7

Automatic Detection of Paroxysmal Atrial Fibrillation

Redmond B. Shouldice, Conor Heneghan and Philip de Chazal
BiancaMed, NovaUCD, Belfield Innovation Park, Ireland

1. Introduction

The purpose of this chapter is to provide a tutorial level introduction to (a) the physiology and clinical background of paroxysmal (intermittent) atrial fibrillation (PAF), and (b) methods for detection of patterns consistent with AF using electrocardiogram (ECG) processing. The document assumes that the reader is familiar with basic signal processing concepts, but assumes no prior knowledge of AF or pattern classification. A practical implementation of an automatic AF detector is presented; a supervised linear discriminant classifier is used to estimate the likelihood of a block of inter-heartbeat intervals being PAF, with accuracies of 92%, 94%, 100% and 100% when the method was used to process the publically available Physionet (Goldberger et al., 2000) signal databases MITDB, AFDB, NSRDB and NSR2DB respectively.

2. Clinical background: Atrial fibrillation

2.1 Physiology, prevalence and health consequences

Atrial fibrillation (AF) is the most commonly found sustained cardiac arrhythmia in clinical practice, and has serious associated mortality and morbidity (Benjamin et al., 1998). For example, about 15% of strokes occur in people with atrial fibrillation, so AF is a significant risk factor for stroke. The prevalence of AF increases with age and is slightly more common in men than in women. The prevalence of AF is 0.5% for the group aged 50 to 59 years and rises to 8.8% in the group aged 80 to 89 years (Kannel et al., 1982). 1–2% of the population suffer from AF at present, and the number of affected individuals is expected to double or triple within the next two to three decades both in Europe and in the USA; in addition, the presence of AF is associated with a marked reduction in everyday functioning and quality of life (Kirchhof et al., 2009).

AF is where the heart's atria quiver instead of beating in a coordinated fashion, reducing effectiveness as a pump, and potentially causing clots to form if blood pools within the atria. Disorganized continuous atrial activity (giving rise to ECG "fibrillatory" waves) may give rise to irregular ventricular rate, visible as "irregularly irregular" variations in the RR ECG interval.

2.2 Clinical implications of automatic detection of PAF

The major hazard of AF is stroke. Frequent episodes of PAF can be captured by 24 hour Holter; however, if episodes are rare or asymptomatic, then long term Holter ECG or ECG

event recorders can help. For instance, Roche et al. (2002) note that in patients still complaining of palpitations after one negative 24-hour Holter, numerous, prolonged, and often asymptomatic episodes of PAF can be revealed by long-term automatic event recorders.
It has been shown that trans-telephonic ECG monitoring can increase the detection rate of PAF in stroke and transient ischaemic attack patients whose 24-hour Holter result was negative (Gaillard et al., 2010). Short duration monitoring (24-72hr Holter) identifies new AF in only about 5% of patients post stroke. In contrast, longer monitoring (e.g., 4 or 7-day event loop recorder) detected new atrial fibrillation/flutter in 5.7%-7.7% of consecutive patients in 2 studies (Liao et al., 2007).

2.3 Diagnostic techniques for PAF

Cardiac event recorders have been shown to have a higher diagnostic yield in patients with intermittent symptoms such as palpitations or dizziness. Since paroxysmal AF may often be the underlying cause of some of these symptoms, event recorders are widely used to document PAF. However, a limitation of patient triggered recorders is that AF may be asymptomatic or occur whilst the patient is sleeping. There may also be inconsistent use of a manual trigger by the elderly. Therefore, the newest generation of AF recorders incorporate auto-trigger features, which will automatically record ECG strips, when certain parameters are recognized – for example, physician defined thresholds of bradycardia or tachycardia.
Accordingly, there is interest in algorithms for the automated recognition of AF in event recorders. As a result, automated methods of detecting these intermittent episodes have been explored. An episode-based classification system which removes the QRS complex is described by Stridh et al. (2001). Other techniques analyse RR interval dynamics to determine AF episodes; for instance the following authors (Dash et al., 2009; Hong-Wei et al., 2009; Lombardi et al., 2004; Maier et al. 2001 & Tateno & Glass, 2001) apply a mixture of time and frequency analysis techniques to RR intervals.
A smart system based on temporal features of RR intervals (Shouldice et al., 2004; Shouldice et al., 2007) for the classification of ECG blocks that contain AF is presented in this chapter.

3. Methods

This section discusses

1. PAF detection block based system
2. Mapping of expert annotations to a quantized signal suitable for the development of a block based system
3. Performance measures
4. Detailed discussion of linear discriminant analysis and an example.

3.1 Classification

A common task in medicine is to distinguish two groups of measurements or subjects. For example, one may wish to identify all people with a particular disease by combining results from blood tests, physical measurements, and symptoms. Such a task is called classification, and is typically the job of the human expert in the system. It is useful to automate the process of classification, particularly where large quantities of data or large numbers of subjects are required. For example, the FDA has recently approved a system (AutoPap from NeoPath) which analyses Pap smears automatically and returns a classification of positive, negative, or indeterminate.

The design of such automated systems has been, and continues to be an area of active research, particularly with increases in computational speed and the increasing use of information technology in medicine.
There are a large number of approaches for developing automated classification systems. Among the most popular are techniques based on discriminant analysis (linear and quadratic), k-means cluster analysis, neural networks, fuzzy logic, and support vector machines. The algorithm outlined here for detection of blocks consistent with PAF uses linear discriminant analysis (LDA), operating on a block by block based measure.

3.2 Block based system

A number of methods exist for processing real time data and then performing classification. For instance, in the case of a PAF classifier, one could analyze incoming QRS detections on a beat by beat basis and make some decision at a particular QRS point based on a small number of previous QRS points. However, in practice some form of windowing (perhaps incorporating many QRS points) is desirable to combat the effect of dropped or erroneous QRS detections, e.g., so as not to be unduly influenced by a profound change due to electromyographic artifact.
The algorithm outlined here uses the concept of per block processing. This involves buffering a certain length of ECG signal and associated QRS detection points and then making a classification decision; in other words, fixed length sequences of ECG beats are inspected. The output classification decision, whether it be PAF, NONPAF or undetermined (i.e. when no QRS points are available due to noise) provides a trigger to store the associated ECG signal to flash memory or not.
A block based system, for instance using non overlapping blocks containing 100 QRS detections, requires that an expert label be appended to each block for performance estimation purposes. This step involves the quantization of the expert annotations (e.g. such annotations might be at the start of a period of atrial fibrillation or at the return to normal sinus rhythm) to the block length. The original expert annotations are also stored on a per beat basis; this step facilitates the estimation of "beat based" performance measures.
Fig. 1 illustrates the segmentation of an expert signal into 5 non overlapping blocks. In this case, a block is tagged as PAF if at least half of the block (i.e. 50 beats) was determined by the expert to be PAF.

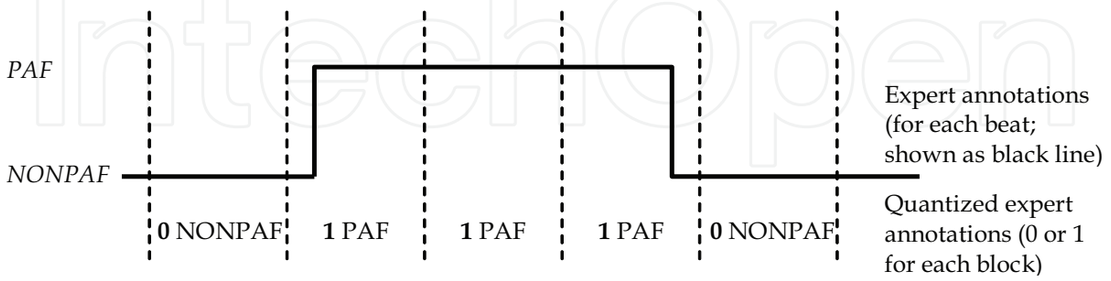


Fig. 1. The quantized expert signal (0 or 1 annotations, perhaps at half the block length) provide the quantized training information (reference) to compare with classifier output. Thus, block performance measures can be calculated from TP, TN, FP, FN figures. However, it should be noted that the inherent quantization will limit the upper bound of performance measures such as 'episode', 'duration' and on a 'beat' basis.

