An evaluation of the generalisability and applicability of the PhysioNet electrocardiogram (ECG) repository as test cases for ECG-based biometrics

Manal M. Tantawi

Faculty of Computer and Information Sciences,
Ain Shams University,
Cairo, Egypt
Email: manalmt@hotmail.com

Kenneth Revett*

Faculty of Informatics and Computer Science,
The British University in Egypt,
El Sherouk City, Egypt
Email: ken.revett@bue.edu.eg
*Corresponding author

Mohammed Fahmy Tolba and
Abdel-Badeeh Salem

Faculty of Computer and Information Sciences,
Ain Shams University,
Cairo, Egypt
Email: fahmytolba@gmail.com
Email: abmsalem@yahoo.com

Abstract: The PhysioNet is a very popular internet-based ECG repository which provides open access to a variety of ECG datasets. The data is collected from subjects within a medical framework, with the intention of acquiring clinically relevant information from patients. Because of the convenience afforded by the internet, literally thousands of ECG records can be downloaded and used for non-medical purposes, such as biometrics. The purpose of this study was to evaluate the applicability and/or suitability of the PhysioNet ECG data for deployment within biometrics. The needs and mindset of a clinician may be quite different from that of a security engineer. This paper therefore attempts to provide a preliminary examination of the PhysioNet ECG data repository along these dimensions, emphasising the need to create methodologies in the context of biometrics that not only take these considerations into account, but integrates them into the biometric methodology.

Keywords: authentication; behavioural biometrics; biometric benchmarks; cognitive biometrics; ECG-based biometrics; equal error rate; false acceptance rate; false rejection rate; fiducial-based feature extraction; PhysioNet ECG repository.

Reference to this paper should be made as follows: Tantawi, M.M., Revett, K., Tolba, M.F. and Salem, A. (2012) ‘An evaluation of the generalisability and applicability of the PhysioNet electrocardiogram (ECG) repository as test cases for ECG-based biometrics’, *Int. J. Cognitive Biometrics*, Vol. 1, No. 1, pp.66–97.

Biographical notes: Manal M. Tantawi is currently a PhD student at the Ain Shams University, Cairo, Egypt. Her PhD dissertation focuses on the role of emotion on ECG-based biometrics. Her research interests include machine learning, biometrics, and ECG-based technologies. She is currently serving as an Assistant Lecturer at the Ain Shams University, where she teaches in subjects such as statistics and machine learning algorithms.

Kenneth Revett is currently actively engaged in research focused in the development of the field of cognitive biometrics, which relies on the deployment of biosignals (EEG, ECG, EDR) for person authentication. He published the first textbook dedicated to Behavioral Biometrics (Wiley & Sons) and is currently finishing up texts in the areas of cognitive robotics, bioinformatics for computer scientists, and cognitive biometrics. He is the editor of two Inderscience journals: *Int. J. of Cognitive Biometrics* and the *Int. J. of Cognitive Performance Support*. He has authored over 130 papers (journal and conference), and served on 30+ international programme committees.

Mohammed Fahmy Tolba has been a Professor of Scientific Computing in the Department of Scientific Computing, Faculty of Computer Science and Information Systems since 1996. He has served as the Information Technology Advisor to the Minister of Education since 1998. He has published over 100 papers and 9 textbooks in the various aspects of computer science. He has served as an information advisor to several government organisations in Egypt, including the Ministry of Finance, and as the Chairman of the Scientific Committee of the Supreme Council for Professorship Promotion in the field of information sciences.

Abdel-Badeeh Salem is a Professor in the Faculty of Computer and Information Systems, Ain Shams University, Cairo, Egypt. He has published over 120 papers in international journals and conferences. He has established the International Conference on Intelligent Computing and Information Systems, held at the Ain Shams University (since 2002). He has taught throughout the Middle East, and has served on over 30 international conference committees. His primary research areas centre on intelligent decision support systems. He has established and continues to direct the *Egyptian Journal of Computer Science*, a monthly journal dedicated to computer science publications.

1 Introduction

Biometrics is a scientific approach for person verification and/or identification based on either anatomical or functional characteristics (motoric or cognitive) of individuals (Revett, 2008). Biometric systems have become integrated into the fabric of everyday life – deployed where and whenever secure access to a trusted instrument is required. Since the discovery of fingerprints over 100 years ago, a variety of approaches to person authentication have been devised. A wide range of methods have been deployed as biometrics, which be broadly classified into anatomical or behavioural methodologies. Anatomical are the most familiar approaches which rely on the constancy of fingerprints, retinal pattern, iris scans, and related anatomical structures. The explicit assumption about this approach is that the measurable features are unique and constant over time. As long as the anatomical structure can be measured and uniquely associated with a given individual, then the acquisition of the anatomical data should be sufficient for identification and/or verification purposes. Behavioural biometrics, a relatively newcomer to the field, relies on the *way* an individual interacts with the authentication device. For instance, signature and voice are classic examples of behavioural biometrics. This approach relies on the uniqueness of motoric function for person authentication/ identification. The lure of behavioural biometrics is its simplicity in terms of large-scale deployment, where in many cases a software-only based approach is required (i.e. keystroke dynamics). Furthermore, behavioural biometrics has a large acceptance base – many people provide their signatures for verification purposes on a regular basis. Anatomical approaches require some form of hardware – some of which can be quite costly (retinal scanner) and difficult to deploy for an internet-based trusted instrument. Further, users may not wish to be subjected to retinal scans all the time – so there is an issue of user *acceptability* that must be addressed. Lastly, not everyone has a fingerprint – or a person may have undergone events that have altered their physical form irreversibly – reducing the likelihood of being able to produce a sample when required for authentication. The issue with behavioural biometrics is one of *reliability* – reflected in the inherent believe that we are not machines – that we do not *reproduce* our signature exactly the same way every time, that our emotional state does affect of speech patterns, etc. An alternative and burgeoning branch of biometrics is termed *cognitive biometrics*, which has been gaining momentum over the past decade (Sufi et al., 2010b). This approach utilises the cognitive, emotive/affective, and conative state of the individual for person verification. The basic approach is to extract some form of a biological signal from the person while performing a given task, such as typing, searching for an element within a picture, or watching a short video clip. Typical biosignals utilised include the electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and blood pulse volume. These and related signals reflect the underlying physiological state of the individual, which should be unique. There is a genetic basis for uniqueness for certain aspects of the EEG, ECG, and GSR that can be exploited if the proper stimulus is deployed. The proper stimulus can induce cognitive and affective states in a fairly reproducible manner, increasing the attractiveness of this approach. The conative state, that is the will or intention of the person, is somewhat more difficult to quantify. Topics such as detection of deception may serve to illustrate the possibilities of this approach.

Regardless of the biometric modality deployed, each must be measured in a universal and quantitative manner if we are to effectively compare the efficacy of a given approach or instantiation of an approach. These functional requirements include: (a) universality, everyone should have the measured property; (b) collectability, the property can be measured quantitatively; (c) acceptance, users should be willing to comply; (d) uniqueness, is different for each person (a genetic basis?); (e) circumvention, how easy it is to fool the system; (f) permanence, invariant over time; and (g) robustness, reproducibility under variable conditions (Agrafioti and Hatzinakos, 2008). Any biometric system should be evaluated rigorously within these seven requirements before one can truly quantify the efficacy of an approach, and hence make claims about the utility of the biometric. In this study, the PhysioNet ECG database will be examined, as an example of a comprehensive, open-access ECG data source. The deployment of ECG as a biometric is certainly not novel, but many published results fail to examine ECG-based biometrics with respect to standard functional requirements (as stated above). Of special note is the permanence of the ECG over time – this is a central issue that will be addressed in this work. Further, one could argue that the lack of detailed functional analysis of ECG-based biometrics reflects deficiencies within the data itself, or is it simply a methodological issue. To address this issue, this study begins with a brief survey of the ECG biometrics-based literature, highlighting the principle findings and in the process, evaluating them with respect to the seven markers. This will highlight the prevalence of these markers within the main stream literature, highlighting any significant trends in terms of central tendencies. This will be followed by our own systematic approach to developing an ECG-based biometric utilising the PhysioNet ECG data repository.

The goal of this study is the development of a rigorous biometric system that benchmarks the functional the requirements of an ECG-based biometric utilising the PhysioNet data repository. The study will utilise a comprehensive set of ECG records from all four of the principal PhysioNet datasets, extracting fiducial-based features. The features space will be used to develop a model for each subject that will be used for training and testing purposes. The results presented in this work were generated by the authors (unless stated otherwise). The resulting classification accuracies (see the discussion below for more details) were used to determine the suitability of this open-access data repository as a data source for developing an ECG-based biometric. In the process, the seven functional features typically deployed as measures of efficacy will be presented. The final analysis will provide a reasonable and well-studied evaluation of the suitability of this data repository in terms of the applicability of the data in developing large-scale ECG-based biometrics.

In order to properly evaluate a biometric system, certain quantitative measures must be instantiated with values. At a minimum, the false acceptance (FAR) and false acceptance rates (FRR) need to be computed from the data. These measures reflect very generally the ease with which an imposter can gain access or the authentic user is denied access to the biometric system respectively. A FAR (Type II error) occurs when a false claim of identity is accepted – an unauthorised user has managed to gain access to the system. A FRR (Type I error) occurs when a true claim of identity is rejected, resulting in inconvenience to the authorised user at the very least. A derivative measure deploying FAR/FRR is the cross-over or equal error rate (CER and EER respectively), which is the value of the intersection of FAR and FRR when plotted on the same graph. This value provides a reliable measure of the trade-off between type I and type II errors as a function of some threshold used in the authentication process (typically the matching scheme threshold). In addition, one may wish to know what the FAR is when the FRR is set to 0.0. Other quantitative issues arise such as the failure to enrol (FTE) and failure to authenticate (FTA). FTE occurs when a user is not able to present their details during the enrolment process, a mandatory stage in most biometric systems. All users of the system must go through an initial enrolment process, which acquires the data from the biometric modality that will be used for subsequent authentication/verification/identification (herewith termed – authentication) purposes. The data acquired from the enrolment process is used to build a model of each user of the system, stored in a secure (encrypted) repository. Once the enrolment data has been acquired, it will be used in the authentication process by comparing the same feature values acquired during an access request to that stored in the data repository. Depending on whether the system is designed for identification (a 1:N mapping) or verification/authentication (1:1 mapping), a matching score is produced which is used to determine the outcome of the request. FTE implies that the user was not able to engage positively in the enrolment process. This may occur because the subject did not possess the biometric feature required (i.e. fingerprints are not available or unusable due to injury or health conditions) or are unable cognitively to comply with the request with respect to stability when behavioural biometrics are deployed (typically requires replicate signatures for example). Another metric deployed in biometrics is the FTA, which reflects a change in the subject subsequent to a positive enrolment phase. For instance, a subject may have been injured and their fingerprints are no longer available, a broken arm may mean the person cannot deploy their dominant hand for signature verification. Lastly, the detection error trade-off (DTE) curve, a modified ROC curve, is sometimes reported. This curve plots the false non-matching rate (FNMR) against the FMR (False Matching Rate) on the x-axis (i.e. matching error rates). This measure attempts to treat the error rates (FNMR and FMR) equally, and is typically plotted using a log scale to spread out the data points to highlight the salient features in the data with respect to matching rates. Note that the EER curve deploys detection rate data, and hence presents complimentary data to the DTE curve. Another potential metric rarely discussed is the cumulative match rate (CMC), which plots the probability of detection as the size of the database increases. Thus, this measure provides information regarding the scalability of the biometric process – what is the likelihood that a match will be found if we start say with 10 users and then increase the database to 10^6 users. This list of metrics is not exhaustive, though it covers the majority of those reported in the biometrics domain (though few papers report multiple metrics). Very few, if any papers, have managed to report all of these measures from results obtained in a single study. It is quite possible that data-based limitations, such as small sample size and single acquisition schemes, precludes acquiring many of the functional metrics feasible. This is why data repositories such as *PhysioNet* are potentially very useful. They provide data that can evaluate the generalisability of the data – a feature rarely reported in the biometrics literature. That generalisability is an important concept in biometrics is beyond question. The issue is how to address and measure it in a scenario that will ultimately lead to large-scale implementations of a biometric approach?

