Signal Quality Indices and Data Fusion for Determining Clinical Acceptability of Electrocardiograms

GD Clifford[1], J Behar[1], Q Li[1,2] and I Rezek[1]

[1] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

[2] Institute of Biomedical Engineering, School of Medicine, Shandong University, Jinan, Shandong, 250012, China

E-mail: gari.clifford@eng.ox.ac.uk

Abstract. A completely automated algorithm to detect poor quality ECGs is described. The algorithm is based on both novel and previously published signal quality metrics, originally designed for intensive care monitoring. The algorithms have been adapted for use on short (5-10s) single-lead and multi-lead ECGs. The metrics quantify spectral energy distribution, higher order moments and inter-channel and inter-algorithm agreement. Seven metrics were calculated for each channel (84 features in all) and presented to either a multi-layer perceptron (MLP) artificial neural network or support vector machine (SVM) for training on a multiple-annotator labelled and adjudicated training data set. A single lead version of the algorithm was also developed in a similar manner.

Data were drawn from the Physionet Challenge 2011 dataset where binary labels were available, on 1500 12-lead ECGs indicating whether the entire recording was acceptable or unacceptable for clinical interpretation. We re-annotated all the leads in both the training set (1000 labelled ECGs) and test data set (500 12-lead ECGs where labels were not publicly available) using two independent annotators, and a third for adjudication of differences.

We found that low quality data accounted for only 16% of the ECG leads. To balance the classes (between high and low quality) we created extra noisy data samples by adding noise from Physionet's noise stress test database (NSTDB) to some of the clean 12-lead ECGs. No data were shared between training and test sets.

A classification accuracy of 98% on the training data and 97% on the test data was achieved. Upon inspection, incorrectly classified data were found to be borderline cases which could be classified either way. If these cases were more consistently labelled, we expect our approach to achieve an accuracy closer to 100%.

PACS numbers: 07.05.Mh, 07.50.Hp, 87.68.+z, 87.80

Keywords: ECG, Machine Learning, mHealth, Neural Networks, Signal Quality

Submitted to: *Physiol. Meas.*

1. Introduction

The explosion of mHealth in both abundant and resource-constrained countries is both a cause for concern and for celebration (Fraser H S and Joaquin B, 2010; Gerber T et al., 2010; Waegemann C P, 2010; T and S, 2011). While mHealth clearly has the potential to deliver information and diagnostic decision support to the poorly trained, it is not appropriate to simply translate the technologies which the trained clinician uses into the hands of non-experts. In particular, it is important that the explosion of access does not lead to a flooding of the medical system with low quality data and false negatives. Clearly for mHealth to expand, a paradigm shift in the way data is analysed must occur. Although telehealth has the potential to connect remote users with little training to trained experts, with patient to doctor ratios as low as 50,000:1 in parts of low-income countries, automated algorithms will be essential to cope with the number of recordings that are likely to be made available. Moreover, since the greatest (and often the only) chance for improving the quality of physiological data is at source, a rapid feedback to the recordist or user concerning the clinical viability of the data is needed. Therefore, data screening must occur at the front end using automated algorithms, prompting the user to re-take recordings when quality is low. Problems of noise are not just confined to ambulatory ECGs however, and issues of noise plague hospitals, causing high levels of false alarms Aboukhalil et al. (2008).

This article addresses these issues in the context of the specific problem of vetting the quality of electrocardiograms (ECGs) collected by an untrained user in ambulatory scenarios. The system described here is intended to provide real-time feedback on the diagnostic quality of the ECG and prompt an inexperienced or lay user to make adjustments in the recording of the data until the quality is sufficient that an automated algorithm or medical expert may be able to make a clinical diagnosis, primarily of arrhythmias, but also for more subtle phenomena such as ischemia.

This was the subject of the PhysioNet/Computing in Cardiology (CinC) challenge 2011 (Silva I et al., 2011), which further specified that the algorithm should be efficient enough to be able to run in near real-time on a mobile phone. At a minimum, the software should be able to indicate within a few seconds, while the patient is still present, if the ECG is of adequate quality for interpretation, or if another recording should be made. Ideally, the software should identify common problems (such as misplaced electrodes, poor skin-electrode contact, external electrical interference, and artefact resulting from patient motion) and either compensate for these deficiencies or provide guidance for correcting them.

Although relatively little has been published on ECG signal quality assessment and noise estimation methods, early published work by Moody and Mark (Moody G B and Mark R G, 1989) used the residual after projecting a QRS complex onto the first five principal components (PCs) of an ensemble of QRS complexes, or the Karhunen-Loève transform (KLT). Standard noise measurement methods for ECG, which can be used as individual signal quality metrics, were reviewed by Clifford *et al* (Clifford G

D et al., 2006), including route mean square (RMS) power in the isoelectric region, ratio of R-peak to noise amplitude in the isoelectric region, the Crest factor or peak-to-RMS ratio and the ratio between in-band (5-40 Hz) and out-of-band spectral power. Concurrently, Li *et al.* (Li Q et al., 2008) developed signal quality metrics which included the ratio of power in various bands of the spectrum, the degree of agreement between different QRS detectors, the degree of agreement between beat detection on different leads, the kurtosis, and the innovation in a Kalman filter. These quality measures were calibrated to provide a mapping between a normalised signal quality index (SQI) and an expected error in heart rate. Redmond *et al* (Redmond S J et al., 2008) used signal masking methods to determine artefacts and the degree of missingness in the ECG. Three feature masks were described in their work: 1) a rail contact mask was used to mark the saturation to 0 or rail voltage; 2) a high frequency mask was obtained by using a 5th order high-pass elliptic forward-backward filter with a cut-off of 40Hz (with a fixed threshold in order to detect muscle and electrode-tissue contact noise), and; 3) a low power mask was employed by using an IIR filter with a pass-band of 0.7-33 Hz and a fixed threshold to locate low power sections in the ECG signal. The algorithm gave sensitivity of 89% and specificity of 98% respectively, although the fixed thresholds required manual tuning for different signal sources.

Multiple signal quality indices (SQIs) and associated classification methods were proposed for the CinC challenge in 2011 (event 1). Of the entrants to this event there were several notable approaches. The highest official score (93.2%) in event 1 was achieved by Xia *et al.* (Xia H et al., 2011) who used SQIs such as: flat line detection, missing lead identification, auto correlation and cross correlation thresholding.