These expert annotations are then available for comparison with the classifier output and thus to generate performance figures. The training process involves the choice of certain characteristics (features) that describe how PAF-like behaviour manifests itself in the block of RR intervals (derived from the difference of the block QRS points). The actual features used in the BiancaMed system are discussed in Section 4; however, it may be useful at this point to inspect a flowchart indicating the main steps involved in the training and testing of the block system used (see Fig. 2).

One can also consider an improvement of the block based system whereby an overlap is introduced. As an example, for a 100 beat block system, an extreme case of overlapping is where each block is used to classify a single middle beat. The block is then shifted by one beat and another classification is performed. Obviously, this is a rather computationally expensive method of classifying on a per beat basis, although it does remove the possible quantization error of the expert signals. A system utilizing somewhat less overlap (say 50% overlap or a shift of 50 beats when a 100 beat block is considered) can be used to trade computational complexity for possible quantization error.

At this stage, it is useful to consider how the power of diagnostic tests can be assessed. Some of the most favoured measures used by clinicians are sensitivity and specificity. These can be explained by reference to Table 1. For a two-class problem, there is an associated table with four cells. There are two rows representing 'true' atrial fibrillation and 'false' non-AF (i.e. all other rhythms).

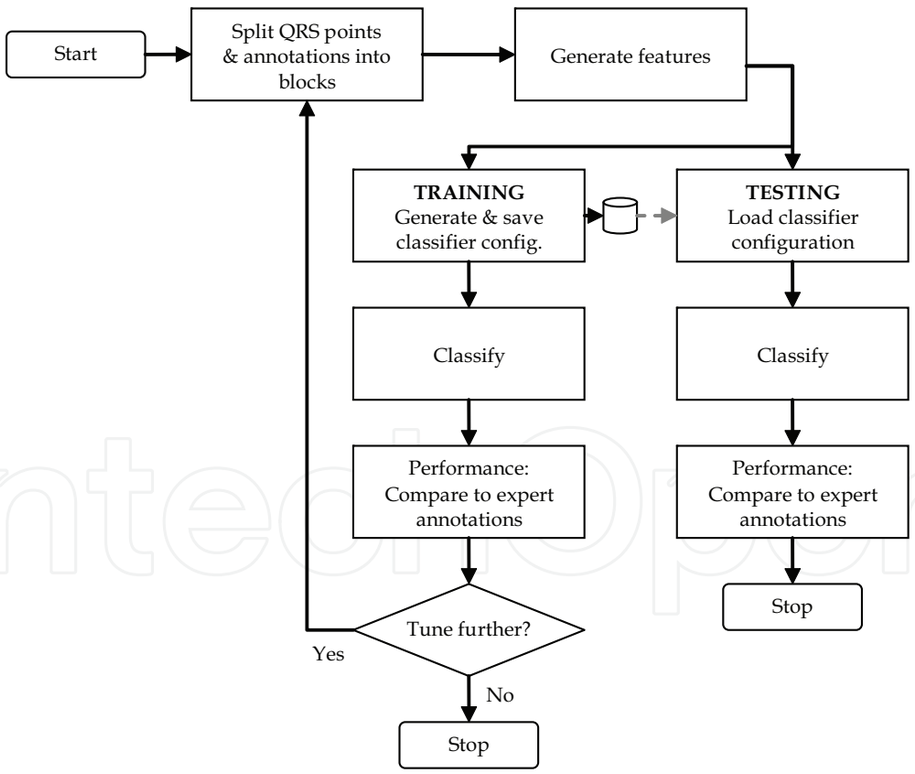


Fig. 2. Main steps involved in both the training and testing of a block based classification system utilizing expert annotations.

There are also two columns representing the assignment of each beat by the test. Note that in the case of a 100 beat block system, a 'PAF' classification leads to all of those 100 beats

being tagged as 'PAF'. Those beats that have been tagged as PAF by a human expert and are subsequently classified as PAF by the algorithm, are called 'true positive' (TP). Beats that are known to be PAF, but which the test classified as control are labelled as 'false negative' (FN). Beats which are NONPAF, but which the test flags as PAF are labelled 'false positive' (FP). Finally, beats which contain rhythms other than PAF, and which the test correctly labels as NONPAF are called 'true negative' (TN).

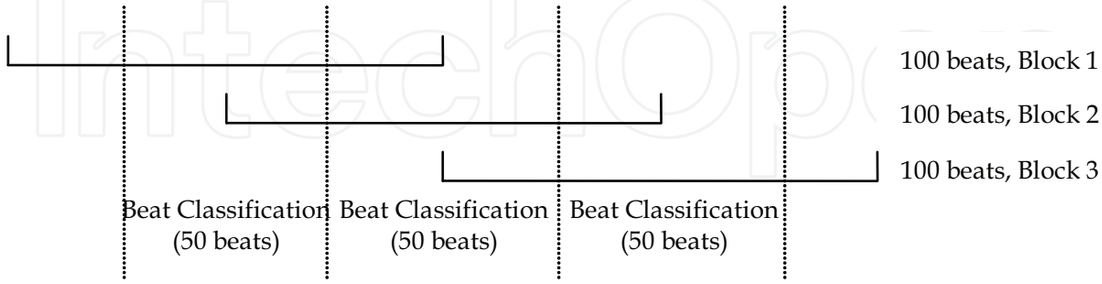


Fig. 3. If 50% overlap is used, each 100 beat block provides a classification of the middle 50 beats. Note: the start and end 50 beat segments should be padded or trimmed.

The sensitivity (*Se*) of a test is then defined as the percentage of PAF beats which are identified as PAF. The specificity (*Sp*) of the test is defined as the percentage of NONPAF beats which are identified (classified) as NONPAF. These may be expressed as

$$Se = \frac{TP}{TP + FN} \qquad Sp = \frac{TN}{TN + FP} \tag{1}$$

Note that it makes no sense to give either figure in isolation. Any diagnostic test can be made to have 100% sensitivity (by simply labelling everyone as having the disease), but such a test would have zero specificity, and would be of no practical benefit.

		EXPERT ANNOTATION	
		PAF	*NONPAF*
TEST RESULT	*PAF*	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
	NONPAF	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

Table 1. Definition of terms used in calculating sensitivity, specificity, positive predictive value, and negative predictive value.

Related measures are the Positive Predictive Value (*PPV*) and the Negative Predictive Value (*NPV*). These are defined respectively as the percentage of PAF-classified beats, which are expert tagged as PAF, and NONPAF-classified beats, which are actually NONPAF. The associated equations are

$$PPV = \frac{TP}{TP + FP} \qquad NPV = \frac{TN}{TN + FN} \tag{2}$$

Finally, the accuracy is given as the overall percentage of correctly classified beats

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In the context of auto-detection of AF using a device with a limited amount of flash memory, the classifier design must balance the need to keep the number of false positive (*FP*) detections as low as possible to avoid exhausting the ECG storage memory while remaining sufficiently sensitive (i.e. keep the number of *FN* detections small) so as to capture real PAF beats.

3.3 Alternative performance measures

As can be seen from the previous section, beat based performance measures can be generated from quantized blocks (e.g. containing 100 beats) of data. However, such measures do not give any indication of how many actual episodes (events) of PAF are captured (i.e. each PAF event can vary widely in length). Jager et al. (1991) propose alternative event based performance measures based on episode and duration, for the evaluation of transient ST elevation detectors. An episode measure is designed to provide information on how well episodes (rather than beats) are captured; a duration measure describes how reliably duration is captured. As an example, such episode and duration performance figures are available for the King of Hearts AF monitor.

