.cps.unizar.es/physiobank/database/ptbdb/`.

[40] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new

research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[41] D. Van Hoogenhuyze, N. Weinstein, G. Martin, J. Weiss, J. Schaad, X. Sahyouni, D. Fintel, W. Remme, D. Singer, Reproducibility and relation to mean heart rate of heart rate variability in normal subjects and in patients with congestive heart failure secondary to coronary artery disease., The American journal of cardiology 68 (17) (1991) 1668–1676.

[42] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, Boca Raton, 1994.

[43] P. Laguna, R. Jané, P. Caminal, Automatic detection of wave boundaries in multilead ECG signals: validation with the CSE database, Computers and Biomedical Research 27 (1) (1994) 45–60.

[44] A. Sagie, M. G. Larson, R. J. Goldberg, J. R. Bengtson, D. Levy, An improved method for adjusting the QT interval for heart rate (the Framingham Heart Study), The American Journal of Cardiology 70 (7) (1992) 797 – 801.