One could argue that quantitative metrics such as those discussed previously provide a mechanism that allows one to quantify suitability of the data. To investigate true generalisability requires large amounts of data collected from multiple subjects over a series of time points. Studies of this nature require careful experimental design approaches, in addition to sound machine learning techniques. The argument is that with properly defined experiments, and acceptable machine learning approaches, quantitative metrics (FAR, FRR, etc.) provide a reliable measure of the generalisability of the data. This is the central question addressed in this work, which will start with a brief description of published work utilising ECG data as the basis for developing a biometric solution.

2 Related work

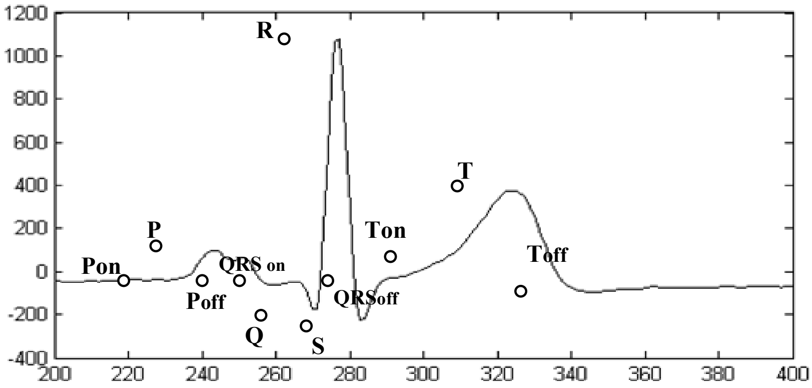
This work assumes that the reader is familiar with the ECG signal waveform, which is a time varying recording of the electrical potential produced by physiological activity of the heart (see Figure 1). Further, it is assumed that there will be an enrolment stage prior to authentication from which a model of the user is generated and deployed in the authentication process. The focus of this section is to provide a detailed summary of published work (not exhaustive unfortunately) which has utilised ECG as a biometric, highlighting the salient features with respect of generalisability.

A typical ECG-based authentication (identification or verification) system usually consists of three crucial stages:

1. pre-processing stage where the signal is filtered for noise and baseline removal
2. feature extraction where a set of features that represent the distinctive characteristic of an individual is extracted (model building)
3. classification, where the extracted features are fed into a classifier assigned labels according to some distance metric from samples stored in the enrolment database.

As with most biosignal-based waveforms, ECG data can be decomposed into a feature space based on either fiducial (i.e. amplitude of landmark features) or non-fiducial based approaches (wavelets/frequency components). This work will focus solely on fiducial-based feature analysis, such as the amplitude of the signal fiducial points: P, Q, R, S and T (see Figure 1) and the durations between them. For classification there are two cases: (a) human identification, a 1:N mapping, where the assumption is that the subject exists within the database and the classification task is to find the closest match to the user requesting access and (b) human verification, a 1:1 mapping, where the task is to compare the access request with the stored model of the user, a much easier task than identification. In both instances, the model is derived from features acquired during the enrolment process. Therefore, the selection of appropriate features is a critical component of any biometric system.

Figure 1 A single heartbeat, illustrating the major landmarks within an ECG. Note the 'on' and 'off' subscripts indicate the start and end of each sub-waveform respectively



2.1 Analytical (fiducial) features based systems

Fiducial features refer to the identification of landmarks that occur within an ECG signal, such as the location of the maximum peak (the 'R' peak) which occurs within a QRS complex (see Figure 1). These features are associated with cardiac physiology such as the contraction and relaxation of the auricles and ventricles. Their identification is important in the diagnosis of a wide variety of heart diseases. A significant amount of research effort has been exerted in order to produce automated fiduciary mark identification systems, which serve as the basis for many machine-based approaches to automated diagnostic systems. These efforts have provided a wealth of information on fiducials that have become useful in the development of ECG-based biometrics. Fiducial features represent temporal and amplitude distances between fiducial points in a heartbeat signal. A crucial issue here is that the reliability of the extracted features is strongly dependent on the accuracy of the detected points, a challenge by itself. For instance, variability in noise levels, recording conditions, number of leads, sampling frequency and related practical considerations introduce a potential for considerable variation in fiduciary extraction methods, especially when trying to acquire them in an automated mode. Further, there are issues of the stability of the ECG within a patient record over time. Each heartbeat contains a stereotyped waveform, but with certain medical conditions (or variations in the emotional status of the subject), the heartbeats exhibit arrhythmic patterns that may require several minutes' worth of beats to identify. This section provides a brief review of *some* of the published work that utilises fiducial features extracted from ECG as the basis of a biometric system.

A set of 30 temporal and amplitude features extracted from recorded ECG data were utilised by Biel et al. (2001). The features were directly extracted from the recording equipment (SIEMENS ECG), and hence there was little control over the selection of features, though they were able to pre/post-process the data. This study yielded 100% human identification rate, achieved via a study deploying 20 subjects, using a multivariate analysis-based method for classification. A Bessel filter of 4th order is applied by Gahi et al. (2008) as a first step to remove noise from ECG signals of 16 persons. A set of 24 fiducial features were utilised and reduced by arranging them according to information gain ratio then the first 9 were chosen. The system was tested and 100% subject identification accuracy was achieved using a Mahalanobis' generalised distance based classifier.

Shen et al. (2002) proposed a two-step scheme for identity identification from one-lead ECG. A template matching method was first used to compute the correlation coefficient for comparison of two QRS complexes. A decision-based neural network (DBNN) approach was then applied to complete the verification from the possible candidates selected with template matching. The inputs to the DBNN were seven temporal and amplitude features extracted from QRST wave. The experimental results from 20 subjects showed that the correct verification rate was 95% for template matching, 80% for the DBNN, and 100% for combining the two methods. Shen (2005) extended the proposed methods in a larger database that contains 168 normal healthy subjects. Template matching and mean square error (MSE) methods were compared for pre-screening, and distance classification and DBNN compared for second-level classification. The features employed for the second-level classification were 17 temporal and amplitude features. The best identification rate for 168 subjects was 95.3% using a template matching and distance classification methodology.

A set of descriptors to characterise ECG trace was proposed by Israel et al. (2005). Input ECG signals were filtered using band pass filter, which were then traced to locate fiducial positions in each heartbeat. 8 points (4 peaks and 4 base points) were detected in the time domain in 2 stages. The peaks were located by finding the local maximum in a region surrounding each of the P, R, and T Complexes. The base positions were identified by tracking downhill and finding the location of the minimum radius of curvature. A set of 15 distance features were then extracted from the distances between fiducial points. Since the distances between the fiducial points and the R position vary with heart rate, the distances were normalised by dividing them by the full heartbeat duration. A Wilk's Lambda method was applied and reduced the number of features to 12. The classification was performed on heartbeats using standard linear discriminant analysis. The system was tested on dataset of 29 subjects, yielding 100% human identification accuracy, while the heartbeat identification rate was 81%. Moreover, the effect of varying ECG lead placement was tested and no significant differences were observed in the results. Finally, the invariance of the normalised extracted features to individual's anxiety states (i.e. reading aloud, mathematical manipulation and driving) was tested and validated.

Singh and Gupta (2009) developed a method for subject verification based on the following scheme: (a) a pre-processing step that includes enhancing the SNR and artefact removal using successive low & high pass filters; (b) QRS detection and a novel P and T delineators (developed by the authors) were utilised to detect QRS, P and T waves, respectively from the ECG records; (c) 19 stable features related to interval, amplitude and angle were computed from each heartbeat. The extracted feature set was utilised via a template matching and an adaptive thresholding scheme. The results yielded an overall classification accuracy of 99%, with FAR 2% and FRR 0% on the data set of 50 ECG datasets from healthy individuals obtained from the QT database (Laguna et al., 1997). It should be noted that the FRR was acquired from the training data, while the FAR was tested by data acquired and processed in the same manner from European ST-T database.

Singla and Sharma (2010) introduced a method that utilised seven amplitude features (notably without any duration based features), utilising a deviation threshold method for the classification task. The system was tested on a database of 20 persons and produced 97% subject identification accuracy. The system was tested for FAR and FRR, yielding values of 3.2% and 3% respectively. The authors used the same data for training and testing purposes, and for calculating the FAR and FRR values. The system utilised 10 samples for each person for testing, and this was subsequently used for measuring FRR (number incorrectly classified out of the 10 samples). The FAR was computed by comparing template of each person with 190 samples from the remaining 19 subjects (10 records each).

2.2 Appearance (non-fiducial) features based systems

Appearance-based approaches do not rely on the direct identification of landmarks (fiducial marks) in the ECG record; rather they deploy a window approach which is moved along the data. The length of the window must be long enough to ensure it contains a complete heartbeat (typically 2–3 seconds is sufficient). Once a heartbeat has been identified within a window, then subsequent heartbeats are acquired by assuming a fixed length heartbeat, and knowing the termination of a single heartbeat, the window can be slid across the data, one heartbeat at a time. In order to determine the length of a single heartbeat, the R-R interval is determined (the temporal distance between successive peak maxima). This provides the length between heartbeats, and then the window is moved such that the R peak is located in the middle of the window. This is an easy computation compared with identifying the set of fiducial marks contained within a heartbeat (which can contain approximately 30 features/metrics), depending on the detail required. Thus, there is no need to detect the full set of fiducial points, a significant advantage of this approach, which is deployed in a significant number of ECG-based biometric systems, a few of which are described next (Khalil and Sufi, 2008; Wang et al., 2008; Fatemian and Hatzinakos, 2009).

Wan and Yao (2008) proposed an artificial neural network (ANN) based approach for verifying ECGs. After removing noise from ECG signals using successive low & high pass filters, 40 heartbeats were extracted for each subject. The 40 heartbeats were reduced to 10 by averaging every 4 heartbeats, which was designed to reduce the noise level. Each of the 10 heartbeats/subject was decomposed into 256 wavelet coefficients. These coefficients were used as input vectors to a 3 layer feedforward neural network. The network input layer accepts 2 heartbeats wavelet coefficients (512-element vector) as the input vector. The output of the neural network generates the discrimination results which indicate whether the two heartbeats in the input vector come from the same or two different individuals. If the two wavelet coefficient structures in the input vector belong to the same individual, the network output is +1; otherwise –1. The system was tested on a database of 23 persons and it was generalised on a database of 15 persons. 100% verification accuracy was achieved in both cases.