Jager et al. (1991) note that there are four possible outcomes in which a detector is presented with an input with is either an event or non event – FP, FN, TP and TN (as per the beat measures). However, they note that in some detectors the concept of a non event cannot be counted (e.g., stating that a subject had four "non-events" of PAF during a nights sleep does not really mean much!) and thus the false negative (FN) figure is undefined. Therefore, they concentrate on obtaining sensitivity (Se, the fraction of events which are detected) and positive predictivity (PPV, the fraction of detections which are events). Furthermore, they provide two different methods of aggregating performance figures to account for the possibly small number of events in each subject dataset. The average statistics give each subject equal weighting; in contrast, the gross statistics give each event or detection equal weighting.

3.3.1 Episode measures

An episode of PAF is deemed to have been successfully detected (a "match") if the period of overlap between the expert annotated signal and predicted signal (i.e., the classifier output) is greater than 50% or when the period of overlap contains the extrema. Of course, the original expert annotations were at a higher resolution (arbitrarily chosen as 1 annotation per sec or 1 Hz based on the per-second expert markings) than the 30 sec quantized sequences. Therefore, the quantized classifier output must be upsampled for calculation of episode and duration measures; as an example, an output of 00010 will become a train of 90 'zeros' followed by 30 'ones' followed by 30 'zeros'.

An episode based sensitivity measure (*ESe*) is an estimate of the likelihood of detecting a PAF episode where *STP* is the number of matching PAF episodes, *FN* is the number of non matching episodes and (*STP*+*FN*) is the total number of episodes. For this measure, a "match" occurs when an expert annotated episode is overlapped by at least 50% of the predicted PAF signal. The defining reference annotations are based on the human expert markings and the detector signal is the quantized classifier output, upsampled to the same rate as the expert signal. The episode positive predictivity (*EPP*) is defined based on *PTP* (capturing the number of matching episodes) and *FP* (the number of non matching episodes). For this measure, a "match" occurs when a predicted episode is overlapped by at least 50% of the expert annotated PAF signal. N.B., this is the reverse of the case for a match when determining STP.

These measures are calculated using

$$ESe = \frac{STP}{STP + FN} \quad EPP = \frac{PTP}{PTP + FP} \tag{4}$$

3.3.2 Duration measures

The total duration of PAF correctly detected can be calculated as follows. The duration sensitivity (*DSe*) and duration positive predictivity (*DPP*) may be calculated using

$$DSe = \frac{\sum D \cap R}{\sum R} \quad DPP = \frac{\sum D \cap R}{\sum S} \tag{5}$$

where $\sum R$ represents the total time of expert determined PAF and $\sum D$ represents the total time of PAF determined by the predictor. $\sum D \cap R$ is the overlap (common region or intersection) between expert and predictor, calculated using a logical AND operation. In general, these duration performance figures might be expected to be broadly similar to the per beat figures.

3.4 Supervised classification using training data

When using a linear discriminant classifier (LDA), one must first derive the covariance matrix and class mean vectors (this is discussed in more detail in a later section). In general, this requires us to have access to a set of data whose class membership is known – i.e., access to expert annotated data. Such a set of data is called a "training set", and the overall approach is called "supervised learning" since the classifier is given some known outputs to begin with. As an aside, it is noted that there is much interest in unsupervised training where the classifier has no knowledge about the class memberships of any data, or perhaps a minimal set of knowledge such as the total number of classes present in the data. However, this more general class of classifier is not considered in this document.

One caveat with using supervised training of a classifier is that the classifier can become "over-trained"; that is it will assign a lot of importance to patterns which arose in the training data through chance alone. As an extreme example, given enough parameters one can draw an arbitrarily complex boundary region between the two classes of data in the training data which will achieve 100% accuracy in identifying class membership. This is shown in Fig. 4 (left) where a complex decision boundary between the two classes using 20 training data vectors is shown. It would appear that the test has 100% performance in all parameters. However, when it is tested on additional data (Fig. 4 (right)), it is clear that in fact the test makes a lot of errors. Therefore, in general one should be cautious in assessing the performance of a classifier on training data alone, since this will tend to bias the results upwards.

A better methodology is to develop the diagnostic test using available training data, and then assess its performance on an independent (test) set, which was never used in the design or training of the classifier. Ideally, the training data and the test data should both be sufficiently large to well represent the general population statistics (where population refers to the subjects likely to be classified using the test, not necessarily the general population). In the system outlined in this document, the population of interest will be subjects with suspected PAF episodes. The values of class mean vectors, and covariance matrices supplied by the authors represent the pool of training data that has currently been well characterized. These values can be altered as the test data set grows, or to deal with specific sub-populations with different statistical distributions.

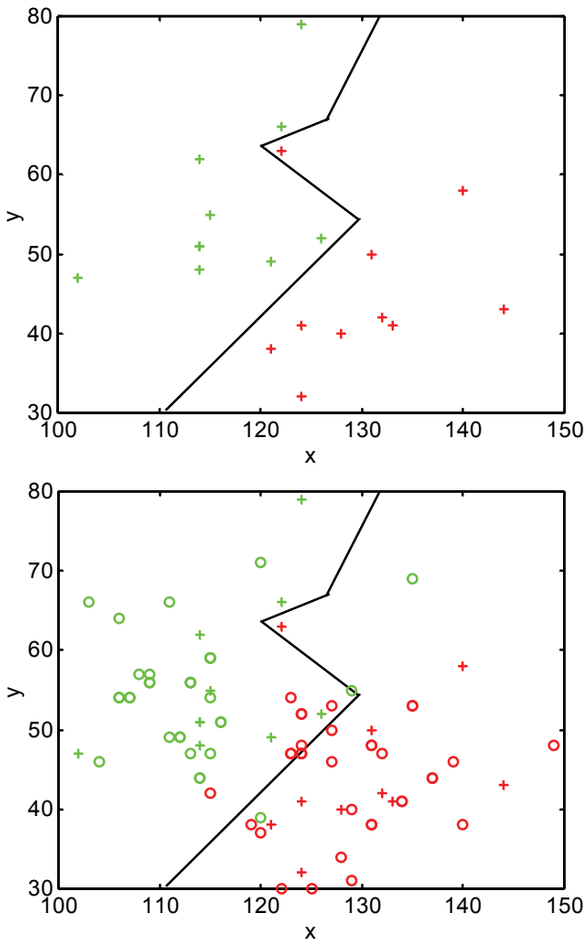


Fig. 4. A relatively complex decision rule to separate two classes with 100% accuracy (upper). Classification performance of the complex decision rule when independent data is tested. The decision rule which produced 100% accuracy for the training data now produces a significant number of errors (lower). This is an example of overtraining.

3.5 Example: Pattern recognition using discriminant analysis

Imagine one is devising a test to identify people with a (hypothetical) disease called "bilurosis". The clinical measurements from each subject consist of two blood measurements: the number of red blood cells (RBC) per unit volume (denoted as x in the following discussion) and the number of lymphocytes per unit volume (denoted as y in the following discussion). For convenience, it will be assumed that there are an equal number (n=40) of control subjects and subjects with the disease. However, in practice this is not likely to be the case. Table 2 gives a truncated view of the measured values from our experimental test. The task is to find an automated means of discriminating control subjects and disease subjects.

In an ideal world, one might hope to make a single measurement that would correctly distinguish the two groups. Initially, one might consider the information offered by the x and y measurements individually. Fig. 5 shows the individual histograms of the x and y variables. This histogram reveals that the distributions of x and y differ for the control and disease subjects, which implies that they are useful diagnostic measurements.

CONTROL SUBJECTS		DISEASE SUBJECTS	
RBC (x)	**LBC (y)**	**RBC (x)**	**LBC (y)**
114	51	128	40
126	52	144	43
…	…	…	…
111.4 (9.8)	**54.4 (9.0)**	**128.4 (7.2)**	**42.1 (9.1)**

Table 2. Data values used for Example 1 (note – these have been truncated at two rows for display here). The values in the last row are mean and standard deviation.

