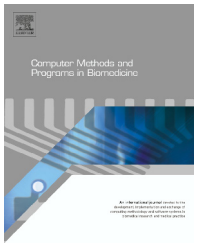
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 117 (2014) 435–447

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



A machine learning approach to multi-level ECG signal quality classification

Qiao Li[a,b], Cadathur Rajagopalan[c], Gari D. Clifford[b,*]

[a] Institute of Biomedical Engineering, School of Medicine, Shandong University, Jinan, Shandong 250012, China
[b] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK
[c] Mindray DS USA, Mahwah, NJ, USA

ARTICLE INFO

Article history:
Received 6 August 2013
Received in revised form
6 September 2014
Accepted 9 September 2014

Keywords:
ECG
Signal quality
Multi-level classification
Machine learning
Support vector machine

ABSTRACT

Current electrocardiogram (ECG) signal quality assessment studies have aimed to provide a two-level classification: clean or noisy. However, clinical usage demands more specific noise level classification for varying applications. This work outlines a five-level ECG signal quality classification algorithm. A total of 13 signal quality metrics were derived from segments of ECG waveforms, which were labeled by experts. A support vector machine (SVM) was trained to perform the classification and tested on a simulated dataset and was validated using data from the MIT-BIH arrhythmia database (MITDB). The simulated training and test datasets were created by selecting clean segments of the ECG in the 2011 PhysioNet/Computing in Cardiology Challenge database, and adding three types of real ECG noise at different signal-to-noise ratio (SNR) levels from the MIT-BIH Noise Stress Test Database (NSTDB). The MITDB was re-annotated for five levels of signal quality. Different combinations of the 13 metrics were trained and tested on the simulated datasets and the best combination that produced the highest classification accuracy was selected and validated on the MITDB. Performance was assessed using classification accuracy (Ac), and a single class overlap accuracy (OAc), which assumes that an individual type classified into an adjacent class is acceptable. An Ac of 80.26% and an OAc of 98.60% on the test set were obtained by selecting 10 metrics while 57.26% (Ac) and 94.23% (OAc) were the numbers for the unseen MITDB validation data without retraining. By performing the fivefold cross validation, an Ac of $88.07 \pm 0.32\%$ and OAc of $99.34 \pm 0.07\%$ were gained on the validation fold of MITDB.

1. Introduction

Electrocardiogram (ECG) recordings are often corrupted by large amounts of noise and artifacts that can be within the frequency band of useful cardiac data and can manifest with similar morphologies to the ECG waveform itself [1]. Not only does the presence of noise and artifact interfere with the correct recognition of QRS, P and T waves of the ECG, but also increases the rate of false alarms for cardiac monitors [2].

Some earlier published works have described ECG signal quality assessment and noise estimation methods, generally as part of a general approach to ECG analysis. Moody and Mark used the residual after projecting a QRS complex onto the first

* *Corresponding author at*: Old Road Campus Research Building, Off Roosevelt Drive, Headington, Oxford OX3 7DQ, UK. Tel.: +44 01865 273000; fax: +44 01865 617701.
E-mail address: gari@robots.ox.ac.uk (G.D. Clifford).

http://dx.doi.org/10.1016/j.cmpb.2014.09.002
0169-2607/

five principal components (PCs) of an ensemble of QRS complexes, or the Karhunen–Loève transform [3]. Standard noise measurement methods for ECG, which can be used as individual signal quality indices (SQIs), were reviewed by Clifford et al., including the root mean square (RMS) power in the isoelectric region, the 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 [1]. Li et al. developed four SQIs 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, and the kurtosis of the ECG [4]. These quality measures were calibrated to provide a mapping between a normalized SQI and an expected error in the heart rate (HR). The normalized SQI was then used as a weighting factor in the procedure of HR estimation by fusing the HR traces from the ECG and arterial blood pressure with a Kalman filter. This SQI-weighted HR estimation method provides an accurate HR estimate even in the presence of high levels of persistent noise and artifact, and during episodes of extreme bradycardia and tachycardia. Redmond et al. used signal masking methods to determine artifacts and the degree of missing information in the ECG [5]. Only 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 fifth-order high-pass elliptic forward–backward filter with a cut-off of 40 Hz (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 segments in the ECG signal. The algorithm yielded a sensitivity of 89% and a specificity of 98% although the fixed thresholds required manual tuning for different signal sources and it was not tested on an independent test set. Moreover, the analysis was not performed on a standard or public database and so the results cannot be compared to any other published method, nor can the results be considered to apply to any data outside their application. Their selection of noise sources was also not comprehensive.