A new wavelet based framework was introduced and evaluated by Fatemian and Hatzinakos (2009). The proposed system utilised a robust pre-processing stage that is directly applied on the raw ECG signal for noise handling. Furthermore, one of the novelties of this system was the design of a personalised heartbeat template so that the gallery set consisted of a single heartbeat per subject. A dyadic wavelet transform (DWT) was applied to the raw ECG signals, and then the signals were reconstructed at the third scale where most of the signal energy was retained. Further smoothing via moving average was also applied. By utilising the computed wavelet coefficients the deflected heartbeats were removed. The remaining heartbeats were then re-sampled, normalised and using the median of the aligned heartbeats, the heartbeat template for each subject was constructed. Finally, classification is accomplished based on the correlation among templates. The system was evaluated over two common databases: MIT-BIH (14 subjects) and the PTB (13 subjects), yielding a maximal accuracy of 99.6%.

Platianiotis and Agrafoti separately utilised the significant coefficients of the discrete cosine transform (DCT) of the auto-correlated heartbeat signals after noise removal (Laguna et al., 1997; Kunzmann et al., 2002). Each ECG signal was divided into windows with the restriction that each window had to be longer than average heartbeat length so that multiple pulses were included (obviating the need for R peak detection). Normalised autocorrelation of each window was computed followed by a discrete cosine transform DCT. The first *C* coefficients that provided the best results are chosen. Plataniotis et al. (2006) used a Euclidean distance as a criterion for discrimination. The system was tested on 13 subjects selected from the PTB database and generalised on 14 subjects from the MIT_BIH database. The results obtained yielded 100% subject identification accuracy for both datasets, and a window recognition accuracy 94.4% for PTB and 97.8% for MIT_BIH. Agrafioti and Hatzinakos (2008) introduced the use of discarding PVC beats from ECG records by checking their DCT and utilised nearest neighbourhood classifier with LDA for reducing DCT coefficients. The system was tested on a database of 56 subjects with 96.4% subject identification accuracy. The same authors also compared using LDA versus DCT for reducing the dimensionality of the normalised auto-correlated signals computed from windowed ECG signals. Template matching was applied for the purpose of classification. The system was tested on 27 subjects from MIT_BIH and PTB databases with 100% subject identification accuracy and 96.6% window recognition accuracy using AC/LDA approach while the results of utilising the AC/DCT approach was 96.3% and 88.9% for subject identification and window recognition accuracy respectively.

A polynomial distance measure (PDM) method for ECG-based biometric authentication was introduced by Sufi et al. (2010a). The method was implemented in a series of steps:

1 for each heartbeat the three complexes P, QRS and T were detected

2 the three complexes were differentiated

3 for each wave an approximated polynomial equation was generated, resulting in a set of stored coefficients.

The coefficients of the three waves were concatenated to form a feature vector for a specific heartbeat. A match between two beats was achieved when the Euclidean distance between their feature vectors was below a parametised threshold. The system was tested on a database of 15 subjects with 100% subject identification accuracy.

A two-stage hierarchical scheme which integrates both analytical and appearance features was proposed by Wang et al. (2008). The first stage utilises 9 analytical features selected from a set of 21 features using Wilk's Lambda method and LDA for classification, while, the second stage utilises PCA feature extraction and a nearest neighbourhood classifier. The hierarchical scheme works as follows: initially only analytic features were used to provide a rough estimate of the potential classes that each entry might belong to. If all the heartbeats are classified as belonging to a single subject, the decision module will return the result without further modification. If some of the heartbeats were misclassified, the PCA-based classification module was applied. The system was tested on 13 subjects of PTB database and it was generalised on 14 subjects of MIT_BIH database with 100% identification accuracy for both and heartbeat recognition accuracy of 98.9% and 99.4% for PTB and MIT_BIH respectively.

2.3 Critical issues of reliability and robustness

The case studies presented in the previous sections indicate that accuracies well above 95% can be obtained, attesting to the level of uniqueness contained within short-term ECG recordings – as well as to the discriminative capacity of the machine learning approaches utilised. Questions arise regarding the reliability of these results – as many of the published reports do not present quantitative reliability data (see Table 1 for details). Critical issues are: (a) most of the existing systems have computed subject identification accuracy, but few of them consider the heartbeat (window) recognition accuracy, which is the number of beats or windows correctly classified as belonging to a particular subject; (b) testing the resulting classifier on heartbeats extracted from records other than those used in the training set, which is only considered in a few of the reported cases discussed in this paper; (c) whether or not the model overfits the data, typically due to an incomplete training/testing protocol, and (d) the response of the system with respect to changes in the number of subjects. Does the system scale when the number of subjects increases significantly? Notwithstanding the issue of heartbeat versus subject recognition, most of the studies discussed in this paper (which is a reasonable survey of published reports) suffer from a small subject base (range from 10 to 50 subjects). These and other issues should be addressed in the experimental design phase of the experiment, and care must be taken to ensure that enough subjects are used and proper validation methods are deployed. These considerations should result in a system which is scalable and generalisable – important properties when deploying a biometric within an environment that may potentially contain a very large number of users (i.e. border patrol).

Table 1 A summary of the implementations discussed in the text, with respect to what biometric quality control measures were reported in the relevant publications

References	*Heartbeat (window) recognition accuracy*	*Training & testing from different records*	*Generalisation to other databases*	*Scalability*	*Testing for FAR & FRR*
Biel et al., 2001; Kyoso and Uchiyama, 2001; Chan et al., 2005; Chan et al., 2008; Boumbarov et al., 2009; Tawfik et al., 2010	✗	✓	✗	✗	✗
Shen et al., 2002; Chuang et al., 2005; Shen, 2005; Khalil and Sufi, 2008; Sufi et al., 2008; Fatemian and Hatzinakos, 2009; Guennoun et al., 2009; Singh and Gupta, 2009; Li and Narayanan, 2010	✗	✗	✗	✗	✗
Singh and Gupta, 2006	✗	✗	✗	✗	✓
Israel et al., 2005; Agrafioti and Hatzinakos, 2008; Irvine and Israel, 2009	✓	✓	✗	✗	✗
Zhang and Wei, 2006; Gahi et al., 2008	✗	✗	✗	✗	✗
Plataniotis et al., 2006; Wang et al., 2008	✓	✓	✓	✗	✗
Singla and Sharma, 2010	✓	✓	✗	✗	✓
Wan and Yao, 2008	✗	✗	√	✗	✗
Wao and Wan, 2010	✗	✓	✗	✓	✗
Agrafioti and Hatzinakos, 2008	✓	✗	✗	✗	✓

Only a few of the studies referred to in Table 1 attempt to the issue of generalisability (Plataniotis et al., 2006; Wan and Yao, 2008; Wang et al., 2008). Very few studies address the issue of scalability, which addresses issues regarding system performance when the number of enrolled users increases. Sufi et al. (2010a) and Khalil and Sufi (2008) introduced systems that utilised a few polynomial coefficients as features, resulting in 100% accuracy in a study involving 10 and 15 subjects respectively. One would hardly expect such a small cohort to yield results that would be applicable to a large population without further testing. Wao and Wan (2010) trained their system with range of subjects from 5 to 30, and found that when the number of subjects was below 15 the accuracy was 90% and 80% for the whole dataset (30 subjects). Moreover, many of fiducial features based systems utilised a reduced set of features selected from larger set determined either empirically or via some statistical method. How many features are sufficient to generate a robust *and* efficient biometric is a perennial question. Shen et al. (2002) worked on 20 subjects only and achieved 100% with 7 features. However, when Shen (2005) increased the number of subjects to 168, the number of features increased to 17 and the accuracy became 95%. Was this change in accuracy due to a change in the number of subjects or the number of features? One way to address this question is through the use of quantitative metrics. These include FAR, FRR, and EER, which effectively normalise the classification accuracy. A few of the systems presented in Table 1 provide such results (Singh and Gupta, 2006; Agrafioti and Hatzinakos, 2008; Singla and Sharma, 2010), but may not acquire this information in a manner that preserves robustness. For instance, the studies reported by Agrafioti and Hatzinakos (2008) and Singla and Sharma (2010) measured the FAR & FRR from the same training database. Singh and Gupta (2006) reported a 2% FAR using one database (different from the one used for training), but reported their FRR results using the training database. In short, none of the published systems (certainly those in Table 1) completely address critical issues of validation, scalability and generalisation. This work was designed to examine how to utilise relatively large data repositories such as the PhysioNet – with its four principal ECG databases, such that scalability and generalisability issues can be addressed. The net result is a comprehensive survey of this data repository – in terms of its applicability as a source of data for developing ECG based biometric solutions. The approach taken in this work is described in detail in the next section.

3 Methodology

The ECG-based authentication system reported in this paper was developed in four stages:

1 pre-processing, which is needed to remove noise and baseline oscillations from ECG records

2 fiducial points detection; a preliminary and crucial stage that includes detecting the peak and the end points of each of the three complexes QRS, P and T

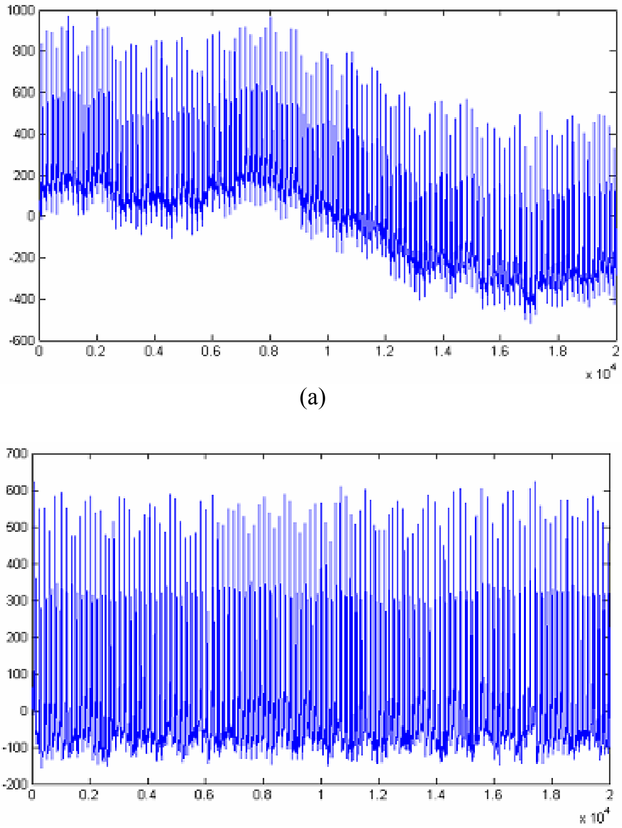
3 feature extraction; it involves extracting 28 fiducial based features from each heartbeat and normalise them to avoid heart rate variability

4 authentication/identification; this stage utilises both a feedforward and Radial base functions (RBF) neural networks as classifiers.

3.1 Pre-processing

ECG records typically contain some sort of noise and baseline wander that should be eliminated as much as possible to increase the accuracy and consistency of detecting fiducial marks. Baseline oscillation, body movements and respiration rhythms contribute noise at low frequencies, while power line interference and digitisation issues contribute to high frequencies in the ECG spectrum (Singh and Gupta, 2006). To address these issues, the data deployed in this study were first pre-processed by applying a Butterworth filter of second order. The cut-off frequencies of the filter were 0.5–40 Hz based on the implemented algorithms for fiducial points detection. Figure 2 shows an ECG segment before and after the filter application, which has corrected baseline drift by band pass filtering.