The histograms can allow one to devise a simple diagnostic test. For instance, it may be noted that the mean x value (111.4) in the control subjects is lower than that of the disease subjects (128.4). Therefore it is possible to make a decision rule that everyone with x values lower than or equal to 120 is a control subject. However, if the x value alone is used as a means of diagnosis, it is noted that 6 control subjects have values greater than this, and 4 disease subjects have lower values, so these subjects would be wrongly classified.
The corresponding performance table for the test (x>120 is disease) in Table 3. The following performance figures are derived:

Sensitivity	90%	Specificity	85%
PPV	85.7%	NPV	84.5%
Accuracy	87.5%		

		REALITY	
		DISEASE	*CONTROL*
TEST RESULT	*DISEASE*	36	6
	CONTROL	4	34

Table 3. Values of TP, FP, FN, and TN for the simple decision rule x>120

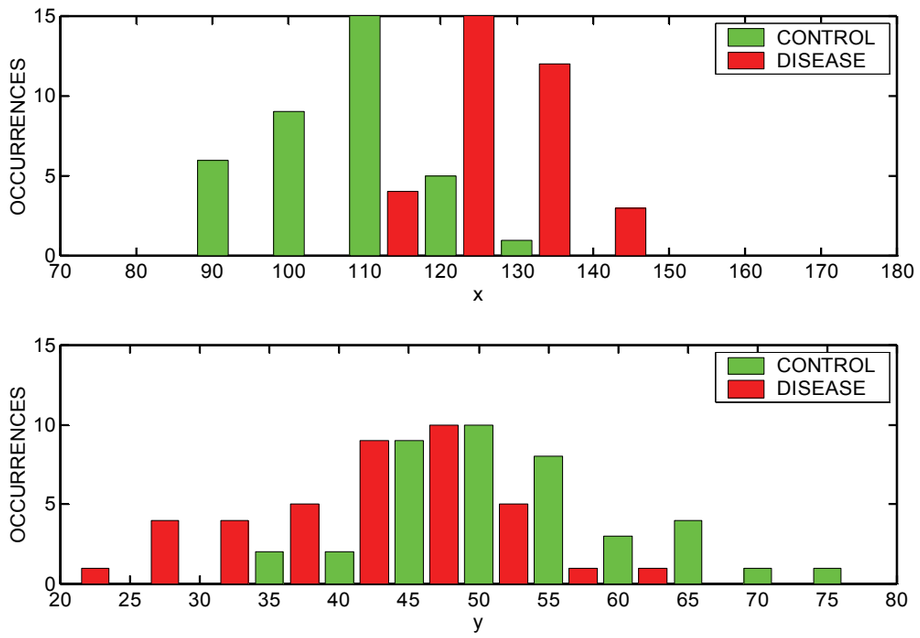


Fig. 5. Histograms of the x and y values for the data contained in Table 2.

It is not hard to see that the exact performance figures of a test are highly dependent on the exact decision rule used. For example, if the decision rule was changed to "x>114 is disease", then every disease subject would be classified as having the disease, but unfortunately 19 control subjects would also be classed as having the disease. In such a case, the performance figures would be:

Sensitivity	100%	Specificity	52.5%
PPV	67.8%	NPV	100%
Accuracy	76.3%		

An ideal diagnostic test would have all of these values equal to 100%. Intuitively, one might suspect that a better diagnostic test would be obtained if information from both the x and y measurements was used. One means of devising such a test is linear discriminant analysis.

The concept of LDA is easy to understand. Fig. 6 shows a scatter plot of the measured x and y data for the control and disease classes. It is easy to see that the two classes fall into two clusters that are well separated. The one-parameter decision rule attempts to find these two clusters by drawing horizontal and vertical decision lines; a better cluster division can be obtained by drawing more general lines of the form $y=mx+c$ as shown in the diagram.

Linear discriminant analysis answers the question of how to find the parameters of the line (or more generally, a hyperplane in n-dimensions). It is referred to as linear since the decision rules it produces will be based on linear combinations of the measured variables.

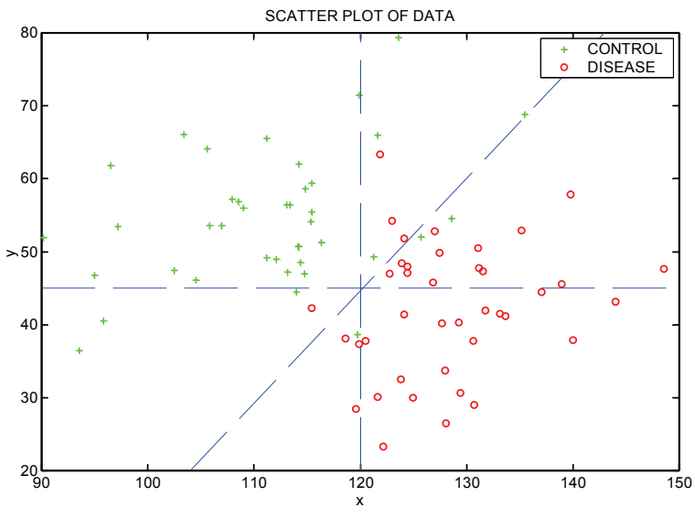


Fig. 6. Scatter plot of the x and y data used in Example 1.

A general derivation of the theory of linear discriminant analysis is beyond the scope of this chapter, but an approach will be used here based on work by Fisher (Fisher, 1936).

Conceptually, one wishes to find a line (or more generally a plane or hyperplane) which optimally separates the two classes of data. This is equivalent to finding the direction of projection onto a line, such that the projected values are well separated.

This is perhaps best explained pictorially, as seen in Fig. 7. In this figure, the two dimensional data points x=(x,y) are projected onto a line. Their value on the line is called their discriminant value. These discriminant values will have their own histograms/ probability density functions (sketched as green [or light grey] and red [or dark grey] pdfs on the discriminant line) for each class. The best choice of projection direction will separate out the two discriminant histograms to the maximum extent possible.

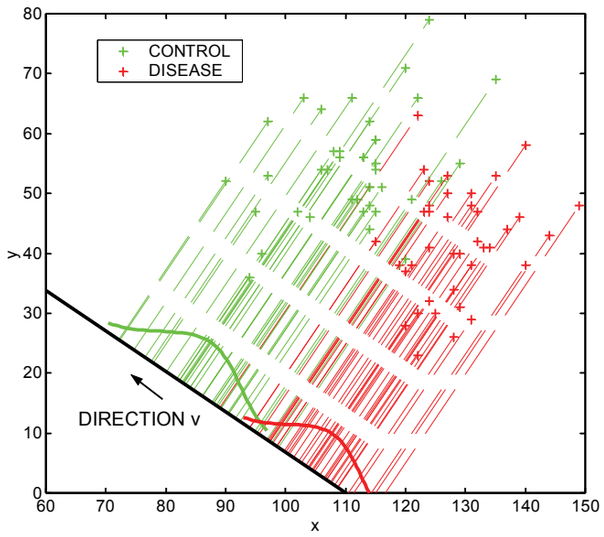


Fig. 7. Projection of the data used in Example 1 onto a vector of direction v, in order to produce a discriminant value for each data point.

This discriminant value will take on a range of (scalar) values, so the decision rule will be based on comparing the discriminant value against this threshold. A reasonable threshold is the discriminant value produced by the point which is equidistant between the two class mean vectors:

$$\frac{(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B)}{2} \tag{6}$$

Therefore, the final decision rule to classify an arbitrary data vector x will be described by

$$(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \quad \begin{matrix} \text{choose classA} \\ > \\ < \\ \text{choose classB} \end{matrix} \quad \frac{(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_A + \boldsymbol{\mu}_B)}{2} \tag{7}$$

If equality is achieved, the test is indeterminate.
One can apply this LDA rule to all 80 data points in the set to produce the following table of results:

		REALITY	
		DISEASE	*CONTROL*
TEST RESULT	*DISEASE*	37	4
	CONTROL	3	36

Table 4. Values of TP, FP, FN, and TN for the LDA rule in Eq. (7)

The corresponding performance values are:

Sensitivity	92.5%	Specificity	90%
PPV	90.2%	NPV	92.3%
Accuracy	91.25%		

These figures are an improvement over the single feature classifiers considered previously.
The LDA decision rule described is termed Fisher's approach. An alternative derivation of a LDA classifier can be obtained using Bayes's rule. It uses the same assumptions about features being normally distributed, and having a common covariance matrix.
It yields a set of discriminant values q_k, one for each class, as follows:

$$q_k = \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \tfrac{1}{2} \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\pi_k) \tag{8}$$

where π_k is the prior probability of class k. In this case, the class with the highest discriminant value is chosen as the output class.