The 2011 PhysioNet/Computing in Cardiology (PCinC) Challenge was to develop an efficient algorithm able to run on a mobile phone and provide useful feedback to indicate if an ECG is of adequate quality for interpretation [6,7]. The Challenge data were standard 12-lead ECG recordings 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 10 s. Every ECG was 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 channel, assigning it to 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 three groups [6,8]:

- Group 1 (acceptable): if $\hat{s} \geq 0.70$, and NF ≤ 1.
- Group 2 (indeterminate): if $\hat{s} \geq 0.70$, and NF ≥ 2.
- Group 3 (unacceptable): if $\hat{s} < 0.70$.

(NF 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. A total of 1500 12-lead ECGs were divided into two sets, in a ratio of 2:1. The larger set of 1000 ECGs was used for training (Set-A), for which binary annotations (acceptable or unacceptable) were available. A smaller set of 500 ECGs was available for testing (Set-B), although competition entrants were blinded to the annotations. Several algorithms from the Challenge were then published in a special focus issue of Physiological Measurement ("Signal Quality in Cardiorespiratory Monitoring") [8–15]. Hayn et al. explored the use of four criteria; a no signal detector, a spike detector, a lead crossing point analysis and a measure of the robustness of QRS detection. An accuracy of 93.3% was achieved on the training set and 91.6% on the test set [9]. Xia et al. developed twelve signal quality heuristics and calculated these for each of the 12 ECG leads, yielding a 12 by 12 matrix. The elements were then summed and a threshold applied to provide a classification for a given 12 lead ECG. After optimization of the threshold, the authors achieved an accuracy of 95%, with a sensitivity of 88% and specificity of 97% on the training set, but the authors did not report the performance on an independent test set [10]. Clifford et al. used a series of SQIs (based on morphological, statistical and spectral characteristics) and a support vector machine (SVM) or multilayer perceptron neural network to combine the SQIs and provide a quality estimate [8]. A classification accuracy of 98% on the training set and 97% on the test set was achieved. Only Clifford et al. studied the effect of the SQIs on arrhythmic data, and showed that arrhythmias led to a drop in accuracy to 93% indicating that algorithms may need retraining for some rhythms [8]. Subsequent work by the same research group showed that classification accuracies of up to 99% on training and test sets were obtained for normal sinus rhythm and up to 95% for arrhythmias on training and test sets [16].

These studies provided a two-class mapping for ECG signal quality: acceptable or unacceptable (as shown in PCinC website, the score is the fraction of group 1 and group 3 ECGs in set-B that were correctly classified by the participant's algorithm). However, clinical usage may demand much more specific signal quality levels according to different requirements [15]. Vaglio et al. studied different noise levels to help the cardiologist to define the set of ECGs that need a manual review after the automatic ECG measurements such as QT, PR and QRS intervals [17]. Vaglio points out that three signal quality bins will be satisfactory: low, average and good quality. ECGs flagged as low-quality will need manual reading, ECGs flagged as average-quality will need cardiologist over-read and good quality ECGs won't need any further review process. Redmond et al. classified the ECG quality into three categories based on the determinability of heart rate: good (HR is easy to determine); average (HR is difficult, but possible to determine); bad (HR cannot reliably be determined) [15]. To analyze low amplitude characteristic segments such as the P wave or ST segment, more levels of signal quality were required. Quesnel et al. used five signal-to-noise ratio (SNR) levels to gate alarms for myocardial ischemia by indicating ST deviations in ambulatory ECG [18]. Annotation of noise is also marked in some public domain ECG databases, such as MIT-BIH

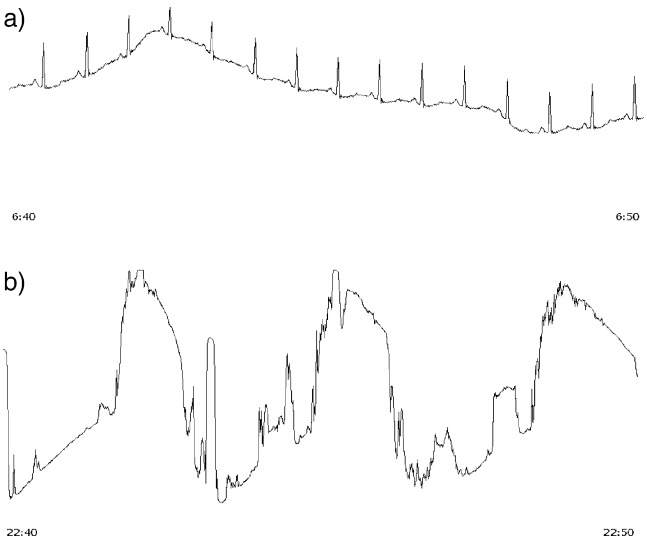


Fig. 1 – Two ECGs marked as noisy in MITDB. (a) Record 234, channel 1, from 6:40 to 6:50 (annotated as Level 2). (b) Record 103, channel 2, from 22:40 to 22:50 (annotated as Level 4).

arrhythmia database (MITDB) [19,20], but no further sub-level of noise was made. Fig. 1 shows two ECGs marked as noisy in MITDB where the noise levels and types of the two ECGs are quite different.