Several other entrants scored only fractions of a percent below these entries, including Ho Chee Tat *et al.* (Ho C T et al., 2011) who notably included heart rate (HR) as one of the features; a HR which is not within the range 30-210 BPM was deemed to be of bad quality. This SQI might give interesting results but attention should be given to the robustness of QRS detector and to the HR range beyond which an ECG sample is rejected. (Arrhythmia records may end up being classified as bad quality.) Other classification methods used included heuristic rules (Moody B E, 2011; Langley P et al., 2011) and random forests (Kalkstein N et al., 2011). Our own official attempt (Clifford G D et al., 2011), which used a neural network approach to combine various SQIs (described in earlier works (Li Q et al., 2008)), achieved the second highest score a score of 92.6%. A subsequent unofficial entry (a few days after the close of the competition) achieved a score of 94.0%. However, we did not consider this improvement significant, since changing the way in which the data were distributed between the training and test sets sometimes resulted in a larger change in performance than this (1.4%) increase in performance. We therefore concluded that it would be difficult to elevate the score towards the high 90's with the existing dataset and that more data, particularly of the under-represented class (unacceptable quality), was needed.

The aim of the CinC challenge 2011 was to classify 12-lead 10s records and not individual leads. However, it is to be noted that sometimes records were judged

acceptable even if one lead was substantially noisy or even a flat line. Since a single lead approach allows the system to be independent of the number of leads recorded, we also present a generalisation of the multi-lead approach and demonstrate results on single lead ECGs for shorter segments of the ECG.

2. Methods

2.1. Data selection and labelling

Data to support development and evaluation of challenge entries were collected by the Sana Project (Celi L A et al., 2009) and provided freely via PhysioNet (Goldberger A L et al., 2000). The data set includes 1500 ten-second (10s) recordings of standard 12-lead ECGs; age, sex, weight, and possibly other relevant information about the patients; and (for some patients) a photo of the electrode placement taken using the mobile phone. Some of the recordings were identified initially as acceptable or unacceptable, but subsequently challenge participants annotated their own annotations to establish a *gold* (or perhaps *silver*) standard reference database of the quality of the recordings in the challenge data set.

The challenge data are standard 12-lead ECG recordings (leads I, II, II, aVR, aVL,aVF, V1, V2, V3, V4, V5, and V6) with full diagnostic bandwidth (0.05-100 Hz). Each lead was sampled at 500 Hz with 16-bit resolution. The leads were recorded simultaneously for a minimum of 10s by nurses, technicians, and volunteers with varying amounts of training in recording ECGs, to simulate the intended target user. In the intended application, the recordists (those making ECG recordings) will not necessarily have had experience. Since the goal of this challenge is to investigate if laypersons can be assisted via software in collecting high-quality ECGs reliably, the recordings gathered for this challenge include ECGs made by volunteers with minimal training.

The binary labels have been available on 1000 leads of the Challenge database 2011 only, the other 500 remaining blind for the participants.

The data were divided into two sets in a ratio of 2:1. The larger set of 1000 12-lead ECGs was used for training (Set-a), for which binary annotations (acceptable or unacceptable) were available. A smaller set of 500 12-leads ECGs were available for testing (Set-b), although competition entrants were blinded to the annotations. The competition required users to submit a list of the files in Set-b together with an estimated classification and an automated scorer posted results immediately.

ECGs collected for the challenge were reviewed by a group of annotators with varying amounts of expertise in ECG analysis, in blinded fashion for grading and interpretation. Between 3 and 18 annotators, working independently, examined each ECG, assigning it a letter and a numerical rating (A (0.95): excellent, B (0.85): good, C (0.75): adequate, D (0.60): poor, or F (0): unacceptable) for signal quality. The average numerical rating, $\hat{s}$, was calculated in each case, and each record was assigned to one of 3 groups:

- Group 1 (acceptable): If $\hat{s} \geq 0.70$, and $N_F \leq 1$.
- Group 2 (indeterminate): If $\hat{s} \geq 0.70$, and $N_F \geq 2$.
- Group 3 (unacceptable): If $\hat{s} < 0.70$.

(N_F is the number of grades that were marked as F.) Approximately 70% of the collected records were assigned to group 1, 30% to group 3, and fewer than 1% to group 2, reflecting a high degree of agreement among the annotators. Challenge participants were also given the opportunity to grade the ECGs in the challenge data sets.

Our team also annotated all data in both Set-a and Set-b using two independent annotators with no previous experience in annotating ECGs, and adjudicated by one engineer with over a decade experience in examining and processing ECGs. We submitted our two independently annotated classifications for both Set-a and Set-b, but not the adjudicated data (because it was not available by the deadline of the 20th July 2011). After the end of the competition, we deduced our annotations were too stringent and therefore, some annotations were adjusted to allow noisier ECGs to be accepted.

Our team also annotated each individual lead, for a total of 18,000 10s ECG segments. Furthermore, we defined (and used) an extended classification scheme, detailed in table 1, which does not render all recordings with a disconnected lead to be unacceptable. This was deemed necessary since a single missing lead should not necessarily be cause for rejection of a 12 lead ECG.

Table 1. Augmented labelling system used in this study.

Quality	Class	Description give to annotators
1.00	A	An outstanding recording with no visible noise or artefact; such an ECG may be difficult to interpret for intrinsic reasons, but not technical ones
0.75	B+	A good recording with transient artefact or low-level noise that does not interfere with interpretation; all leads recorded well
0.5	B-	Same as above with missing lead(s)
0.25	C+	An adequate recording that can be interpreted with confidence despite visible and obvious flaws, but no missing signals
-0.25	C-	Same as above with missing lead(s)
-0.5	D+	a poor recording that may be interpretable with difficulty, or an otherwise good recording with one or more missing signals
-0.75	D-	A poor recording that may be interpretable with difficulty
-1.00	F	an unacceptably poor recording that cannot be interpreted with confidence because of significant technical flaws

To map our annotations back to the competition annotations, B-, C- and D- became D, and B+ and C+ became B and C respectively. Note also that each class of acceptability was mapped to a numerical quality rating between -1 (worst quality) to +1 (best quality) in order to provide a less quantized set of targets for the MLP and to allow our continuous classifiers the option to predict individual classes. The ECGs were not pre-processed prior to annotation. All results in this paper were generated using this new set of annotations.