3.6 Modifications to linear discriminant analysis

The classification rule given above does not account for any knowledge one may have about the likely distribution of the classes prior to conducting the test (known as *the prior probabilities*). It is not unreasonable that knowledge of these distributions can be used to help the classifier (i.e., if 99% of people undergoing the test are normal, and one wishes the specificity to be good, then it may make sense to move the threshold for the discriminant value closer to the normal class). In other words, moving the discriminant value towards the normal class will result in higher specificity and lower sensitivity. As a specific example, in the above test one might set the prior probabilities of the two possible outcomes (denoted as π_A and π_B) of the test to 0.5, as based on the data available, there are equal numbers of control and disease subjects.
Alternatively, one may know from other experience that approximately 5% of the general population has the hypothetical disease "bilurosis" and hence set the prior probability of the disease outcome to 0.05 and of a normal outcome to 0.95. Knowledge or estimates of the prior probabilities, therefore, can be used to modify the discrimination rule as follows:

$$(\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} \quad \begin{array}{l} \text{choose class A} \\ > \\ < \\ \text{choose class B} \end{array} \quad \frac{(\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_\mathbf{A} + \boldsymbol{\mu}_\mathbf{B})}{2} - \log\left(\frac{\pi_B}{1-\pi_B}\right) \tag{9}$$

Note that $\pi_A = 1 - \pi_B$. Finally, since the discriminant value is not simply a hard decision ("yes" vs. "no"), but rather a number, it may be used to estimate the confidence of the decision. Alternatively, this can be thought of as the *posterior probability* of a test vector belonging to a specific class.
A Bayesian formulation of the linear discriminant analysis test allows a convenient means to capture the confidence of our decision. It can be shown that the posterior probability of a test vector being in class k is given by

$$P_k = \frac{\exp(q_k)}{\sum_k \exp(q_k)} \tag{10}$$

where q_k is its corresponding discriminant value for membership in set k. If any of the P_ks are close to one, then the classifier is very confident in assigning it to class k.

3.7 Techniques for improving classification performance – conversion to gaussianity

An important assumption in the derivation of a linear discriminant classifier is that the features have a Gaussian (normal) distribution. In other words, if the feature is x, then the probability density function (pdf) of x is given by

$$p_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right) \tag{11}$$

This is an assumption that is often violated by features which have useful classification information, and will lead to reduced performance of the classifier. For example, one feature which is used by the BiancaMed PAF detection system is based on the range of spread in a block of RR intervals. The resulting histogram can be more closely made to approximate a normal distribution by the application of the square root transformation.

Other transformations that are commonly used are the log transformation or the cube root. Features which are likely to be non-Gaussian include power measurements, envelope measurements, and counts (e.g., counting the number of premature ventricular contractions within a window).

4. Implementation

4.1 Block based PAF detector

Atrial fibrillation is known to be capable of causing ventricular depolarization timing changes (filtering action of the AV node) that can give rise to an irregularly irregular RR interval (difference between successive QRS complexes). As such, it is this type of RR behavior that is desirable to discriminate using suitable temporal features.

A block diagram showing input, preprocessing, feature generation, classification and output for a block based detector is provided in Fig. 8.

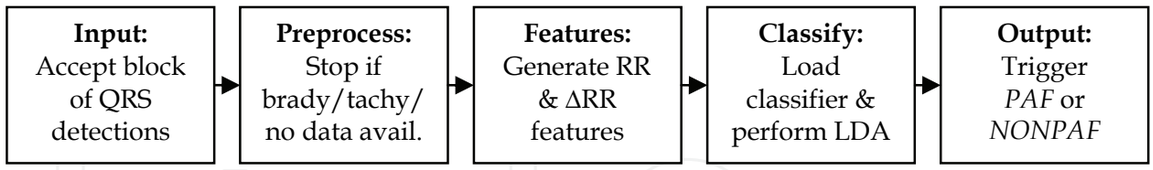


Fig. 8. Block based classification system

4.2 Deriving features using the RR series

The discussion in Section 3.4 on linear discriminant analysis describes how to derive a linear classifier from a set of features extracted from the data. The real key to a successful diagnostic test, however, is in identifying features which provide classification information. This is an art rather than a science, and in general features are found through a combination of a systematic approach and experience of the classifier designer. There are some techniques for finding optimal subsets of features, but in general it is the job of the researcher to seek out the most likely candidate features.

A first step towards selecting features for the final classifier was the formation of class-dependent histograms of each feature (a selection of these are shown in Fig. 9). Inspection of the histograms allowed features that had well separated means and approximately Gaussian distribution to be chosen.

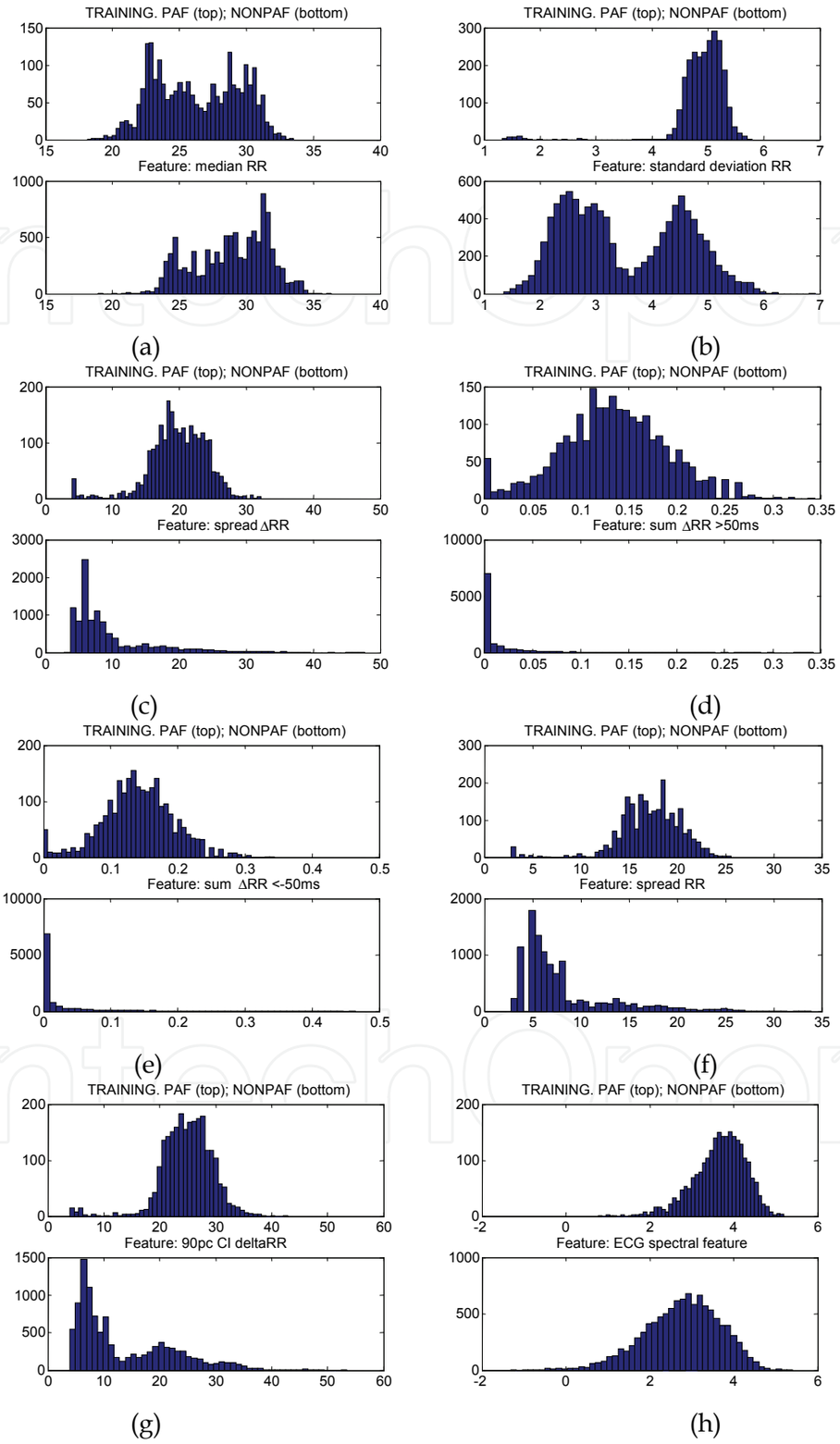


Fig. 9. Histograms of candidate features from training set.