In this paper, we present a five-level ECG signal quality classification algorithm using a machine learning approach. A total of 13 signal quality metrics were derived from segments of ECG waveforms and were presented to a support vector machine to perform training on a simulated dataset and validated on the MIT-BIH arrhythmia database.

2. Methods

2.1. *Signal annotation*

2.1.1. *Signal quality categories*

By consulting our clinical partners and improving on the quality rating from the PCinC challenge, we define the single lead ECG signal quality categories as follow.

- Level 0 (clean): An outstanding recording with no visible noise or artifact
- Level 1 (minor noise): A good recording with transient artifact or low-level noise that does not interfere with interpretation or recognition of P, T or atrial flutter waves
- Level 2 (moderate noise): An adequate recording that can be interpreted with confidence despite visible and obvious flaws, but does not interfere with recognition of QRS complexes or ventricular flutter waves
- Level 3 (severe noise): A poor recording that may be interpretable with difficulty. Noise interferes with QRS or ventricular flutter recognition
- Level 4 (extreme noise): An unacceptably poor recording that cannot be interpreted with confidence because of significant technical flaws.

Table 1 – Required SNR to create noisy ECGs for each type of noise in the NSTDB. BW, baseline wander; EM, electrode motion; MA, muscle artifact.

Quality level	Description of noise	SNR levels (dB)		
		BW	EM	MA
Level 1	Minor	12	6	12
Level 2	Moderate	6	0	6
Level 3	Severe	0	−6	0
Level 4	Extreme	−6	−12	−6

2.1.2. *Annotation*

Two biomedical engineering master's degree students with moderate training in ECG analysis annotated each ECG lead separately and an ECG expert reviewed each of these annotations. When there was disagreement between the two students and the expert agreed with one of them, the majority decision ruled. If the expert disagreed with both of the annotators, (even though the two students agreed), a discussion between all annotators was conducted to reach a consensus.

2.2. *Data set*

2.2.1. *Simulated dataset*

The PCinC Challenge dataset was re-annotated by our group. The simulated dataset was composed by selecting the clean ECG leads (annotated as Level 0) from the PCinC Challenge dataset and adding different types of noise at varying SNR levels from the MIT-BIH Noise Stress Test Database (NSTDB) [20,21]. There were 1895 and 763 single leads of ECG marked as Level 0 in the PCinC Challenge training set (Set-A) and test set (Set-B) respectively, and these were defined to be Level 0 of the simulated dataset. Each lead was 10 s in duration. Three types of real ECG noise namely BW (baseline wander), EM (electrode motion artifact) and MA (muscle artifact), were added to the clean leads with different levels of SNR as shown in Table 1. Therefore a total of 5685 (1895×3) training segments and 2289 (763×3) test segments were created for each noise level. An illustration of signal quality at each level of the simulated dataset is shown in Fig. 2. A segment of 5 s ECG signal in the middle of each lead was used in this study, and so the first and last 2.5 s were discarded. Note that the dominant rhythm of the simulated dataset is of normal sinus rhythm.

2.2.2. *Real dataset*

The MITDB was re-annotated to include the five levels of signal quality by our group and was used as the real validation dataset. Every lead of the 48 half-hour ECG recordings in the MIT database was annotated using WAVE [22], which is an extensible interactive graphical environment for manipulating sets of digitized signals with optional annotations provided by Physionet. The annotator marked the beginning and the ending time of each different signal quality level by click the mouse buttons at the corresponding position of the ECG lead. Then a second-by-second annotation file was created automatically based on these fiducial markers. A five second non-overlapping window was used to segment the ECG data into different quality levels by a program designed by our group that reviewed the annotation file automatically. Note the 5 s segments were annotated according to

the annotation of the consistent continuous seconds. The segments which maintained the same quality level for at least 5 s were selected. Since the more noisy types of ECG tend to be less frequent or last a shorter amount of time in the MITDB, a modified selection approach was taken. Segments with four continuous seconds at a consistent quality level and a fifth second where the level was lower than that of the other 4 s was selected. Additionally, segments with 3 s at a consistent quality level and the first and the fifth seconds of lower quality than that of the central 3 s were selected for use. Since overlapping windows will cause the longer segment (which happens to be clean) to repeat many more times than the shorter segment (which happens to be noisy), we used non-overlapping windows in this work. A summary of the number of ECG segments of each quality level in the real dataset is to be found in Table 2. An illustration of each quality level of the real dataset is to be found in Fig. 3.