2.2. Balancing data

It is well-known that building classifiers using imbalanced classes, i.e. when one class greatly outnumbers the other classes, causes bias and results in a poor generalisation

ability of the classification model. When prior probabilities (and a Bayesian training paradigm) are not used to overcome this problem, the alternative is to balance the training classes. We therefore balanced the dataset by bootstrapping the unrepresented class to be equal to the more numerous class using additive real noise on clean data.

In order to generate additional noisy records we used the PhysioNet noise stress test database (NSTDB) (Moody G B et al., 1984; Goldberger A L et al., 2000) noise samples. The database contains samples for three types of noise; record *bw* contains baseline wander noise, record *em* contains electrode motion artefact with a significant amount of baseline wander and muscle noise as well. Finally record *ma* contains mainly muscle noise. For the purpose of our experiment we used *em* and *ma* noise. Baseline wander was not considered because it does not often render an ECG unacceptable. In order to prevent correlation between the noise added in the training and test sets, we divided the *em* and *ma* files in two at the midpoint of the recordings; half to be used with the training set and half with the test set.

As only two leads were provided with both the *em* and *ma* records from PhysioNet, principal component analysis (PCA) was used to generate a third orthogonal lead along the dominant principal component from the two leads in the recording. Assuming the resultant leads form an orthogonal lead set with arbitrary orientation, the Dower transformation (Dower G E et al., 1980) was then used to create realistic correlated 12-lead sets of noise. The purpose of this step was to generate 12-lead noise records with realistically correlated, but not identical noise on the different leads. This is particularly important for assessing the performance of iSQI which is based on inter-lead QRS detection. Herein, *em* and *ma* refer to 12-leads records generated from the PhysioNet samples and using PCA and the Dower transformation.

Figure 1 introduces the workflow for generating the additional noisy records. A total of 18000 leads were available from the challenge. We annotated all the leads individually, thus providing a training set composed of 2020 noisy and 9980 good quality leads. 4000 good quality leads were selected from a subset of 334 patients in the training set that had their 12 leads of good quality ($334 * 12 = 4008$). For each of these patients, 10s from a noisy record (*em* or *ma*) were selected at random and a calibrated amount of this 10s randomly selected 12-lead noise samples were added to the clean 12-lead record. The noise was added so that the signal to noise ratio (SNR) was equal to $-6db$ on each generated lead. After adding the noise and computing the SQI values, 8 leads were randomly dropped making 4000 samples for each *em* and *ma* noise. This resulted in 8000 additional leads for the training set (4000 with *em* noise and 4000 with *ma* noise). The SNR, S, was controlled by computing the coefficient a for each lead as follows:

$$y = x + a * v$$

$$a = \sqrt{\exp\left(\frac{-\ln(10) * S}{10}\right)\frac{P_x}{P_v}}$$

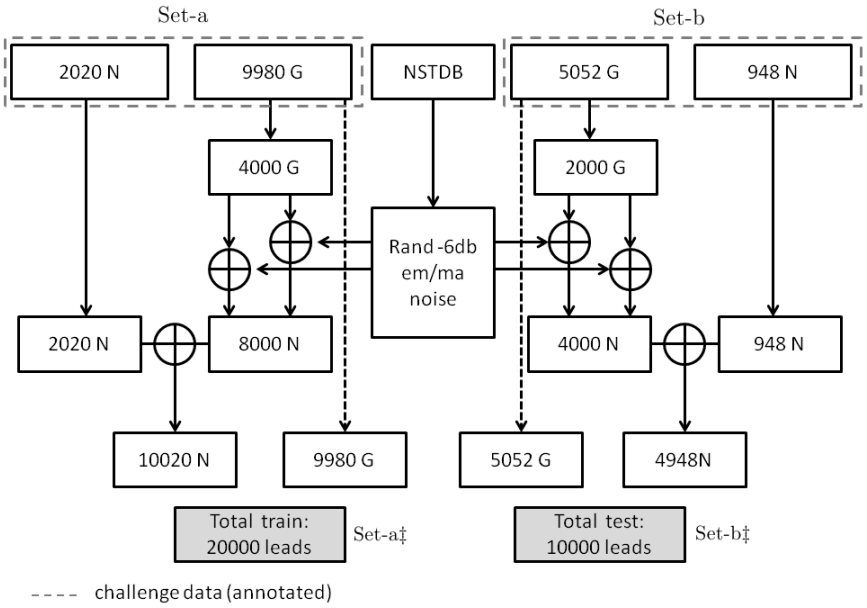

Figure 1. Flow diagram to illustrate how the data from Set-a and Set-b were generated to balance data sets Set-a‡ and Set-b‡. G: Good samples, N: Noisy samples, NSTDB: PhysioNet noise stress test database.

with y the output signal (i.e. the noisy signal generated from the clean sample), x the initial clean signal, v a 10s noisy sample from *em* or *ma* selected at random, a the amount of the noisy sample that we add to the clean signal, P_x the power of the clean signal and P_v the power of noisy signal.

The same steps were followed in order to generate 4000 bad quality leads from clean leads in the test set (Set-b). As a result we had an overall of 20000, 10s ECG samples for the training set and 10000 for the test set with about half the samples being of good quality and half being of poor quality thus resulting in balanced classes. Figure 2 gives an example of samples generated by this method when adding *em* and *ma* noise on one lead.

The training and test set for 12 leads classification were also balanced in a similar manner. The re-annotated data set included 752 good quality and 248 low quality cases in the training set and 400 good quality and 100 low quality cases in the test set. For the training set, 252 good quality cases were selected to add -6dB *em* and *ma* noise on each lead respectively to generate 504 noise cases. So the balanced training set has both 752 good quality cases and noisy low quality cases. For the test set, 150 good quality cases were selected at random from the test set to which noise was added in order to generate 300 more cases of noisy 12 lead ECG.

The balanced training and test sets are denoted Set-a‡ and Set-b‡ from herein.