Some features of RR and ΔRR (i.e., delta or difference between successive RR intervals) considered were:

- Mean (Huang et al., 1998; Lombardi et al., 2004)
- Root Mean Square (RMS)
- Median
- Standard deviation (Capucci et al., 1992)
- Interquartile range
- Other quantile ranges
- First 4 serial correlation coefficients (de Chazal & Heneghan, 2001)
- Power spectral density estimates (Bettoni et al., 2002; Herweg et al., 1998)
- PNN50, %NN >50 ms different that previous interval (Maier et al., 2001)
- PNN50a, %NN >50 ms longer than previous interval
- PNN50b, %NN >50ms shorter than previous interval
- Skewness
- Kurtosis
- Allan factor
- Approximate entropy

Note that power spectral (frequency based) features were excluded from the final feature set due to extra computational complexity. Also, in the case of multiple features that measure similar basic characteristics of the signal (e.g. trimmed mean, median and mean are all measures of central tendency) only one feature was chosen. The bimodal distribution of the log standard deviation of RR intervals (NONPAF) in Fig. 9(b) makes a Gaussian fit difficult. It should be noted that features which ignore extreme values (whether high or low) are likely to be more robust in a Holter based system where a certain level of QRS misdetection is possible. After analysis of the features, five were chosen that provide an insight into the spread, central tendency and count of thresholded differences.

A first preprocessing step prior to calculating ECG features is to verify that the sequence of QRS detections is within certain bounds. For instance, if the block beats per minute (bpm) value is greater than a particular threshold (suggesting tachycardia) or below a certain threshold (suggesting bradycardia), then the brady/tachy function should already have been activated; thus, it does not make sense to go through the classification process.

The RR interval series is used to generate the classification features. Since the classification is performed on 100 beat blocks, this number of QRS detections must be buffered along with the associated ECG signal. Features are generated independently for each of the blocks, whether overlapping or not.

The time-domain features were calculated as follows. The QRS intervals series, denoted by $QRS[n]$, and assuming n intervals contained within each block, the RR ($RR[n\text{-}1]$) and difference in RR ($\Delta RR[n\text{-}2]$) series were generated.

Subsequently, the following time domain measures were calculated for each block:

1. Trimmed mean (i.e. deletion of outliers) of $RR[n\text{-}1]$ at 10%.
2. Spread in $\Delta RR[n\text{-}2]$ from 5% to 95%.
3. Spread in $RR[n\text{-}1]$ from 5% to 95%.
4. Sum of $\Delta RR[n\text{-}2]$values greater than 50 ms (scaled to operating sampling rate) divided by the total number of $\Delta RR[n\text{-}2]$ values within the block, i.e. ($n\text{-}2$). Upper bound set to twice the trimmed mean of $RR[n\text{-}1]$.
5. Sum of $\Delta RR[n\text{-}2]$values less than –50 ms (scaled to operating sampling rate) divided by the total number of $\Delta RR[n\text{-}2]$ values within the block, i.e. ($n\text{-}2$). Upper bound set to twice the trimmed mean of $RR[n\text{-}1]$.

Transformations were applied to these features where appropriate. It should be noted that each of the five features operates on a subset of the block RR sequence; therefore, they might be expected to be quite robust to outlying (and perhaps erroneous) values.

4.3 Training and test data

The classifier parameters used in the block detector must first be generated. This is achieved using supervised learning on available annotated data, specifically a "training" subset. A separate "test" (i.e., independent and withheld) subset is kept separately from the "training" data; these "test" data are then used for performance validation.
The class mean vectors were calculated using

$$\boldsymbol{\mu}_{\mathbf{A}} = \frac{1}{N_A}\sum_{k=1}^{N_A}\mathbf{x}_{\mathbf{k,A}} \qquad \boldsymbol{\mu}_{B} = \frac{1}{N_B}\sum_{k=1}^{N_A}\mathbf{x}_{\mathbf{k},B} \tag{12}$$

where A denotes the normal class, B is the PAF class, N_A is the number of training vectors in class A, and $\mathbf{x}_{\mathbf{k,A}}$ is the kth vector from Class A. The common covariance matrix was calculated using

$$\Sigma = \frac{1}{N_A + N_B - 2}\left(\sum_{k=1}^{N_A}\left(\mathbf{x}_{\mathbf{k,A}} - \boldsymbol{\mu}_{\mathbf{A}}\right)^T\left(\mathbf{x}_{\mathbf{k,A}} - \boldsymbol{\mu}_{\mathbf{A}}\right) + \sum_{k=1}^{N_B}\left(\mathbf{x}_{\mathbf{k,B}} - \boldsymbol{\mu}_{\mathbf{B}}\right)^T\left(\mathbf{x}_{\mathbf{k,B}} - \boldsymbol{\mu}_{\mathbf{B}}\right)\right) \tag{13}$$

Using the calculated values of $\boldsymbol{\mu}_K, \Sigma$, an incoming feature $\mathbf{x}$ may be classified by calculating the discriminant value:

$$q = \left(\boldsymbol{\mu}_B - \boldsymbol{\mu}_A\right)^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{x} - \tfrac{1}{2}\left(\boldsymbol{\mu}_B - \boldsymbol{\mu}_A\right)^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\left(\boldsymbol{\mu}_B + \boldsymbol{\mu}_A\right) + \log\left(\frac{\pi_B}{1-\pi_B}\right) \tag{14}$$

and posterior probabilities for class membership calculated using

$$P_A = 1 - P_B$$
$$P_B = \frac{\exp(y)}{1+\exp(y)}$$

The databases available for the training and testing of the algorithm are available online through the Physionet resource (Goldberger et al., 2000). These databases, containing expert annotated atrial fibrillation segments, are as follows:

1. **Database AFDB** is the MIT-BIH Atrial Fibrillation Database and includes 25 long-term ECG recordings of human subjects with atrial fibrillation (mostly paroxysmal). Of these datasets, 2 have no original ECG signal for validation with the QRS detector.
2. **Database MITDB** is the MIT-BIH Arrhythmia Database contains 48 half-hour excerpts of two-channel ambulatory ECG recordings, obtained from 47 subjects. 8 of the datasets contain periods of atrial fibrillation.
3. **Database NSRDB** is the MIT-BIH Normal Sinus Rhythm Database. This database includes 18 long-term ECG recordings of subjects. Subjects included in this database were found to have had no significant arrhythmias; they include 5 men, aged 26 to 45, and 13 women, aged 20 to 50.

4. **Database NSR2DB** is the Normal Sinus Rhythm RR Interval Database. This database includes beat annotation files for 54 long-term ECG recordings of subjects in normal sinus rhythm (30 men, aged 28.5 to 76, and 24 women, aged 58 to 73).

A PAF episode was defined to begin at either an atrial fibrillation "AFIB" or atrial flutter "AFL" annotation. Any other rhythm annotation was a non-event (i.e. NONPAF). An assumption is made here that the target hardware for the PAF classification samples ECG at 128 Hz. Due to the fact that the Physionet databases were sampled at greater than the desired sampling rate (Fs) of 128 Hz, the initial classifier development was performed by dividing the supplied database QRS detections by their respective Fs values, multiplying by 128, rounding to the nearest integer, and then scaling to milliseconds. This is an attempt to provide a rough simulation of the increased sampling (but not quantization) error that is expected in hardware. For more realistic "real world" performance estimates, it is intended that the ECG data available as part of the Physionet datasets be resampled and quantized to the expected input values and then run through the target hardware's QRS detector. This step should serve to highlight the influence (if any) of QRS misdetections/lead dropout on classification performance.