Figure 2 An example of an ECG time series before (a) and after baseline the application of a 2nd order Butterworth filter (b) (see online version for colours)



(a)

(b)

3.2 *Fiducial point detection*

3.2.1 *QRS complex detection*

A *Detection of R peak*: The Pan and Tompkins (1985) algorithm was applied for QRS detection due to its robustness. The algorithm is typically implemented as follows: (a) low & high pass filtering; (b) differentiation; (c) squaring; (d) moving-window integration; (e) thresholding; (f) search back for missing R waves and remove misclassified T waves as R waves due to their height. Figure 3 shows an ECG segment at various steps in R detection.

B *Detecting QRS onset and offset*: After detecting R peaks, the onset and the offset of all QRS waves is implemented using the Illanes-Manriquez and Zhang (2008) algorithm. The algorithm computes an envelope of the QRS complex from the filtered ECG signal, it represents the modulus of the complex signal formed by the filtered ECG signal (the real part) and the Hilbert transform of the filtered ECG signal (the imaginary part). During the occurrence of a QRS complex, the envelope signal has a bell-shaped concave wave, and the beginning and the end of this concave wave correspond respectively to the QRS onset and offset (Figure 4). The onset of a QRS wave is detected after the computation of an indicator related to the area covered by the QRS complex envelope inside a moving window and based on the fact that the computed value will reach its maximum value when the window reaches the QRS onset. The QRS offset is detected in the same way, but the ECG signal is filtered by a band pass filter with cut-off frequencies (5–30 Hz) before computing the envelope.

C *Detecting Q&S points*: In the proposed work, the Q point was defined as the local minimum (valley) between the R peak and the onset of the considered QRS wave, while, the S point is the local minimum between the R peak and the offset point.

Figure 3 R peak detection algorithm steps for a normal ECG segment: (a) original signal; (b) after successive low & high pass filtering; (c) differentiated signal; (d) output of squaring process; (e) results of moving-window integration; (f) position of R peaks after the thresholding step and delay removal (see online version for colours)

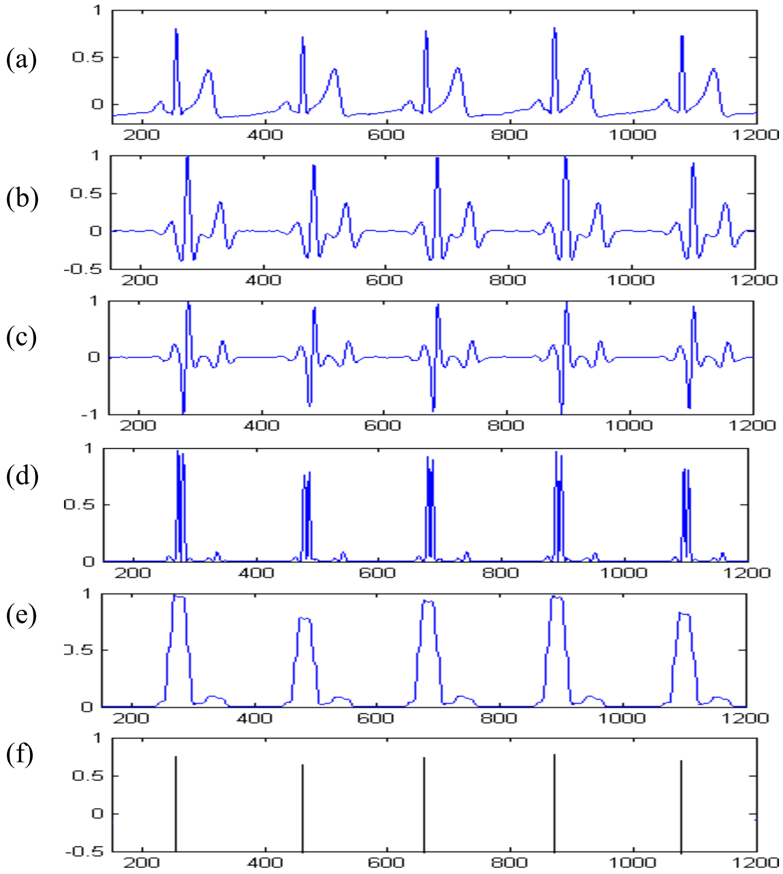
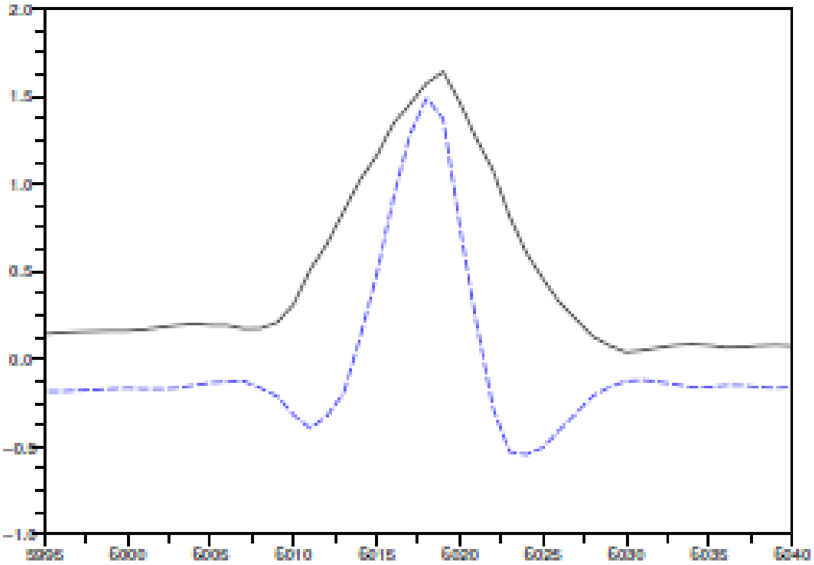


Figure 4 The lower trace is the QRS signal and upper trace is the resulting envelope (see online version for colours)



3.2.2 P wave detection

A *P peak*: A search window was defined for the P wave before the QRS complex. It begins from the QRS onset with width 200 ms (Figure 5a). The peak of the P wave is the local maximum in the search window. If two peaks are found with ratio between them is less than Θr and duration between them less than Θd then this P wave has an 'M' shape. In this case the P peak is considered the one with higher amplitude. The width of the window and values of thresholds were computed empirically through experimentation.

B *P wave end points*: The onset and offset of the P wave was defined using the method of minimum radius of curvature (Israel et al., 2005; Wang et al., 2008; Guennoun et al., 2009; Tawfik et al., 2010). This method proved more robust to local noise than the more obvious derivative measures (Israel, et al., 2005). The P wave onset was defined by tracking downhill from the right side as shown in Figure 6. The X and Y were fixed and then the minimum radius of curvature was found by maximising the value of δ using the vector cross product between the two directed line segments. The P wave offset is defined in the same way but this time by tracking downhill from the left side of the P wave. For more accurate detection of the end points, a small modification is introduced to consider the case of M shape P wave. Regardless, which of the two peaks is considered the peak of the P wave, to define the onset we track downhill before first peak, while for the offset we track downhill after the second peak.

Figure 5 (a) Window location used for detecting a P wave and (b) search window for T wave (see online version for colours)

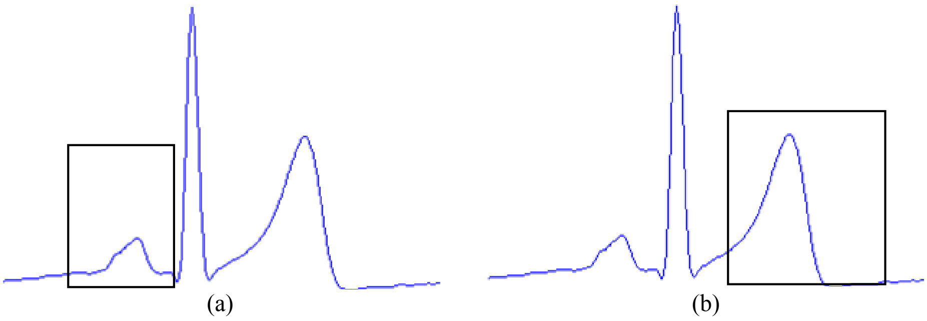
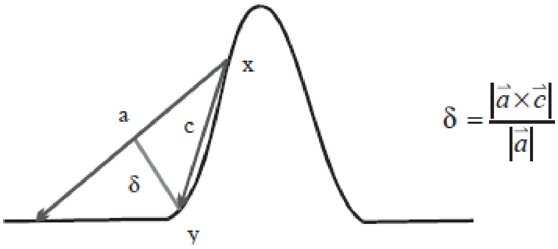


Figure 6 Radius of curvature and the resulting computation of curvature



$$\delta = \frac{|\vec{a} \times \vec{c}|}{|\vec{a}|}$$

3.2.3 T wave detection

A *T peak detection*: A search window was defined after the QRS complex to detect T peak. It begins from the QRS offset and remains for 400 ms (Figure 5b). This width of the window was computed empirically through experiments. The T peak was detected using the wing functions proposed by Zarrini and Sadr (2009). This method was chosen due to its efficiency, simplicity and its ability to detect both positive and negative (inverted) T waves. The 'wings' function obtained at each successive sample in the search window, wings were modelled with 2 neighbouring segments W_1 and W_2 as follows:

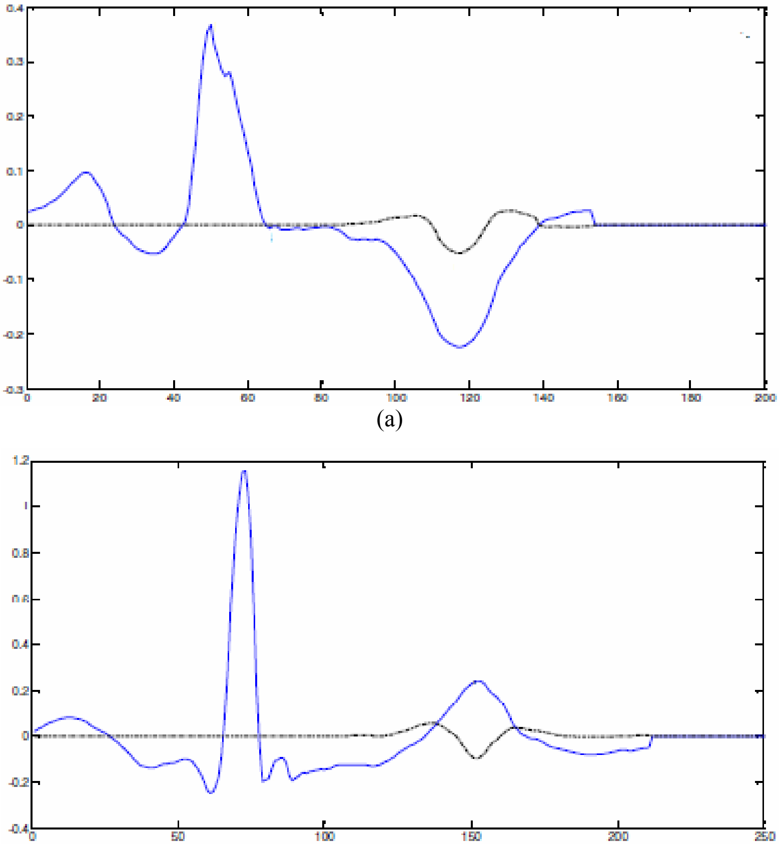
$$W_1 = X_{i-16} - X_i \tag{1}$$

$$W_2 = X_i - X_{i+16} \tag{2}$$

where, X is the input sample to the wing function. A Wing function W is computed by multiplying W_1 and W_2. The minimum point of the W signal inside the search window is considered the T peak as shown in Figure 7.