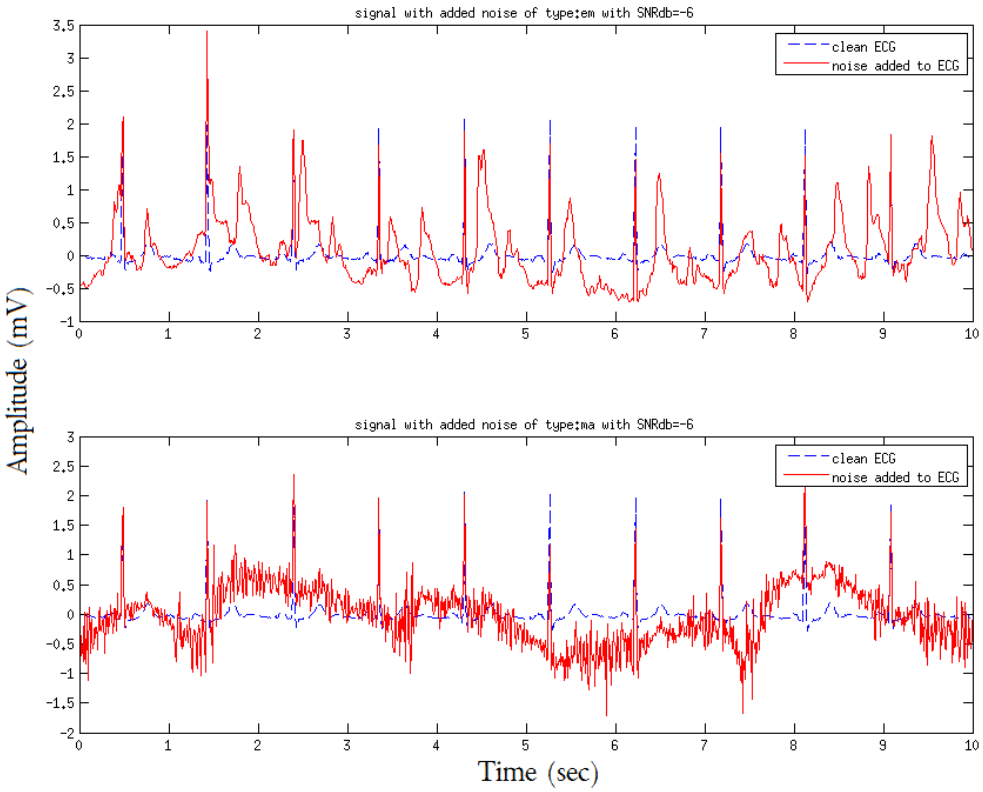

Figure 2. Examples of noisy ECG generation. The original clean ECG lead is displayed (dashed blue line) with the resultant ECG after addition of noise with SNR= $-6db$ noise (red). Upper plot illustrates the addition of *em* noise and the lower plot the addition of *ma* noise. Note that these are just examples, and the same section of ECG is never used more than once.

2.3. Pre-processing of ECGs

Each channel of ECG was downsampled to 125 Hz using an anti-aliasing filter. QRS detection was performed on each channel individually using two open source QRS detectors (*eplimited* and *wqrs*) since *eplimited* is less sensitive to noise (Li Q et al., 2008). Note that *eplimited* is a QRS detector based on Pan and Tompkins (P & T) algorithm.

2.4. Signal quality indices

Seven signal quality indices were chosen based on earlier work (Li Q et al., 2008) and run on each of the $m = 12$ leads separately, producing 84 features per recording:

(i) iSQI: The percentage of beats detected on each lead which were detected on all leads.

(ii) bSQI: The percentage of beats detected by *wqrs* that were also detected by *eplimited*.

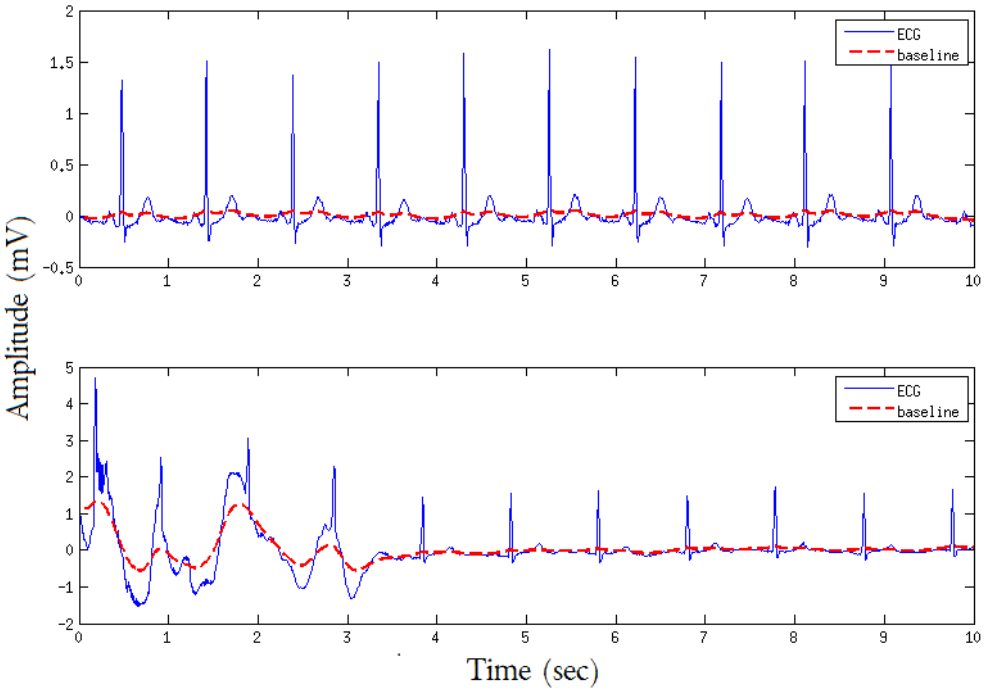

Figure 3. Example of baseline for a good quality ECG sample (upper plot) and bad quality ECG sample (lower plot) taken from Set-a of the PhysioNet/Computing in Cardiology Challenge 2011 dataset (Silva I et al., 2011). The baseline plotted corresponds to the signal filtered with a cutoff frequency at $1Hz$ using a zero-phase second order low pass filter. In this example, the good quality sample (upper plot) has a $basSQI$ (see text) of 0.966 whereas the poor quality ECG (lower plot) has $basSQI$ of 0.5.