The 25 AFDB datasets were used to select the classifier features. This was achieved by using leave-one-out cross fold validation to provide a measure of generalized error and then tuning features to maximize accuracy. The leave-one-out cross-validation scheme works as follows: one subject's data is withheld and classifier parameters are obtained by using all the other available training data. The classification performance of the resulting classifier can then be estimated on the one withheld set. Overall estimated classification performance can then be obtained by averaging the results over all possible withheld sets.

When the (empirically determined) 5 features (as listed previously) were determined, the final classifier parameters were stored based on all of the AFDB datasets. Testing can then be performed on the same database, although some bias is to be expected. The other three databases, MITDB, NSRDB and NSR2DB represent independent test data.

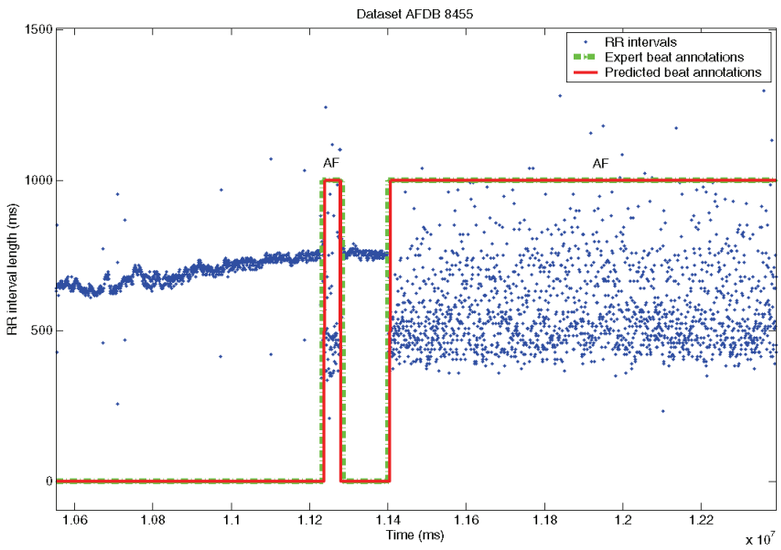


Fig. 10. BiancaMed algorithm processes multiple measures from the RR series and accurately predicts PAF events. This record excerpt is part of the AFDB.

5. Test results: Performance of the system

The classifier parameters saved during the training process were then used to classify each of the withheld test datasets from MITDB (48 records, 8 with AF), NSRDB (18 records) and NSR2DB (54 records). ROC plots of the performance on the AFDB and MITDB are presented in Fig. 11.

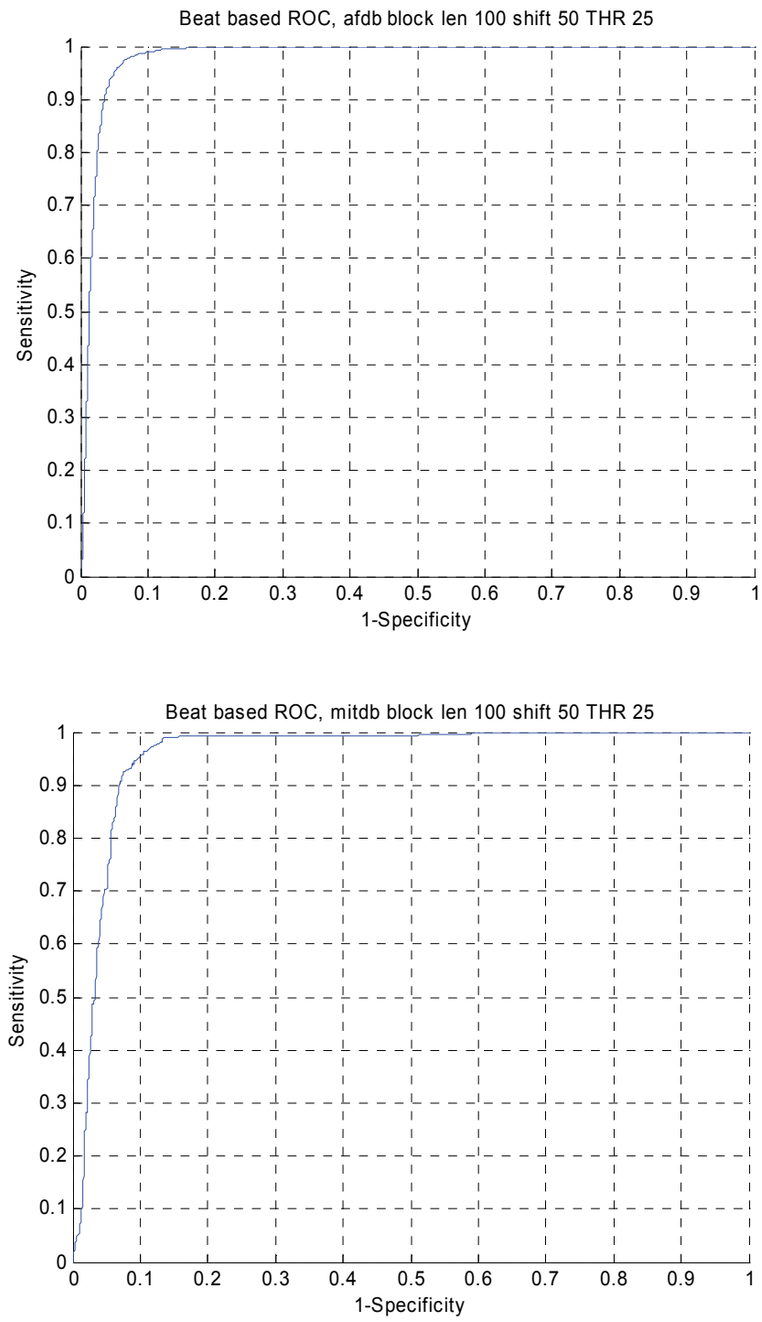


Fig. 11. Receiver operating characteristics: AFDB (left) and MITDB (right) on a per beat basis

Database	Aggregation	ESe	EPP	DSe	DPP
MITDB	Gross	81	72	74	54
	Average	88	70	81	55
AFDB	Gross	81	54	75	59
	Average	79	78	80	71

Table 5. Published 'King of Hearts Express AF' (c.f. 'Physician's Operation Manual') database test results including all events (all figures are percentages)

Database	Beat Se	Beat Sp	Beat Acc
MITDB	88	94	93
AFDB	94	97	96* estimated

Table 6. Published 'Tateno and Glass' database test results (beat basis)

Database	Aggregation	ESe	EPP	DSe	DPP
MITDB	Gross	81	90	93	89
	Average	88	88	86	87
AFDB	Gross	71	80	92	95
	Average	76	67	83	80

Table 7. BiancaMed system, 100 beat epochs with 50% overlap, episode and duration figures (compared to expert beat annotations) tested on subjects with AF episodes (i.e. all AFDB records and 8 MITDB records).

An alternate method of measuring the performance is on a beat-by-beat basis. These measures allow the determination of the expected number of false triggers on recordings free of AF such as the NSRDB, NSR2DB and those records with other arrhythmias in MITDB). Table 8 shows results for the system processing epochs containing 100 beats.

Database	Beat Se	Beat Sp	Beat Acc
MITDB	93	92	92
AFDB	92	96	94
NSRDB	--	100*	100
NSR2DB	--	100**	100

[*=1,600 FP beats out of total 1,582,715 beats; **=10,150 FP beats out of total 5,258,870 beats]

Table 8. BiancaMed system, 100 beat epochs with 50% overlap, beat figures (compared to expert beat annotations). All database records analyzed.

6. Conclusion

On a beat-by-beat basis the systems boasts high sensitivity and specificity on the MITDB (arrhythmia) and AFDB (atrial fibrillation) datasets as well as a low false trigger rate on the normal sinus rhythm records in NSRDB and NSR2DB.

As can be seen from the gross and average sensitivity and positive predictivity figures the PAF detector described in this chapter does, on average, perform well at discriminating patterns consistent with atrial fibrillation in the withheld test set.