B *T wave end points*: The onset and the offset of the T wave were also detected using the method of minimum radius of curvature, even for the negative T wave, but in this scenario, the direction is reversed – climbing up the valley not tracking downhill as in the positive case.

Figure 7 (a) ‘Wings’ functions of a negative T wave, (b) ‘wings’ functions of a positive T-wave (see online version for colours)



(a)

(b)

3.3 Feature extraction and normalisation

After detecting the fiducial points, features were extracted from each heartbeat. In the literature, the fiducial features based systems typically utilised a small set of features (by design or through dimensionality reduction from a larger set of features). Typically, some information based approach (such as information gain or PCAS) is used to reduce the feature space. This work utilises a set of 28 features that represents the majority of features utilised in the literature. As shown in Figure 8, these features encompass 19 temporal features (distances between fiducial points), 6 amplitude features and 3 angle features. All the features are listed in Table 2. The temporal features are normalised using the full heartbeat duration to avoid heart rate variability. While, the amplitude features are normalised by the R amplitude to avoid signal attenuation, and the angle features are used as raw features. In order to have the same impact on the classifier, the range of all features was scaled to a range from 0 to 1.

Figure 8 Details of the 28 fiducial features used in this work

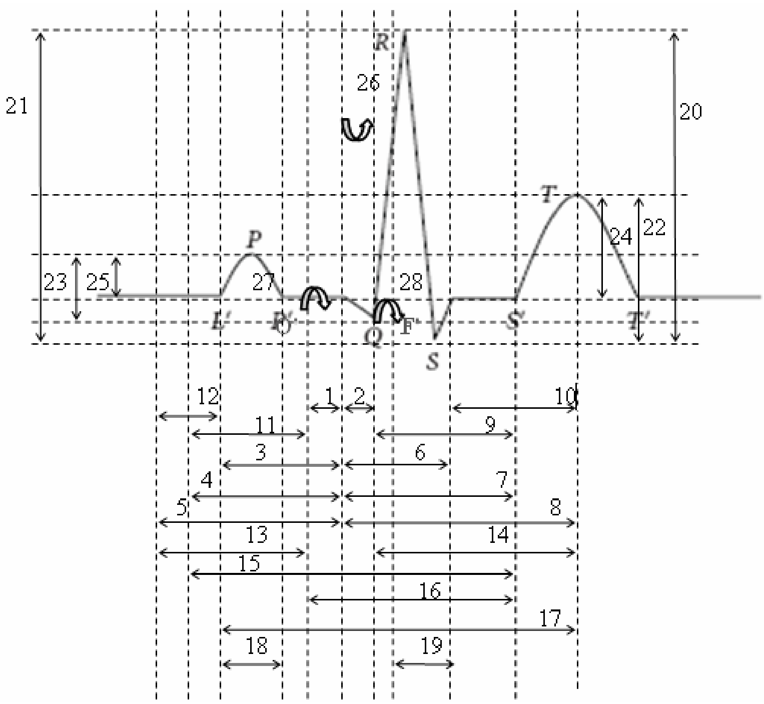


Table 2 Summary of the 28 fiducial features, categorised according to type of features

Feature type	*Extracted features*			
Temporal features	1) RQ	6) RS'	11) PQ	16) QT
	2) RS	7) RT	12) P'L'	17) P'T'
	3) RP'	8) RT'	13) QL'	18) P'O'
	4) RP	9) ST	14) ST'	19) F'S'
	5) RL'	10) S'T'	15) PT	
Amplitude features	20) RS	22) TS	24) TT'	
	21) RQ	23) PQ	25) PL'	
Angle features	26) < R	27) < Q	28) < S	

3.4 Subject identification

In this stage, the 28 extracted features of a heartbeat are fed into the classifier to be identified. Two classifiers were utilised in this work: Feedforward and Radial basis Functions (RBF) neural networks.

3.4.1 Feedforward neural network classifier

Feedforward neural networks (FFN) are one of the most popular and most widely used models in many practical classification applications. FFN distinguishes itself by its nonlinear processing elements via the inclusion of one or hidden layers. Most bare trained using some form of the back propagation algorithm (Haykin, 1998). In this work, results are not improved with the addition of a hidden layer, so one was not included in any of the work presented in this paper – resulting in a single layer perceptron with nonlinear activation function. The number of output units was set to the number of classes (subjects to be identified) contained within the data. The Softmax nonlinear function was chosen as the activation function for all elements in the network. The choice of this function was suitable in this study, as the outputs compete, yielding values between 0...1 that can be used in a winner-takes-all strategy. The function formula is as follows:

$$\sigma(y_i) = \frac{\exp^{y_i}}{\sum_j \exp^{y_i}} \quad \textit{where } y_i = \sum_{k=1}^{n} x_k w_k \tag{3}$$

where x is the input vector of length n, w is the weight vector (same length as x) associated with the element number i and $\sum_j \exp^{y_i}$ is the summation of the exponential of the outputs of the j processing elements in the layer.

3.4.2 RBF neural network classifier

Radial basis functions (RBFs) have attracted a great deal of interest due to their rapid training, generality and simplicity. It has been proven that RBF networks, with enough hidden neurons, are also universal approximators. The RBF network is based on the simple idea that an arbitrary function $y(x)$ can be approximated as the linear superposition of a set of localised basis functions $\varphi(x)$ (Haykin, 1998). The RBF is composed of three different layers: the input layer in which the number of nodes is equal to the dimension of input vector. In the hidden layer, the input vector is transformed by a radial basis activation function (Gaussian function):

$$\varphi(x; c_j) = \exp\left(-\frac{1}{2\sigma^2}\left\|x - c_j\right\|^2\right) \tag{4}$$

where || || denotes the Euclidean distance between the input data sample vector x and the center c_j of Gaussian function of the j-th hidden node; finally the outer layer with a linear activation function, the k-th output is computed by equation

$$F_k(x) = \sum_{j=1}^{m} w_{kj} \cdot \varphi(x; c_j) \tag{5}$$

where w_{kj} represents a weight synapse associates with the j-th hidden unit and the k-th output unit with m hidden units (Haykin, 1998) and the orthogonal least square algorithm (Chen and Chang, 1996) is applied to choose the centres, which is a very crucial issue in RBF training due to its significant impact on the network performance. This algorithm is chosen for its efficiency and because there are no parameters to be defined or randomly initialised.

4 Experiments and results

The PhysioNet ECG repository was used as a source of ECG data throughout the experiments reported in this paper. The repository consists of four principal datasets, the characteristics of which can be summarised as follows:

1. The QT database (Laguna et al., 1997) consists of 105 records, each 15 minutes long, and sampled at 250 Hz. The recordings have been chosen to include a broad variety of ECG morphologies and have been extracted from other existing ECG signal databases as follows: 10 from MIT-BIH Arrhythmia, 6 from MIT-BIH ST Change, 13 from MIT-BIH Supraventricular Arrhythmia, 10 from MIT-BIH Normal Sinus Rhythm, 33 from European ST-T, 24 'sudden death' patients from BIH and 4 from MIT-BIH Long-Term ECG databases. For each record, between 30 and 100 representative beats were manually annotated by cardiologists for all fiducial points except T wave onsets, for which the annotation was incomplete.
2. The PTB database (Bousseljot et al., 1995) includes 549 records from 294 subjects. Each record consists of the conventional 12-leads and 3 Frank leads ECG, sampled at 1000 Hz. The majority of the subjects (244) were pathological, exhibiting a variety of medical pathologies, while the remaining 51 subjects were healthy. In addition, some of the subjects' (13) records were acquired over multiple sessions, with a variable delay (in most cases several months) between recordings. These subjects were distinguished from the remainder of the subjects (38), from whom ECG records were acquired within a single day. These two sub-datasets are termed PTB_1 and PTB_2 respectively in this work. Note that for all subjects (51) in this group, the ECG was collected for 1.5–2 minutes, yielding approximately 100 heartbeats. The pathological records from this dataset were not utilised.
3. The MIT BIH Normal Sinus Rhythm database (Goldberger et al., 2000) contains 18 ECG (2 leads) recordings from different clinically normal subjects (at least no arrhythmias were annotated in the records), sampled at 128 Hz. This dataset was acquired as a continuous 24 hour record, which provides data to evaluate local stability. It should be noted, however, that the majority of the records contain significant numbers of artefacts and noise (this was especially obvious in 4 of the subjects). This database, along with 6 subjects selected from the MIT BIH long-term database, were utilised as a source of impostors only (Goldberger et al., 2000).
4. The Fantasia database (Iyengar et al., 1996, Goldberger et al., 2000) contains records for 40 subjects, 20 young (21–34 years) and 20 elderly (68–85 years old). It is noteworthy that there were equal numbers of male and female subjects in this dataset. There is only a single 2-hour record for each subject, which was sampled at 250 Hz. Each record includes ECG, respiration, and for half of the dataset the records include a blood pressure waveform.

A series of five experiments were performed in this work. The initial experiment addressed the performance assessment of the implemented algorithms for detection of fiducial points, while the remaining experiments were concerned with the exploring the robustness and generalisability of the PhysioNet datasets. In terms of classification techniques, for a comparison, the feedforward and RBF neural networks were utilised. They were selected because of their widespread use and their generality as classifiers.

4.1 Fiducial point detection

The performance of the implemented fiducial detection algorithms was assessed on QT database. This database is commonly used for validation and assessment of such algorithms, because of the extensive annotation associated with the subject records. The performance of the R wave detector was assessed as in Martinez et al. (2004), by calculating the sensitivity and positive predictive values (99.6% and 99.4% respectively). More generally, with respect to automated fiducial point isolation, the best one can do is to compare the computed results with those obtained by manual inspection by a domain expert. The mean of the errors μ was taken as the time differences between automated and cardiologist annotations; the mean standard deviation of the error σ which was computed by averaging the intra-recording standard deviations (Laguna et al., 1994; Kunzmann et al., 2002; Martinez et al., 2004; Martinez et al., 2010). Table 3 includes results of our detectors and related results in the literature for comparison. Results are given in milliseconds in the form of $\mu \pm \sigma$. In this work, each 5 ms in a single sample point (Fs = 200 Hz), while in literature reports indicated in Table 3, one sample point is 4 ms (Fs = 250 HZ). The Q, S and T onset results are not included because these points are annotated in only a few records in the QT database. The accepted σ of measurements recommended by CSE committee (CSE, 1985) is also given in the last row of Table 3.

Table 3 Summary of a selection of published results regarding the accuracy of automated fiducial point detection from studies deploying the QT database

	P_{onset}	P_{peak}	P_{offset}	QRS_{onset}	QRS_{offset}	T_{peak}	T_{offset}
This work	8±12.37	3.13±11.34	11±13.1	6.8±6.47	–3.5±5.77	8.9±11.89	–4.5±14.48
Martinez et al. (2004)	2±14.8	3.6±13.2	1.9±12.8	4.6±7.7	0.8±8.7	0.2±13.9	–1.6±18.1
Martinez et al. (2010)	2.6±14.5	32±25.7	0.7±14.7	–0.2±7.2	2.5±8.9	5.3±12.9	5.8±22.7
Kunzman et al. (2002)	3±68	–12±40	–19±35	–20±10	26±11	–40±35	–68±46
Plataniotis et al. (2006)	14±13.3	4.8±10.6	–0.1±12.3	–3.6±8.6	–1.1±8.3	–7.2±14.3	13.5±27
Tolerance (CSE, 1985)	10.2	–	12.7	6.5	11.6	–	30.6

Note: All values are in milliseconds.

4.2 Testing stability of ECG

For this analysis, the PTB_1 dataset was utilised, as it provides multiple subject recordings which span several months (in some cases more than 1 year elapsed between recordings). It should also be noted that this (and related internet-based ECG data repositories) are collected for medical diagnostic purposes, and hence contain records with the potential for inherent variability due to the pathological status of the subjects. In particular, the PTB_1 dataset provides more than one record for 13 of its subjects (PTB_1): 7of them have only records recorded in same day and 3 of them have only records recorded several months apart, while the remaining subjects have both kind of records. This dataset therefore serves as a ‘worst’ case scenario – the results should be considered as the floor in terms of classification accuracy. From the PTB_1 dataset, 13 records were selected for training. From each record, 100 beats were utilised. 28 fiducial features are extracted and normalised from each beat, which was utilised as the training set. A two-layer feedforward neural network was trained separately with both normalised and non-normalised training data, to quantify the effect of normalisation on the results. The network was tested using 23 records. In this experiment, both individual heartbeats as well as subject recognition was measured. A subject was considered correctly classified if more than half of his/her beats were correctly classified and heartbeats were classified by majority voting. In order to evaluate the stability (temporal) of the data, the NN was trained on data acquired on a given day, and tested on records extracted from that same day or from records acquired with a significant temporal delay (in some cases more than 1 year separated training from testing records). The results for normalised and non-normalised feature sets are shown in Table 4. A FFN (trained with SoftMax) was adequate at recognising all subjects, even those for which the testing data was temporally isolated from the training data. Normalisation did not appreciably affect heartbeat recognition accuracy, though it tended to reduce subject recognition. In addition, a Radial Basis Function (RBF) network was deployed, in order to compare another neural network approach commonly deployed in literature reports. The results were generally superior to the FFN based approach (trained using the SoftMax algorithm), and in this case, non-normalised data yielded better heart beast and subject recognition accuracy (see Table 5 for details).

Table 4 The effect of normalisation on classification accuracy of proximal (same day) and distal records (those recorded after a significant time delay from the training set), on heartbeat and subject classification accuracy

	Same day records		*Long time apart*	
	Subject recognition accuracy %	*Heartbeat recognition accuracy %*	*Subject recognition accuracy %*	*Heartbeat recognition accuracy %*
Normalisation	100 (10/10)	94.8	83.3 (5/6)	87
No Normalisation	100 (10/10)	96.5	100 (6/6)	79.4

Note: These results were obtained using a feedforward neural network trained using the SoftMax algorithm

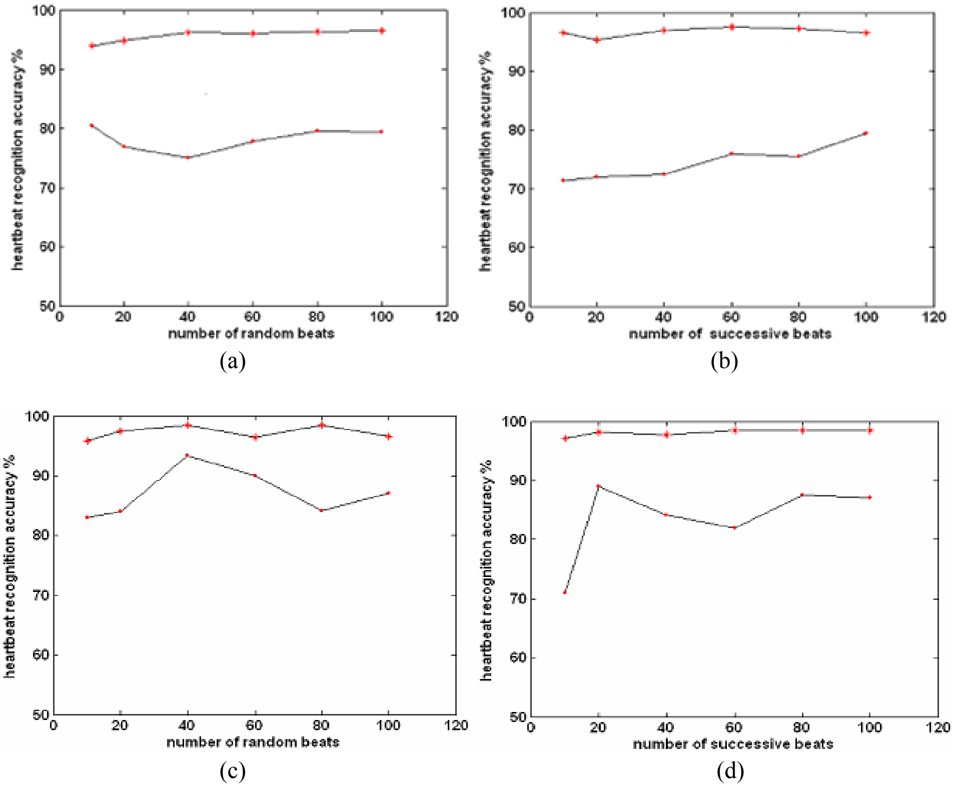
Table 5 The same experiment that produced the data in Table 4, except an RBF neural network was utilised for the classification task. All other aspects of the experiment that produced these results was the same as those used to produce the results presented in Table 4

	Same day records		*Long time apart*	
	Subject recognition accuracy %	*Heartbeat recognition accuracy %*	*Subject recognition accuracy %*	*Heartbeat recognition accuracy %*
Normalisation	100 (10/10)	97.8	83.3 (5/6)	82
No Normalisation	100 (10/10)	98.7	100 (6/6)	86.3

4.3 Reducing training set

This experiment investigated the effect of the number of heartbeats used for training on the resulting classification accuracy. Two sorts of selection criteria were utilised: (a) selecting the beats successively from each record (100 beats); (b) selecting heartbeats randomly from the whole record. Both classifiers were trained in the same way discussed in Section 4.2, except for the number of beats utilised for training, which was varied from 10 to 100 beats/subject. The results for both classifiers show that reducing the number of training beats for each subject only affect the heartbeat recognition accuracy, while the subject identification accuracy was not affected. Figures 9 and 10 present the impact of reducing the number of training beats on heartbeat recognition accuracy achieved by both classifiers with non normalised and normalised feature sets. A summary of the best results achieved is provided in Table 6.

Figure 9 The impact of reducing the number of training beats on heartbeat recognition accuracy. The left column presents un-normalised results, while the right presents the normalised results. The upper row of figures present the FFN-based results while the lower row the results obtained by the RBF classifier. Further, the upper line in each graph shows accuracies achieved by testing with 'proximal' records and the lower line shows accuracies achieved by testing with 'distal' records. In (a) and (c) beats are randomly chosen, while in (b) and (d) beats are successively chosen (see online version for colours)



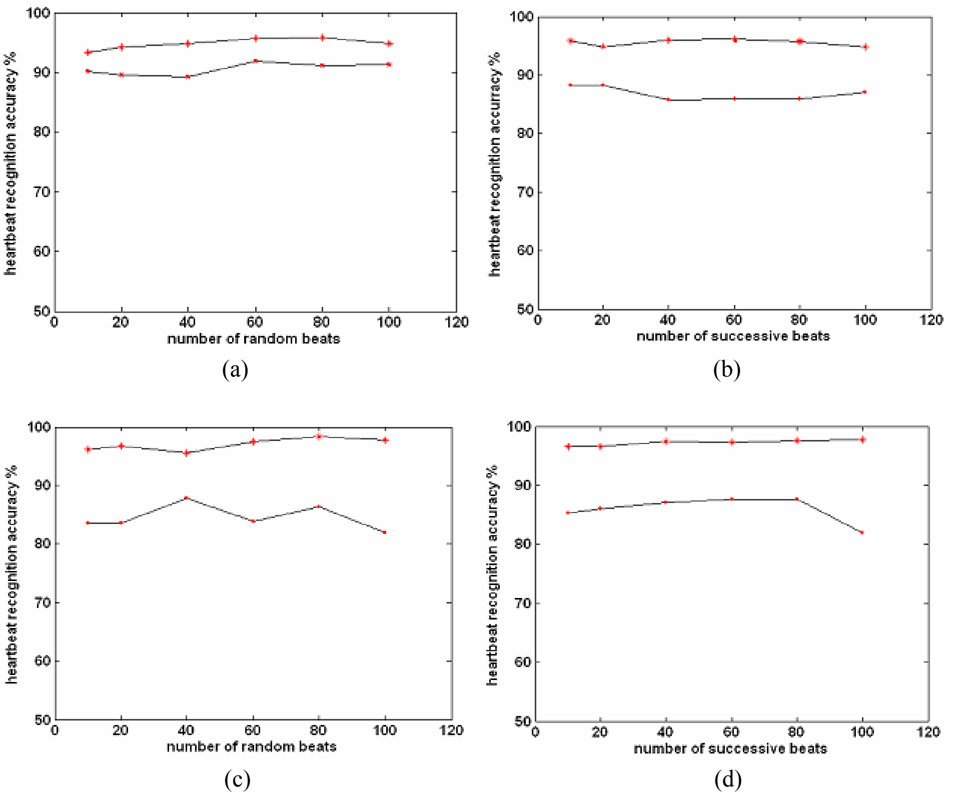
(a)

(b)

(c)

(d)

Figure 10 Depicts the impact of reducing the number of training beats on heartbeat recognition accuracy in case of normalised feature set. Figures (a) and (b) depict the FFN results, while (c) and (d) depict the RBF results. The upper line in each graph shows accuracies achieved by testing with 'proximal' records and the lower line shows accuracies achieved by testing with 'distal' records. In (a) and (c) beats are randomly chosen, while in (b) and (d) beats are successively chosen (see online version for colours)



(a) (b)

(c) (d)

Table 6 Summary of the best classification accuracies from the FFN & RBF with reduced training sets

	Same day records		*Long time apart*	
	Subject recognition accuracy %	*Heartbeat recognition accuracy %*	*Subject recognition accuracy %*	*Heartbeat recognition accuracy %*
FFN + normalised features (10 successive beats)	100 (10/10)	95.8	100 (5/6)	88.3
RBF + non-normalised features (40 random beats)	100 (10/10)	98.4	100 (6/6)	93.3
RBF + normalised features (80 random beats)	100 (10/10)	98.3	83.3 (5/6)	92.2

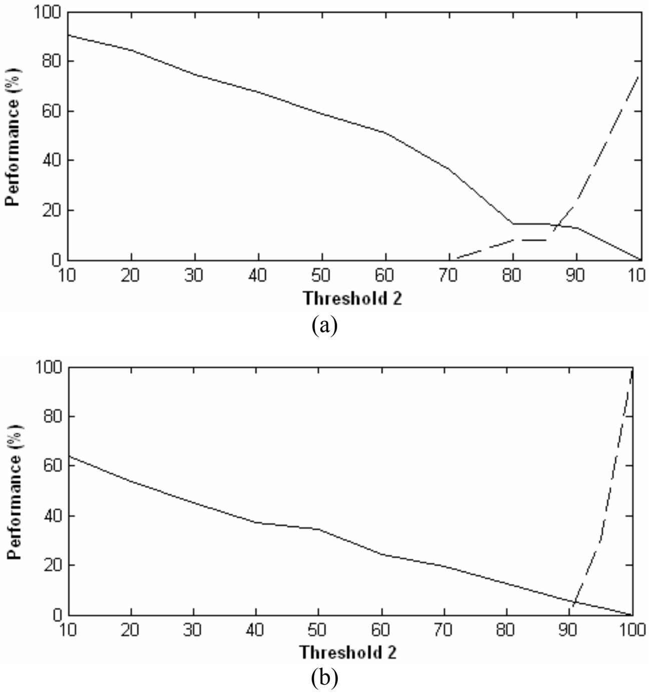
The results of this experiment indicate that the RBF classifier performed better than a FFN trained with the SoftMax algorithm. Further, paradoxically, reducing the number of training beats for each subject not only preserves the results but also improves the heartbeat recognition accuracy. In addition, this experiment also suggests that normalisation of the data reduces the classification accuracy. As shown in Figure 10, utilising the normalised feature set provides more stable results along the considered range and also decreases the gap between the accuracies achieved by testing with 'distal' and 'proximal' records in comparison with those achieved by non normalised set (cf. Figure 9). Finally, choosing the beats randomly or successively had no significant impact on the classification results with normalised feature set, however that was not the case with the non normalised feature set.

4.4 Measuring FAR, FRR and EER

In spite of their significance in assessing any biometric system, few published reports have provided values for the FAR, FRR and ERR, key measures of the suitability of a biometric. One must keep in mind that when calculating these measures, the training and testing data should be quite distinct whenever and as much as possible. For instance, calculating FRR or FAR from the training set may provide a biased result. In the context of ECG based authentication schemes, one should consider the deployment of two separate thresholds. One is typically used (call it Θ1) for classifying individual heartbeats, while another threshold (call it Θ2) represents the minimum number of correctly classified beats out of a fixed number of testing beats needed for a subject to be considered identified. From our initial trials for measuring for measuring FAR, it was found that FAR decreases significantly when the number of beats for training was reduced to 10 beats for each subject (while maintaining a reasonable value for FRR). The system developed by Singh and Gupta (2006) appears to be one of the few reports in the available literature that reported FAR which was calculated from subjects not contained in the training set. However, it is not mentioned how many impostors were considered in the test and also the FRR was measured using beats from the same training records. For comparison, though the FRR results in this paper were calculated by using the remaining beats of training records (90 beats), the FAR was calculated from subjects in three databases: 38 subjects of PTB_2, 24 subjects of MIT_BIH and 40 subjects of Fantasia (100 beats for each subject). For both classifiers the optimum value for Θ1 is 0.5. While for Θ2, 80 and 90 beats are the optimum values for FFN and RBF respectively as shown in Figure 11. The results in Table 7 demonstrate the superiority of RBF with normalised features, while Table 8 depicts the FAR for each database separately. The average FAR calculated in this work is higher than that measured by Singh and Gupta (2006), which may reflect the increase in the number of impostors (more than one hundred), acquired from different databases. This approach may enhance the reliability of the results and it should be noted that the FAR was 0% when deploying only the data (impostors) from the MIT_BIH database.

In order to evaluate the veracity of the results, the RBF was re-deployed using the same testing records utilised in previous experiments (Sections 4.2 and 4.3), and the FRR values were computed. Figure 12 shows the variation of FRR and FAR with Θ2 when Θ1 is 0.5 and 0.4 respectively. The best results were achieved with Θ1=0.4 and Θ2 = 90. The FRR increases to 7.69% which means one authorised subject is not allowed to log in, while the FAR increases to 9.8 % (due to considering Θ1 with lower value (0.4) in order to preserve FRR) and EER equals to 9.8%. Table 9 shows FAR for each database separately.

Figure 11 Results obtained for FAR (solid line) and FRR (dashed line) varies with Θ2 when Θ1 equals 0.5 using FFN in (a) and RBF in (b)



(a)

(b)

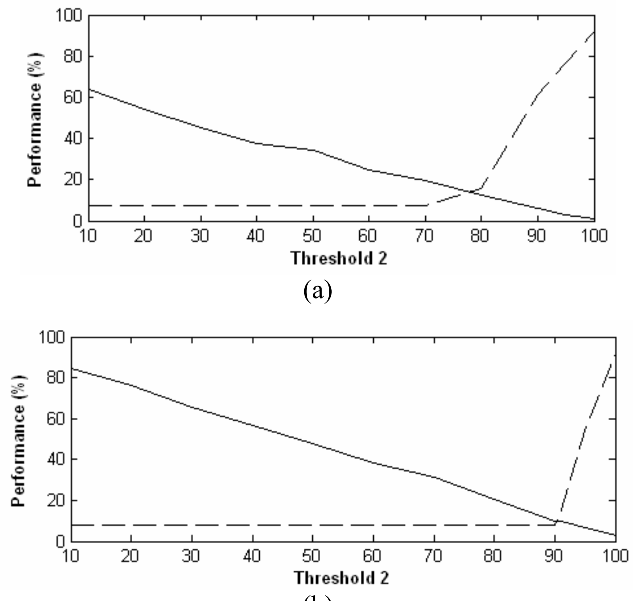
Table 7 The best values for FRR and FAR achieved using FFN and RBF, given values for thresholds (see text for details)

	FRR (%)	*FAR (%)*	*EER (%)*
FFN with 10 non-normalised features	7.69 (1/13)	14.7 (17/102)	13.7
RBF with 10 normalised features	0 (0/13)	5.8 (6/102)	4.9

Table 8 Summary of the smallest FAR value for each database separately using RBF

Database	*FAR (%)*	
PTB_2	3/38	7.89%
MIT_BIH	0/24	0%
Fantasia	3/40	7.5%
Total	*6/102*	*5.58%*

Figure 12 Relationship between the magnitude of the FAR (solid line) and FRR (dashed line) changes as a function of Θ2 when trained using the RBF, with Θ1 set to to 0.4 (a) and 0.5 (b)



(a)

(b)

Table 9 A summary of the FAR for each database separately using RBF with Θ1 = 0.4 and Θ2 = 90

Database	*FAR (%)*	
PTB_2	4/38	10.5%
MIT_BIH	0/24	0%
Fantasia	6/40	15%
Total	*10/102*	*9.8%*

4.5 Generalisation to other standard databases

In the context of biometrics, it is very crucial that the system is able to generalise across large populations of unseen data, as the cost for misclassification can be very expensive. In the present context, testing for generalisation entails training on one dataset and testing the system with data from a different dataset. This is one of the key benefits of the PhysioNet suite of ECG databases – it provides a wide range of ECG data acquired under a variety of conditions. Factors such as age, gender, length of time over which data from the same subject was collected, and number of replicates provides a wide range of factors that can be explored with respect to their impact on the generalisability of a classifier. The number of potential combinations of factors presented by the PhysioNet database is very large, so only a few of the factors were examined in this study. Further, it must be noted, that in the approach adopted in this work, the model parameters were acquired from a particular dataset, and used in testing mode on other datasets without any modification of the parameters used to develop the model. After instantiating the parameters of the system and the thresholds as obtained from previous experiment, the generalisability of the classifier was tested by their application of two databases: PTB_2 and Fantasia. First for the PTB_2, since there is only one record for each subject, we trained the system using 10 beats and we used the remaining beats of the record for testing FRR (keeping the same neural network architecture and training parameters such as momentum etc.). While for the FAR, the subjects of PTB_1, MIT_BIH and Fantasia are considered. The system achieved FRR = 2.5%, while the corresponding was FAR = 10.3%. Since the Fantasia database contains one two-hour record is available per subject, 10 randomly selected beats were used for training and 100 (randomly selected beats) were used for testing and measuring FRR. In order to measure the FAR, data from subjects of the PTB_2 were used instead of those of Fantasia which were used for training. The system achieved FRR = 10%, with an FAR = 9.3%. Tables 10 and 11 show the FAR for each database separately using RBF trained with PTB_2 and Fantasia datasets respectively. These results produce FAR/FRR values that are reasonably high, but one must remember that the training parameters of the neural networks were not tuned for each dataset, that the architectures remained the same across all the tests, that the training and testing data were derived from very different datasets whenever possible. That the data used in this experiment probably represents a more complete picture of a population at large, these results are to be considered more realistic than what which would be obtained by using a dataset tuned model.

Table 10 Summary of the FAR for each database separately using RBF trained with PTB_2 dataset

Database	*FAR (%)*	
PTB_1	1/13	7.69%
MIT_BIH	3/24	12.5%
Fantasia	4/40	10%
Total	*8/77*	*10.3%*

Table 11 Summary of the FAR for each database separately using RBF trained with Fantasia dataset

Database	*FAR (%)*	
PTB_2	5/38	13.15%
MIT_BIH	2/24	8.34%
PTB_1	0/13	0%
Total	*7/75*	*9.3%*

5 Conclusion

ECG-based biometrics has become an established methodology which relies on the individuality of cardiophysiology. The approach requires that a series of heartbeats are collected from the subject, which although inconvenient, does not pose any health risks. Once the data is collected, a model of the subject is constructed and subsequently used for the authentication task. The model is generated by acquiring pertinent features from the ECG data – using either fiduciary and/or appearance based. This study focused solely on the deployment of the fiducial approach, which requires that a set of landmarks be identified, which is quantified according to their magnitude and/or timing. These fiducials can be used to build a model of the subject – or may be used directly, as was the case in this study. The fiducials then are used to generate a classifier which is used for the purposes of automated authentication. In the present study, the ECG data was acquired from an internet based repository of medical ECG data – the PhysioNet ECG databases, which contains a collection of ECG datasets.

In the development of this ECG-based biometric system, crucial issues such as the stability of ECG, quantitative metrics such as FAR and FRR, and the ability to generalise were investigated. These issues play a fundamental role in the assessment of the reliability and robustness of a biometric system. A set of 28 fiducial features that represent the majority of features utilised in the literature were used in this study. Two 'mainstream' classifiers (FFN and RBF neural networks) were deployed as they are easy to implement and are general purpose. Both classifiers were trained with normalised and un-normalised feature sets extracted from PTB_1 records and tested with records recorded either on the same day or after several months (to several years in many cases). The results show that both classifiers have the ability to recognise all subjects tested using proximal ('same day') records and also distal ('long time apart') records but with lower heartbeat recognition accuracy (for the latter case). In general, normalisation tended to reduce the classification accuracy somewhat – which is unexpected, as this pre-processing step tends to eliminate variations due to heart rate variability. This scaling (similar to the effect of normalisation) may reduce the effects of heart rate variability. In addition, the effect of reducing the number of training beats for each subject was examined. This parameter is important and will certainly influence the classification accuracy to a certain extent. The RBF outperforms the FFN in this experiment and normalising the features provides higher and more stable heartbeat recognition accuracy for distal records than without normalising regardless of the criteria used for choosing the training beats. Moreover, the RBF with the normalised feature set achieved the best results using ten beats only for training when FAR & FRR are considered. The FAR was computed from data selected randomly (approximately 100 impostors) from three different databases (PTB_2, MIT_BIH and Fantasia), while the FRR was computed with beats from same training records. The results were 0%, 5.8% and 4.9% for FRR, FAR and EER respectively. Later, the FRR was computed with beats from another records recorded in same day or after few months (years) and the results were 7.6%, 9.8% and 9.8% for FRR, FAR and EER respectively. In addition, the generalisability of the classifier (RBF) was examined by utilising training and testing data from different databases, keeping the same model parameters and architectures. After fixing the parameters and the final values of thresholds defined using the PTB_1 dataset, the system was trained with a subset of the databases and used the other datasets for impostor testing. FAR = 9.3% and FRR = 10% are achieved when Fantasia dataset is used for training. While in case of using PTB_2 dataset for training we achieved FAR = 10.3% and FRR = 2.5%. These FAR and FRR results are not as stellar as many, but considering how they were obtained, probably reflect a more accurate estimate of PhysioNet ECG data repository when treated as if it reflects a true and relatively large population of subjects.

As a derivative result from this work, the applicability of the PhysioNet ECG data repositories has been examined. There is a wide variation in the coverage of the major PhysioNet ECG datasets. Some contain only single records from the same individual, some contain both short-term and long-term recordings, and some have a considerable amount of variation when long records are collected from the same individual. These issues must be taken into account if the data repository is to be used as the subject data for development of a biometric system. Comparisons between studies in terms of quantitative results, when they are reported, must be considered in light of the exact samples used in the development and testing phases of the system. Comparing records obtained from differing time scales will certainly have an impact on the results – whether one only uses training and test samples that are temporally close or spread apart will certainly yield an impact on the FAR/FRR results. This issue of data stability must be highlighted in studies in order to allow different classification approaches to be compared on the same ground. The issue of sampling frequency, noise and other signal acquisition factors may play a minor role, as there are many tried and tested approaches that can be applied to ameliorate these factors. But clearly the health status of the subjects deployed in the study, the duration between ECG acquisitions, the number of replicates, training/testing paradigm will have a significant impact on the results. These issues draw attention to the anticipated generalisability of the resulting biometric system. Data repositories, which are a luxury in many domains, provide the opportunity to estimate the search space a classifier can properly explore. In biometrics, generalisability is an extremely critical issue – as one tends to develop a system using a small number of subjects, and then deploy it on a much larger population. Unfortunately, much of the relevant biometrics literature has ignored this aspect – and it was the intended purpose of this paper to simply highlight the potential of internet based data repositories for such an exploration. This effort will be continued in our lab, where issues such as relevance of and the dimensionality of the features, the role of the classifier with respect to the classification task, and the role of the health and emotional status of the subjects, will be explored in depth.

References

Agrafioti, F. and Hatzinakos, D. (2008) 'ECG based recognition using second order statistics', *6th Annual Conference on Communication Networks and Services Research (CNSR)*, 5–8 May, Halifax, Canada.

Biel, L., Petersson, O. and Philipson, L.P. (2001) 'Wide: ECG analysis: a new approach in human identification', *IEEE Transactions on Instrumentation and Measurement*, Vol. 50, No. 3, pp.808–812.

Boumbarov, O., Velchev, Y. and Sokolov, S. (2009) 'ECG personal identification in subspaces using radial basis neural networks, intelligent data acquisition and advanced computing systems: technology and applications', *IDAACS 2009, IEEE*, pp.446–451.

Bousseljot, R., Kreiseler, D. and Schnabel, A. (1995) Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das Internet', *Biomedizinische Technik*, Band 40, Ergänzungsband 1,S 317

Chan, A., Hamdy, M.H., Badre, A. and Badee, V. (2005) 'Wavelet distance measure for person identification using electrocardiograms', *IEEE Transaction on Instrumentation and Measurement*, Vol. 57, No. 2, pp.248–253.

Chan, A., Hamdy, M., Badre, A. and Badee, V. (2008) 'Person identification using electrocardiograms', *IEEE Transactions on Instrumentation and. Measurement*, Vol. 57, No. 2, pp.248–253.

Chen, S. and Chang, E. (1996) 'Regularized orthogonal least squares algorithm for constructing radial basis function networks', *International Journal of Control*, Vol. 64, No. 5, pp.829–837.

Chuang, C., Hsu, C. and Chien, C. (2005) 'A novel personal authentication approach using one-lead ECG signal', *IEEE ICSS2005 International Conference on Systems & Signals*.

Fatemian, S. and Hatzinakos, D. (2009) 'A new ECG feature extractor for biometric recognition', *Proceedings of the 16th international conference on Digital Signal Processing*, IEEE Press Piscataway, NJ, USA, pp.323–328.

Gahi, Y., Lamrani, M., Zoglat, A., Guennoun, M., Kapralos, B. and El-Khatib, K. (2008) 'Biometric identification system based on electrocardiogram data, new technologies, mobility and security', *NTMS'08*, 5–7 November, pp. 1–5.

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., and 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.

Guennoun, M., Abbad, N., Talom, J., Rahman, M. and El-Khatib, K. (2009) 'Continuous authentication by electrocardiogram data, science and technology for humanity (tic-sth), *IEEE Toronto International Conference*, pp.40–42.

Haykin S. (1998) *Neural Networks*, 2nd ed., Prentice Hall.

Illanes-Manriquez, A. and Zhang, Q. (2008) 'An algorithm for robust detection of QRS onset and offset in ECG signals', *Computers in Cardiology*, Vol. 35, pp.857−860.

Irvine, J.M. and Israel, S.A. (2009) 'A sequential procedure for individual identity verification using ECG', *Journal on Advances in Signal Processing*, Vol. 2009, No. 13, pp.123–132.

Israel, S.A., Irvine, J.M., Cheng, A., Wiederhold, M.D. and Wiederhold, B.K. (2005) 'ECG to identify individuals', *Pattern Recognition*, Vol. 38, No. 1, pp.133–142.

Iyengar, N., Peng, C-K., Morin, R., Goldberger, A.L. and Lipsitz, L.A. (1996) 'Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics', *American Journal of Physiology*, Vol. 271, pp.1078–1084.

Khalil, I. and Sufi, F. (2008) 'Legendre polynomials based biometric authentication using QRS complex of ECG', *4th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP*, Melbourne.

Kunzmann U, Schochlin, G. and Bolz, A. (2002) 'Parameter extraction of ECG signals in real-time', *Biomed Tech (Berl)*, Vol. 4, No. 2, pp.875–878.

Kyoso, M. and Uchiyama, A. (2001) 'Development of an ECG identification system', *Proceedings of 23rd IEEE Engineering in Medicine and Biology Conference*, Vol. 4, pp.3721–3723.

Laguna, P., Jané, R. and Caminal, P. (1994) 'Automatic detection of wave boundaries in multi-lead ECG signals: validation with the CSE database', *Computers and Biomedical Research*, Vol. 27, No. 1, pp.45–60.

Laguna, P., Mark, R.G., Goldberger, A.L. and Moody, G.B. (1997) 'A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG, *Computers in Cardiology*, pp.673–676.

Li, M. and Narayanan, S. (2010) 'Robust ECG biometrics by fusing temporal and cepstral information', *Proceedings of 20th International Conference on Pattern Recognition (ICPR), Turkey*.

Martinez, J.P., Almeida, R., Olmos, S., Rocha, A.P. and Laguna, P. (2004) 'A wavelet-based ECG delineator: evaluation on standard databases', *IEEE Transactions on Biomedical Engineering*, pp.570–581.

Martínez, A., Alcaraz, R. and Rieta, J. (2010) 'Application of the phasor transform for automatic delineation of single-lead ECG fiducial points', *Physiological Measurement Journal*, Vol. 31, No. 11, pp.1467–1485.

Pan, J. and Tompkins, W.J. (1985) 'A real time QRS detection algorithm', *IEEE Transactions on Biomedical Engineering*, Vol. 33, No. 3, pp.230–236.

Plataniotis, K., Hatzinakos, D. and Lee, J.K.M. (2006) 'ECG biometric recognition without fiducial detection', *Proceedings of Biometrics Symposiums (BSYM'06)*, September, Baltimore, MD, USA.

Revett, K. (2008) *Behavioral Biometrics: A Remote Access Approach*, John Wiley & Sons.

Shen, T.W., Tompkins, W.J. and Hu, Y.H. (2002) 'One-lead ECG for identity verification', *Proceedings of 2nd Joint EMBS/BMES Conference*, pp.62–63.

Shen, T.W. (2005) *Biometric Identity Verification Based on Electrocardiogram (ECG)*, PhD. Thesis, University of Wisconsin, Madison.

Singh, Y.N. and Gupta, P. (2006) *ECG to Individual Identification*, Prabhu Goel Security Center, IIT, Kanpur.

Singh, Y.N. and Gupta, P. (2009) 'Biometrics method for human identification using electrocardiogram', in Tistarelli, M. and Nixon, M.S. (Eds): *ICB 2009, LNCS 5558*, pp.1270–1279.

Singla, S. and Sharma, A. (2010) 'ECG based biometrics verification system using LabVIEW', *Songklanakarin Journal of Science and Technology*, Vol. 32, No. 3, pp.241–246.

Sufi, F. Khalil, I. and Habib, I. (2010a) 'Polynomial distance measurement for ECG based biometric authentication', *Security and Communication Networks*, Vol. 3, No. 4, pp. 303–319.

Sufi, F., Khalil, I., and Hu J., (2010b) *ECG based Authentication, Handbook of Information and Communication security*, Springer.

Tawfik, M., Selim, H. and Kamal, T. (2010) 'Human identification using time normalized QT signal and the QRS complex of the ECG', *Proceedings of the 7th International Symposium on Communication Systems Networks and Digital Signal Processing, CSNDSP*, 21–23 July, Newcastle upon Tyne, pp.755–759.

The CSE Working Party (1985) 'Recommendations for measurement standards in quantitative electrocardiography', *European Heart Journal*, Vol. 6, pp.815–825.

Wan, Y. and Yao, J. (2008) 'A neural network to identify human subjects with electrocardiogram signals', *Proceedings of the World Congress on Engineering and Computer Science 2008, WCECS 2008*, 22–24 October, San Francisco, USA.

Wang, Y., Agrafioti, F., Hatzinakos, D. and Plataniotis, K. (2008) 'Analysis of human electrocardiogram for biometric recognition', *EURASIP Journal on Advances in Signal Processing*, Vol. 1, pp.1–6.

Wao, J. and Wan, Y., (2010) 'Improving computing efficiency of a wavelet method using ECG as a biometric modality', *International Journal of Computer and Network Security*, Vol. 2, No. 1, pp.15–20.

Zarrini, M. and Sadr, A. (2009) 'A real-time algorithm to detect inverted and symmetrical T-wave', *2nd International Conference on Computer and Electrical Engineering*, 28–30 December, Dubai, pp. 318–322.

Zhang, Z. and Wei, D., (2006) 'A new ECG identification method using Bayes' theorem', *TENCON 2006, IEEE Region 10 Conference*, pp.1–4.