(iii) pSQI: The relative power in the QRS complex: $\int_{5Hz}^{15Hz} P(f)\,\mathrm{d}f / \int_{5Hz}^{40Hz} P(f)\,\mathrm{d}f$.

(iv) sSQI: The third moment (skewness) of the distribution.

(v) kSQI: The fourth moment (kurtosis) of the distribution.

(vi) fSQI: The percentage of the signal x_m which appeared to be a flat line.

(vii) basSQI: The relative power in the baseline: $1 - \int_{0Hz}^{1Hz} P(f)\,\mathrm{d}f / \int_{0Hz}^{40Hz} P(f)\,\mathrm{d}f$

Note that a low $basSQI$ means that the power within the band $[0; 1Hz]$ is abnormally high with respect to the power in the $[0; 40Hz]$ interval, which is likely to be caused by an abnormal shift in the baseline. Figure 3 illustrates what type of event this SQI is designed to catch. On this example, the good quality sample (upper plot) has a $basSQI = 0.966$ whereas the poor quality ECG (lower plot) has $basSQI = 0.5$.

2.5. Machine learning for classifying ECG

Note that all the SQIs except kSQI and sSQI possess values within the range [0 1] by definition. Thus we normalized the kSQI and sSQI by subtracting the median value

(less prone to outliers than the mean) and dividing by the standard deviation. The mean and standard deviation from the training set were used for both the training and test set when normalizing.

The resultant 84 features were then used to train various machine learning algorithms to classify the data as acceptable (1) or unacceptable (-1). To compare possible inconsistencies in labelling between the sets we compared results for training on Set-a‡ and testing on Set-b‡ against training on Set-b‡ and testing on Set-a‡. We compared two different classifiers; a support vector machine (SVM), and a multi-layer perceptron (MLP) artificial neural network.

We tested two classification approaches: a single classifier trained on all 12 leads combined (to classify records) and classifiers trained on the individual leads (to classify single leads). In the 12-lead classifier the input data consisted of 84 features (7 per lead, see section 2.4) whilst the single lead classifiers were trained on the 7 features extracted for each lead individually. All classifiers were provided with the same class-labels *1:Acceptable or -1: Unacceptable*, as described in section 2.1.

2.5.1. MLP Artificial Neural Network A standard three-layer feed-forward MLP was used in which the input nodes were fully connected to the next hidden layer and in turn, to the output layer. The output layer consisted of a single node. Thus, the full network function for M hidden and D input nodes is given by

$$y(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

where $\sigma(a) = \frac{1}{1+exp(-a)}$ is the sigmoid mapping function and $w_{ji}^{(1)}$,$w_{kj}^{(2)}$ are the weights to the hidden and output layers, respectively, and $w_{j0}^{(1)}$ and $w_{k0}^{(2)}$ are respectively the input and hidden layer bias terms.

Training of the neural network (determining the values for $w_{ji}^{(1)}$,$w_{kj}^{(2)}$,$w_{j0}^{(1)}$ and $w_{k0}^{(2)}$) was based on the Levenburg-Marquat algorithm (Moré J J, 1977). The stopping criteria were; maximum of 100 epochs, or error $\leq 10^{-5}$, or gradient $\leq 10^{-5}$.

During MLP training, the training set (Set-a‡) was further divided automatically into 70% training, 15% validation and 15% testing to determine the optimal architecture. Using the validation set, the number of nodes (J) in the hidden layer was chosen to be the one which provided the highest accuracy. To that end note that the input comprised of 84 nodes and the output layer was a single node, representing the probability of good (+1) or bad (-1) quality data. Since we have at most 1504 training examples in Set-a‡ and we want the number of free parameters (weights in the MLP) to be approximately one tenth of this or less (Bishop C M, 2006), then the number of hidden nodes must be restricted to about 17 ($<$ 1504/86 if we include the bias weights). We therefore exhaustively tested 84-J-1 architectures for $J = 2, 3, 4, ...17$.

2.5.2. Support Vector Machine When using SVM, the user has to make a certain number of choices such as how the data should be pre-processed, what kernel to use and

how to choose the hyperparameters (soft margin constant C and any parameters of the kernel). Wrong choices may result in severely reduced performance (Hsu C W et al., 2010). In this work we used the libSVM library (Chang C C and Lin C J, 2011).

For the SVM we first performed test using linear and non linear kernels. As the non linear kernels performed better than the linear one we choose to work with a Gaussian (non linear) kernel defined by:

$$k(\mathbf{x}, \mathbf{x}^{'}) = \exp\left(\gamma \|\mathbf{x} - \mathbf{x}^{'}\|^2\right).$$

where γ controls the width of the Gaussian and plays a similar role as the degree of the polynomial kernel in controlling the flexibility of the resulting classifier (Ben-Hur A and Weston J, 2012). $\mathbf{x}^{'}$ and $\mathbf{x}$ are two vectors expressed in the feature space.

The SVM with an Radial Basis Function (RBF) kernel has two parameters: C and γ. The constant $C > 0$ defines the relative importance of maximizing the margin and minimizing the amount of slack. A large value of C will assign a high penalty to errors and margin errors (Ben-Hur A and Weston J, 2012). The best combination of the two parameters depends on the problem and consequently model parameters selection must be performed. As recommended by Chih-Wei Hsu *et al* (Hsu C W et al., 2010) we performed a grid search on C and γ using cross-validation and by trying exponentially growing sequences of the parameters.

3. Results

In this work we concentrate on reporting sensitivity (Se), which measures the proportion of poor quality signals that have been correctly identified as poor, specificity (Sp), which measures the proportion of good quality records that have correctly been identified as acceptable, and accuracy (Ac) corresponds to the proportion of ECGs that have correctly been classified.

3.1. Results on single lead ECGs

Initially the SVM was run on each individual SQI in order to analyse their ability to individually distinguish between good and bad quality samples. bSQI, kSQI and basSQI gave the best result with $Ac = 0.899$, $Ac = 0.917$ and $Ac = 0.919$ on the test set respectively (See Table 2). All combinations of pairs, triplets etc. of SQIs were then used in order to identify which SQI were the most relevant and also if a sub combination of the 7 SQI provided better results than all SQIs together. We found that the best result was obtained when considering four SQIs (bSQI, basSQI, kSQI, pSQI) with 0.958 accuracy on the test Set-b‡ (see Table 3).

We then performed a grid search to tune the SVM parameters using 90% of the initial training set for training and the remaining 10% as the validation set. For each combination of the parameters the area under the receiver operating characteristic (ROC) curve was computed. We identified $C = 25$ and $\gamma = 1$ as performing well on the training and validation set. We then retrained the SVM using the above parameters

(defaults were $C = 1$ and $\gamma = 0.14$) and considering the four best SQIs as previously identified. This resulted in $Ac = 0.965$ on the test set, that is an increase of close to 0.7% from using the SVM with its default parameters.

Table 4 present the final results obtained using the MLP and SVM classifiers and when considering the best four SQIs. The overall best results was achieved for the SVM with $C = 25$ and $\gamma = 1$. We achieved $Ac = 0.965$, $Se = 0.972$, $Sp = 0.958$ on the test set.

Using the MLP or SVM it is possible to return probability estimates instead of binary classification, allowing us to build the ROC curves. Figure 4, shows the ROC curves for the training set using both the MLP and SVM approach. The dots on the curves indicate the value of sensitivity and specificity that correspond to the maximum accuracy on the training set.

Table 2. Results of single lead classification process and when considering each SQI individually. Results are given for the SVM classifier.

		bSQI	iSQI	kSQI	sSQI	pSQI	fSQI	basSQI
Train Set-a‡	Ac	0.910	0.780	0.916	0.756	0.681	0.599	0.932
	Se	0.900	0.609	0.923	0.844	0.405	0.937	0.956
	Sp	0.921	0.951	0.909	0.667	0.958	0.260	0.906
Test Set-b‡	Ac	0.899	0.765	0.917	0.741	0.689	0.592	0.919
	Se	0.904	0.587	0.934	0.841	0.414	0.939	0.953
	Sp	0.894	0.940	0.900	0.642	0.958	0.250	0.886

Table 3. Single lead classification using SVM with several combinations of SQI.

	selected SQI	Ac training Set-a‡	Ac test Set-b‡
pairs	bSQI, basSQI	0.951	0.938
triplets	bSQI, basSQI, kSQI	0.962	0.956
quadruplets	bSQI, basSQI, kSQI, pSQI	0.964	0.958
quintuplets	bSQI, basSQI, kSQI, pSQI, fSQI	0.963	0.958
sixtuplets	bSQI, basSQI, kSQI, pSQI, fSQI, iSQI	0.962	0.956
all SQI		0.960	0.955

3.2. Results on 12 lead ECGs

Table 5 shows the results of classification on 12 leads of balanced dataset when considering each SQI individually with the SVM classifier using default parameters

Table 4. Single lead classifier results after optimization and when considering the best four SQIs (bSQI, basSQI, kSQI, pSQI). Best results are underlined.

	Training Set-a‡		Test Set-b‡		Training Set-b‡		Test Set-a‡	
	MLP	SVM	MLP	SVM	MLP	SVM	MLP	SVM
Ac	0.972	<u>0.971</u>	0.963	<u>0.965</u>	<u>0.973</u>	0.971	<u>0.962</u>	0.964
Se	0.971	0.974	0.958	0.972	0.977	0.978	0.955	0.960
Sp	0.972	0.963	0.968	0.958	0.969	0.964	0.969	0.968

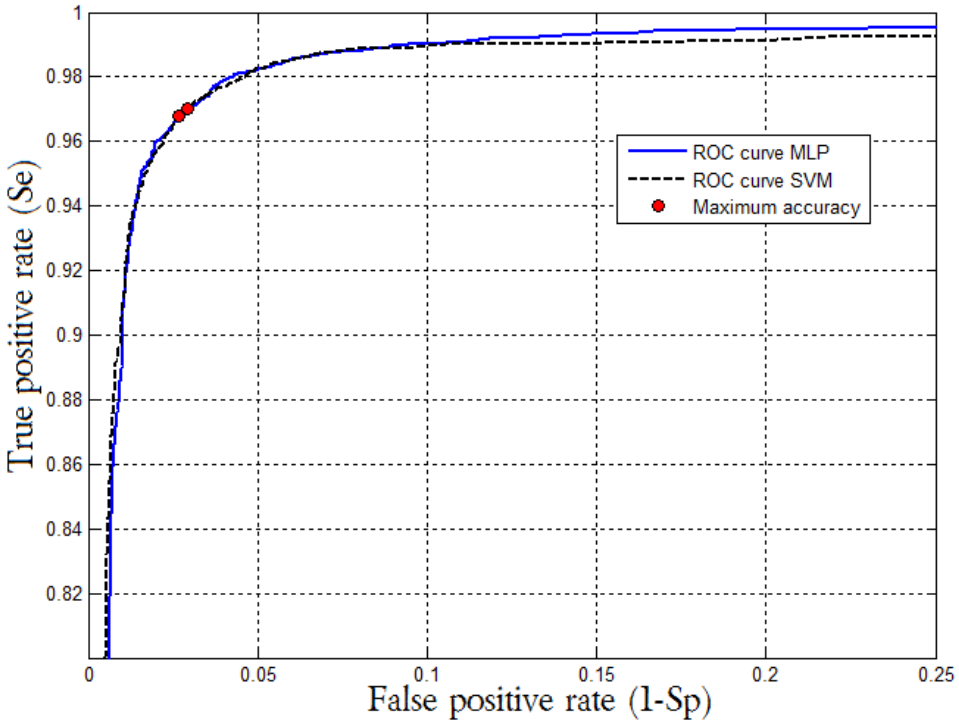

Figure 4. ROC curves for the training set (Set-a‡) using both the MLP (continuous curve) and SVM (dashed curve) approach when considering the best four SQIs on single lead data. The large red dots on the curves indicate the value of sensitivity and specificity that correspond to the maximum accuracy.

setting. bSQI, kSQI, sSQI and basSQI gave the best results with a maximum Ac of 0.93.

Table 6 shows the results of classification on 12 leads using SVM with all combinations of SQIs. We found that the best result was obtained when considering five SQIs (bSQI, basSQI, kSQI, sSQI and fSQI) with 0.949 accuracy on the test set. Then we performed grid search on SVM and hidden nodes selection on MLP to find the best results using these five SQIs and the results are shown in table 7. The MLP provided the best result (Ac=0.959) testing on Set-b‡, although the best competition entry (0.952) was found using SVM and training on Set-b‡.

Table 5. Results of classification process on 12 leads and when considering each SQI individually. Results are given for the SVM classifier.

		bSQI	iSQI	kSQI	sSQI	pSQI	fSQI	basSQI
Train Set-a‡	Ac	0.920	0.813	0.911	0.918	0.704	0.713	0.922
	Se	0.901	0.734	0.859	0.888	0.417	0.933	0.888
	Sp	0.939	0.892	0.963	0.948	0.99	0.493	0.955
Test Set-b‡	Ac	0.909	0.805	0.925	0.906	0.723	0.730	0.933
	Se	0.910	0.742	0.905	0.905	0.463	0.967	0.943
	Sp	0.908	0.868	0.945	0.907	0.983	0.493	0.923

Table 6. Classification on 12 leads using SVM with several combinations of SQI, training on Set-a‡ and testing on Set-b‡.

	selected SQI	Ac training Set-a‡	Ac test Set-b‡
pairs	iSQI, basSQI	0.940	0.945
triplets	bSQI, basSQI, pSQI	0.938	0.948
quadruplets	bSQI, basSQI, kSQI, sSQI	0.945	0.948
quintuplets	bSQI, basSQI, kSQI, sSQI, fSQI	0.949	0.949
sixtuplets	bSQI, basSQI, kSQI, sSQI, fSQI, iSQI	0.946	0.948
all SQI		0.944	0.946

Table 7. Results of selected five SQIs on 12 leads after optimising hyperparamters γ and/or C . ‡ indicates balanced data. Best results are underlined. PNet indicates the PhysioNet competition entries.

	Training Set-a		Test Set-b		Training Set-b		Test Set-a	
	MLP	SVM	MLP	SVM	MLP	SVM	MLP	SVM
Ac	0.962	0.999	0.940	0.916	0.994	0.998	0.910	0.891
Se	0.899	0.990	0.890	0.941	0.980	0.998	0.686	0.888
Sp	0.983	1	0.953	0.807	0.998	1	0.984	0.908
PNet			0.910	0.886	0.948	0.952		
Ac‡	0.978	0.997	0.959	0.953	0.995	0.999	0.928	0.934
Se‡	0.963	0.993	0.960	0.961	0.993	0.998	0.889	0.889
Sp‡	0.993	1	0.958	0.944	0.998	1	0.968	0.976
PNet‡			0.904	0.894	0.946	0.952		

3.3. Varying window length and rhythm

Two further tests were performed across the single lead data to demonstrate the generality of the algorithm. First, the window was reduced from 10s to 5s in second intervals. We noted that the MLP dropped its performance from 97% to 93% on the test data (Set-b‡).

We also tested the algorithm across data in the MIT-BIH arrhythmia database (without retraining on the arrhythmia database) and noted an average accuracy of 93% for 10s windows.

3.4. Average process time

Finally, we tested the execution time of the individual components of our algorithm to process 10s of single lead data. Table 8 provides timings when executing the routines written in Matlab, running on a 3.10 GHz Intel(R) Xeon(R) CPU E31225. Although this processor is several times faster than a traditional mobile phone processor, coding the routines in C would lead to significant improvements in execution time. However, the total execution time is only around 38ms, with the slowest routines being the QRS detectors, followed by the power estimation routine, all of which many monitoring systems routinely already compute.

Table 8. Average process time required for computing the SQI and classifying a new occurrence. Note that both basSQI and pSQI use the power spectral density so we only need to compute it one time. All the computation were performed using Matlab and a 3.10 GHz Intel(R) Xeon(R) CPU E31225.

	kSQI	sSQI	pSQI	fSQI	P&T	wqrs	SVM	MLP
Time (ms)	0.33	0.29	1.92	0.07	2.46	33.18	0.10	0.001

4. Discussion and conclusions

The method for classifying the quality of ECGs presented in this paper is a novel and completely general approach to the problem of automatically identifying trustworthy signals or events. Essentially the covariant structure of how the noise manifests in the multi-lead ECG is learned to provide a context for when a signal is trustable or not.

We made several improvements in this work to improve the general performance of the classifier over our original submission to the CinC challenge 2011. In particular we upsampled the unrepresented class (the noisy data) using a database of realistic noise (the NSTDB). This approach allowed us to improve both our internal test scores and our (unofficial) PhysioNet score. However, we note that our training and test scores are marginally higher than our PhysioNet competition scores (98%, 97% and 95% respectively). There are two major reasons for this. First, our higher results

are reported on balanced data (equal numbers of acceptable and unacceptable ECGs) to reflect the performance of the algorithm on any individual ECG. The PhysioNet Challenge data included far more acceptable ECG examples, and therefore performance scores on the challenge data reflect the performance of an algorithm on a cohort of data with a presumed quality distribution. Secondly, since we relabelled the data, we may have been more consistent than the aggregate of the competition entrants, and therefore our labels may have been easier to separate than those provided for the competition.

The achieved training accuracies of 98% and test set accuracies up to 97% indicate that extremely accurate classification of noisy ECGs is possible. We note also that our approach is only marginally affected by window size and abnormal rhythms. On inspection of the incorrectly classified data, we found that the labels were 'borderline' and could be relabelled either way and the test accuracy considered to approach 100%. Another related issue was that noise was often transient, and hence a label may apply only to a section of a window, and hence the classifier could become confused in such cases. With more accurate and consistent labelling, it is likely that the test accuracy will approach the training accuracy.

Currently we observe similar results using a MLP and SVM, although the SVM provides slightly better specificity. The MLP is faster than the SVM in execution time (after training) and may therefore provide a suitable on-line approach. However, we note that the second QRS detector is by far the largest consumer of CPU cycles in our system (see table 8). Removing the second QRS detector drops the processing time from 38ms to 5ms, with a drop in accuracy of only 0.01% on the test data (Set-b‡).

A particular strength of the approach described here, is that as more data are made available, and more data are annotated, our algorithm can be updated (by incremental training on new examples) without the need for full retraining, and hence be used adaptively in a prospective manner, profiting from user feedback. We have therefore begun to implement this system on an Android phone application for adaptive use in the field.

It should be noted that for collecting data in the field, initially it is perhaps more important to reject noisy ECGs than it is to retain clean ECGs, since falsely accepting a noisy ECG is usually worse than rejecting a clean ECG (as long as this is not too frequent) since the user is available to re-record the ECG (assuming that the algorithm does not classify an pathological recordings as noise, and hence force all subjects to look normal). As time progresses and the user learns from the algorithm what is acceptable or not, the user will be able to provide oversight for the algorithm and flag erroneous classifications. This learning feedback between human and computer may provide a general approach for all intelligent mobile applications.

A final important note is that the seven quality metrics we chose, based partially on earlier work (Li Q et al., 2008), are unlikely to be the optimal set of quality indicators, and may require changing for different contexts, equipment, diagnostic outcomes, or patient populations (particularly for patients with many arrhythmias). Indeed, metrics suggested by other entrants in the Computing in the PhysioNet Computing

in Cardiology Competition 2011 may well add additional information and improve on the results reported here. The general framework we have described in this paper is sufficiently flexible to allows for the use of an arbitrary number of quality metrics, selecting those that are most appropriate for a given situation. In associated work (Monasterio V and Clifford G D, 2012) we describe a method which employs feature selection to determine the most relevant group of features, including both physiological and noise metrics. In this way, a multidimensional threshold can be found which uses the temporal association of noise, signal and their covariances, mimicking the human approach to separating noisy from clinically useful ECGs.

References

Aboukhalil, A., Nielsen, L., Saeed, M., Mark, R. G. and Clifford, G. D. (2008). Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform, *Journal of Biomedical Informatics* **41**(3): 442–451.

Ben-Hur A and Weston J (2012). A User's Guide to Support Vector Machines, Technical report http://pyml.sourceforge.net/doc/howto.pdf.

Bishop C M (2006). *Pattern Recognition and Machine Learning*, Springer Verlag.

Celi L A, Sarmenta L, Rotberg J, Marcelo A and Clifford G D (2009). Mobile Care (Moca) for Remote Diagnosis and Screening, *Journal of Health Informatics in Developing Countries* **3**(1): 17–21.

Chang C C and Lin C J (2011). LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2**(3): 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Clifford G D, Azuaje F and McSharry P E (2006). *Advanced methods and tools for ECG data analysis*, Artech House, Norwood, MA.

Clifford G D, Lopez D, Li Q and Rezek I (2011). Signal Quality Indices and Data Fusion for Determining Acceptability of Electrocardiograms Collected in Noisy Ambulatory Environments, *Computing in Cardiology* **38**: 285–288.

Dower G E, Machado H B and Osborne J A (1980). On deriving the electrocardiogram from vectoradiographic leads, *Clinical Cardiology* **3**(2): 87–95.

Fraser H S and Joaquin B (2010). Implementing medical information systems in developing countries, what works and what doesn't, *AMIA Annu Symp Proc* **2010**: 232–236.

Gerber T, Olazabal V, Brown K and Pablos-Mendez A (2010). An agenda for action on global e-health., *Health Aff (Millwood)* **29**(2): 233–236.

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* **101**(23): e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215 PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Ho C T, Chen X and Lim E T (2011). Physionet Challenge 2011: Improving the Quality of Electrocardiography Data Collected Using Real Time QRS-Complex and T-Wave Detection, *Computing in Cardiology* **38**: 441–444.

Hsu C W, Chang C C and Lin, C J (2010). A Practical Guide to Support Vector Classification, *Technical report* .
URL: *http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf*

Kalkstein N, Kinar Y, Naaman M, Neumark N and Akiva P (2011). Using Machine Learning to Detect Problems in ECG Data Collection, *Computing in Cardiology* **38**: 437–440.

Langley P, Marco L, King S, Duncan D, Maria C, Duan W, Bojarnejad M, Zheng D, Allen J and Murray A (2011). An Algorithm for Assessment of Quality of ECGs Acquired via Mobile Telephones, *Computing in Cardiology* **38**: 281–284.

Li Q, Mark R G and Clifford, G D (2008). Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter, *Physiological measurement* **29**(1): 15–32.

Monasterio V and Clifford G D (2012). Robust Apnoea Detection in the Neonatal Intensive Care Unit, *Physiol Meas* . In Submission.

Moody B E (2011). Rule-Based Methods for ECG Quality Control, *Computing in Cardiology* **38**: 361–363.

Moody G B and Mark R G (1989). QRS morphology representation and noise estimation using the Karhunen-Loève transform, *Computers in Cardiology* **16**: 269–272.

Moody G B, Muldrow W E and Mark R G (1984). A noise stress test for arrhythmia detectors, *Computers in Cardiology* **11**: 381–384.

Moré J J (1977). The Levenberg-Marquardt Algorithm: Implementation and Theory, *in* G. A. Watson (ed.), *Numerical Analysis*, Vol. 630 of *Lecture Notes in Mathematics*, Springer Verlag, pp. 105–116.

Redmond S J, Lovell N H, Basilakis J and Celler B G (2008). ECG quality measures in telecare monitoring, *Conf Proc IEEE Eng Med Biol Soc* pp. 2869–72.

Silva I, Moody G B and Celi L (2011). Improving the Quality of ECGs Collected Using Mobile Phones: The PhysioNet/Computing in Cardiology Challenge 2011, *Computing in Cardiology* **38**: 273–276.

T, T. and S, K. (2011). Special delivery: An analysis of mHealth in maternal and newborn health programs and their outcomes around the world, *Maternal and Child Health Journal* pp. 1–10. In Press: doi:10.1007/s10995-011-0836-3.

Waegemann C P (2010). mHealth: the next generation of telemedicine?, *Telemed J E Health* **16**(1): 23–25.

Xia H, Garcia G A, McBride J C, Sullivan A, Bock T D, Bains J, Wortham D C and Zhao X. (2011). Computer Algorithms for Evaluating the Quality of ECGs in Real Time, *Computing in Cardiology* **38**: 369–372.