However, it must be noted that other rhythms (whether they be arrhythmias or arising during periods that would be classed as "normal") that give rise to irregular RR interval times may also trigger the detector. Runs of ventricular premature contractions (PVCs) can cause false triggering. In addition, if the QRS detector incorrectly detects noise/artifact or T waves as QRS complexes, the false impression of "irregularly irregular" RR intervals may cause the algorithm to incorrectly trigger.

The BiancaMed PAF classification system is competitive with the published 'King of Hearts' episode and duration figures. On a beat-by-beat basis (Table 5.3) the systems boasts high sensitivity and specificity on the MITDB (arrhythmia) and AFDB (atrial fibrillation) datasets as well as a low false trigger rate on the normal sinus rhythm records in NSRDB and NSR2DB.

The episode, duration, and block accuracies of the system indicate that it performs very well in detecting PAF, and also at achieving a low false detection rate in normal subjects. As the method described only relies on inter-heartbeat intervals and low complexity time-domain features it is well suited to power-dependent applications such as ECG event recorders. It is also suited to other sensors of heart-rate, such as photoplethysmogram and ballistocardiogram.

The system outlined in this chapter has clear clinical relevance as an enabler for long term automatic capture of rare or asymptomatic paroxysmal atrial fibrillation events which might otherwise be missed by short Holter recordings or manual event recorders.

7. References

Benjamin, E.J.; Wolf, P.A.; D'Agostino, R.B.; Silbershatz, H.; Kannel, W.B. & Levy, D. (1998). Impact of atrial fibrillation on the risk of death. *Circulation*, Vol.98, No.10, pp. 946–952.

Bettoni, M. & Zimmermann, M. (2002). Autonomic tone variations before the onset of paroxysmal atrial fibrillation. *Circulation*. Vol.105, No.23, pp.2753-9.

Capucci, A.; Santarelli, A.; Boriani, G. & Magnani, B. (1992). Atrial premature beats coupling interval determines lone paroxysmal atrial fibrillation onset. *Int J Cardiol.* Vol. 36, No.1, pp.87-93.

Dash, S.; Chon, K.H.; Lu, S. & Raeder, E.A. (2009). Automatic Real Time Detection of Atrial Fibrillation. *Annals of Biomedical Engineering*. Vol.37, No.9, pp.1701-1709.

de Chazal, P. & Heneghan, C. (2001). Automated Assessment of Atrial Fibrillation, *Computers in Cardiology*, Vol.28, pp.117-120.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Vol.7, pp.179–88.

Gaillard, N.; Deltour, S. et al. (2010). Detection of paroxysmal atrial fibrillation with transtelephonic EKG in TIA or stroke patients. *Neurology*. Vol.74, No.21, pp.1666-70.

Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.Ch.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K. & Stanley, H.E. (2000). PhysioBank,

PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation.* Vol.101,No.23,pp.e215-e220.

Herweg, B.; Dalal, P.; Nagy, B. & Schweitzer, P. (1998). Power spectral analysis of heart period variability of preceding sinus rhythm before initiation of paroxysmal atrial fibrillation. *Am J Cardiol.* Vol.82, No.7, pp.869-74.

Huang, J.L.; Wen, Z.C.; Lee, W.L.; Chang, M.S. & Chen, S.A. (1998). Changes of autonomic tone before the onset of paroxysmal atrial fibrillation. *Int J Cardiol*. Vol.66, No.3, pp.275-83.

Jager, F.; Moody, G.B.; Taddei, A. & Mark, R.G. (1991). Performance measures for algorithms to detect transient ischemic ST segment changes. *Computers in Cardiology*, Los Alamitos. pp.369–372.

Kannel, W.B.; Abbott, R.D.; Savage, D.D.; McNamara, P.M. (1982) Epidemiologic features of chronic atrial fibrillation: the Framingham study. *N Engl J Med.* Vol.306,pp.1018-22.

Kirchhof, P.; Bax, J. et al. (2009). Early and comprehensive management of atrial fibrillation: proceedings from the 2nd AFNET/EHRA consensus conference on atrial fibrillation entitled 'research perspectives in atrial fibrillation'. *Europace*. Vol.11, No.7, pp.860-85.

Liao, J.; Khalid, Z.; Scallan, C.; Morillo, C. & O'Donnell, M. (2007). Noninvasive cardiac monitoring for detecting paroxysmal atrial fibrillation or flutter after acute ischemic stroke: a systematic review. *Stroke*. Vol.38,No.11,pp.2935-40.

Lombardi, F.; Tarricone, D.; Tundo, F.; Colombo, F.; Belletti, S. & Fiorentini, C. (2004). Autonomic nervous system and paroxysmal atrial fibrillation: a study based on the analysis of RR interval changes before, during and after paroxysmal atrial fibrillation. *Eur Heart J*. Vol.25,No.14,pp.:1242-8.

Hong-Wei, L.; Ying, S.; Min, L.; Pi-Ding, L. & Zheng, Z. (2009). A probability density function method for detecting atrial fibrillation using R–R intervals. *Medical Engineering & Physics*. Vol.31,No,1,pp.116-23.

Maier, C.; Bauch, M. & Dickhaus, H. (2001). Screening and prediction of paroxysmal atrial fibrillation by analysis of heart rate variability parameters. *Computers in Cardiology*. Vol 28,pp.129-32.

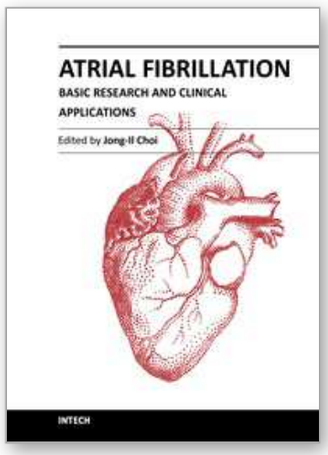
Roche, F.; Gaspoz, J.M.; Da Costa, A.; Isaaz, K.; et. Al (2002). Frequent and prolonged asymptomatic episodes of paroxysmal atrial fibrillation revealed by automatic long-term event recorders in patients with a negative 24-hour Holter. *Pacing Clin Electrophysiol.* Vol.25, No.11,pp.1587-93.

Shouldice, R.B.; O'Brien, L.M.; O'Brien, C.; de Chazal, P.; Gozal, D. & Heneghan, C. (2004). Detection of obstructive sleep apnea in pediatric subjects using surface lead electrocardiogram features. *Sleep*. Vol.27,No.4,pp.784-92.

Shouldice, R.B.; Heneghan, C. & de Chazal, P. (2007). Automated detection of paroxysmal atrial fibrillation from inter-heartbeat intervals. *Conf Proc IEEE Eng Med Biol Soc*. 2007, pp.686-9.

Stridh, M. & Sornmo, L. (2001). Spatiotemporal QRST cancellation techniques for analysis of atrial fibrillation. *IEEE Trans on Biomed Eng*, Vol.48, pp.105-111.

Tateno, K. & Glass, L. (2001). Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and deltaRR intervals. *Med Biol Eng Comput.*, Vol.139; pp.664-71.



Atrial Fibrillation - Basic Research and Clinical Applications

Edited by Prof. Jong-Il Choi

ISBN 978-953-307-399-6
Hard cover, 414 pages
Publisher InTech
Published online 11, January, 2012
Published in print edition January, 2012

Atrial Fibrillation-Basic Research and Clinical Applications is designed to provide a comprehensive review and to introduce outstanding and novel researches. This book contains 22 polished chapters and consists of five sections: 1. Basic mechanisms of initiation and maintenance of atrial fibrillation and its pathophysiology, 2. Mapping of atrial fibrillation and novel methods of signal detection. 3. Clinical prognostic predictors of atrial fibrillation and remodeling, 4. Systemic reviews of catheter-based/surgical treatment and novel targets for treatment of atrial fibrillation and 5. Atrial fibrillation in specific conditions and its complications. Each chapter updates the knowledge of atrial fibrillation, providing state-of-the art for not only scientists and clinicians who are interested in electrophysiology, but also general cardiologists.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen