IOP PUBLISHING PHYSIOLOGICAL MEASUREMENT

Physiol. Meas. **33** (2012) 1419–1433 doi:10.1088/0967-3334/33/9/1419

Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms

G D 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, People's Republic of China

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

Received 29 February 2012, accepted for publication 10 May 2012
Published 17 August 2012
Online at stacks.iop.org/PM/33/1419

Abstract

A completely automated algorithm to detect poor-quality electrocardiograms (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 (5–10 s) single- 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 artificial neural network or a support vector machine (SVM) for training on a multiple-annotator labelled and adjudicated training dataset. 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 dataset (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 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 were 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%.

Keywords: ECG, machine learning, mHealth, neural networks, signal quality

(Some figures may appear in colour only in the online journal)

Table 1. Augmented labelling system used in this study.

Quality	Class	Description given 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

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 generalization 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 *et al* 1984, Goldberger *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 PC from the two leads in the recording. Assuming the resultant leads form an orthogonal lead set with arbitrary orientation, the Dower transformation (Dower *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-lead records generated from the PhysioNet samples and using PCA and the Dower transformation.

Figure 1 depicts the workflow for generating the additional noisy records. A total of 18 000 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. Four thousand 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, 10 s from a noisy record (*em* or *ma*) were selected at random and a calibrated amount of these 10 s randomly selected 12-lead noise samples were added to the clean 12-lead record. The noise was added,

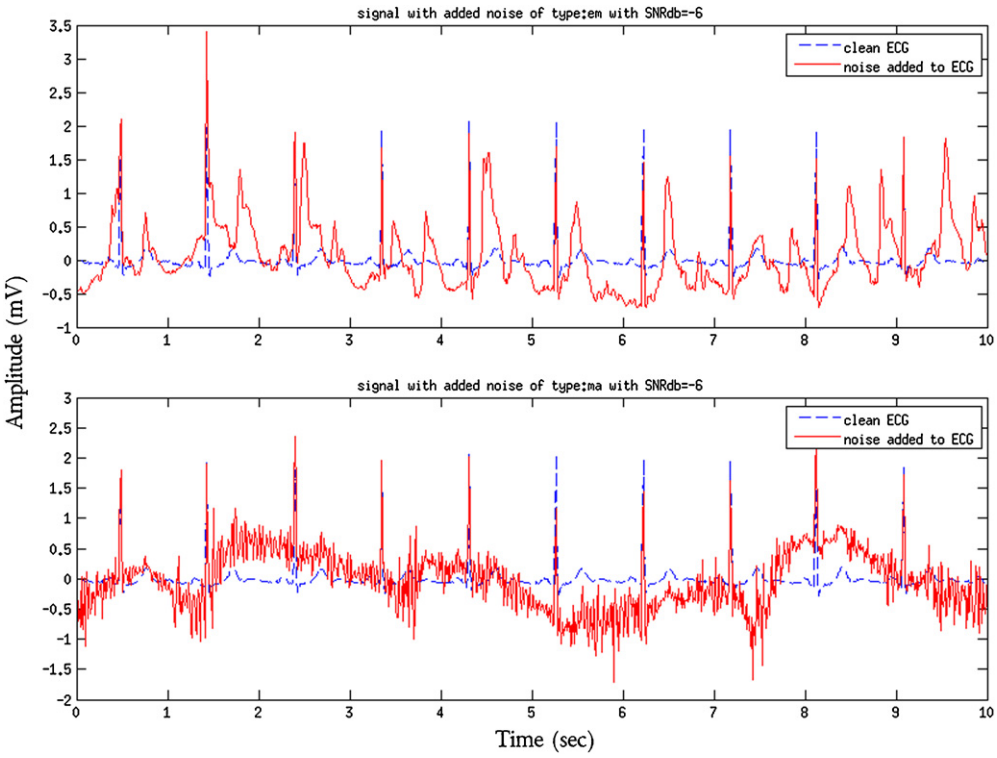

Figure 2. Examples of noisy ECG generation. The original clean ECG lead is displayed (dashed line) with the resultant ECG after addition of noise with SNR $= -6$db noise (solid line). Upper plot illustrates the addition of *em* noise and the lower plot the addition of *ma* noise. Note that these are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

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‡ hereafter.

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 *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 *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*.
(iii) pSQI: the relative power in the QRS complex: $\int_{5\text{Hz}}^{15\text{Hz}} P(f)\,\mathrm{d}f / \int_{5\text{Hz}}^{40\text{Hz}} P(f)\,\mathrm{d}f$.
(iv) sSQI: the third moment (skewness) of the distribution.

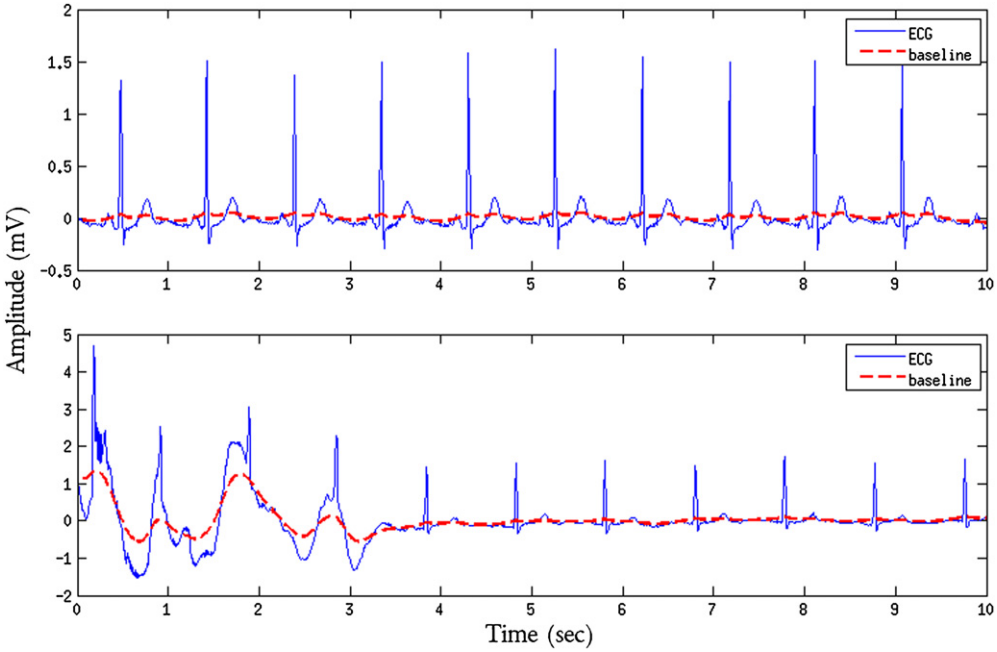

Figure 3. Example of baseline for a good-quality ECG sample (upper plot) and bad-quality ECG sample (lower plot) taken from Set-a of the PhysioNet/Computing in Cardiology Challenge 2011 dataset (Silva *et al* 2011). The baseline plot corresponds to the signal filtered with a cutoff frequency at 1Hz using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

(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: $1 - \int_{0\text{Hz}}^{1\text{Hz}} P(f)\,\mathrm{d}f / \int_{0\text{Hz}}^{40\text{Hz}} P(f)\,\mathrm{d}f$.

Note that a low *basSQI* means that the power within the band [0; 1 Hz] is abnormally high with respect to the power in the [0; 40 Hz] 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.

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	*Ac*	0.910	0.780	0.916	0.756	0.681	0.599	0.932
Set-a‡	*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	*Ac*	0.899	0.765	0.917	0.741	0.689	0.592	0.919
Set-b‡	*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 the 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

and consequently model parameters selection must be performed. As recommended by Hsu *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 was the most relevant and also if a combination of the seven SQI provided better results than all SQIs together. We found that the best result was obtained when considering four SQIs (bSQI, basSQI, kSQI and 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

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	*Ac*	0.920	0.813	0.911	0.918	0.704	0.713	0.922
Set-a‡	*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	*Ac*	0.909	0.805	0.925	0.906	0.723	0.730	0.933
Set-b‡	*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 the SVM with several combinations of SQIs, 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		

Table 6 shows the results of classification on 12 leads using the 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 the SVM and hidden nodes selection on the 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 the SVM and training on Set-b‡.

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 10 to 5 s in second intervals. We noted that the MLP dropped its performance from 97% to 93% on the test data (Set-b‡).