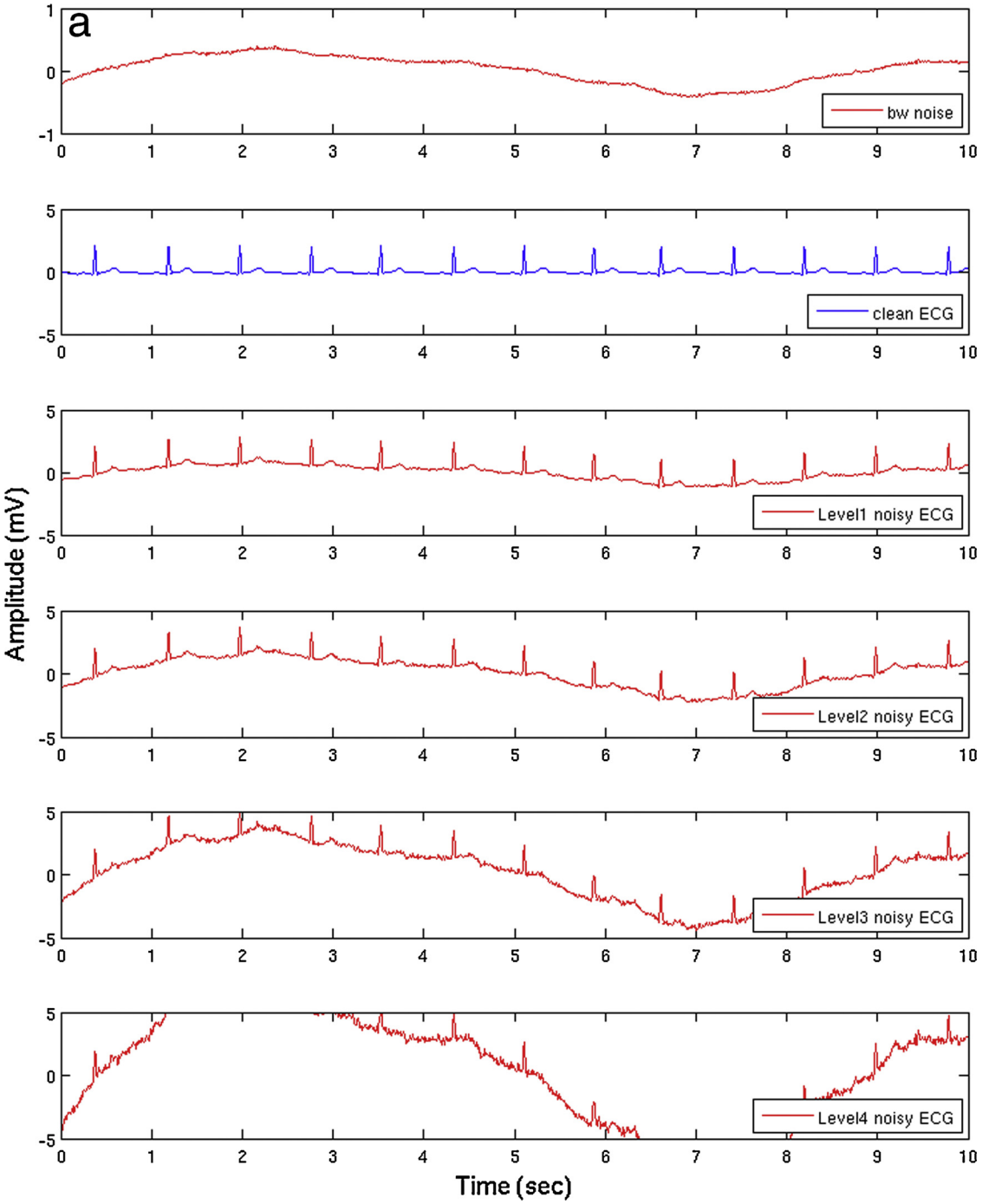


Fig. 2 – An illustration of signals corresponding to each quality level of the simulated dataset for (a) baseline wander (BW), (b) electrode motion (EM) and (c) muscle artifact (MA) noise. Note that for clarity the vertical axis was fixed at ±5 mV and values outside this range were clipped.

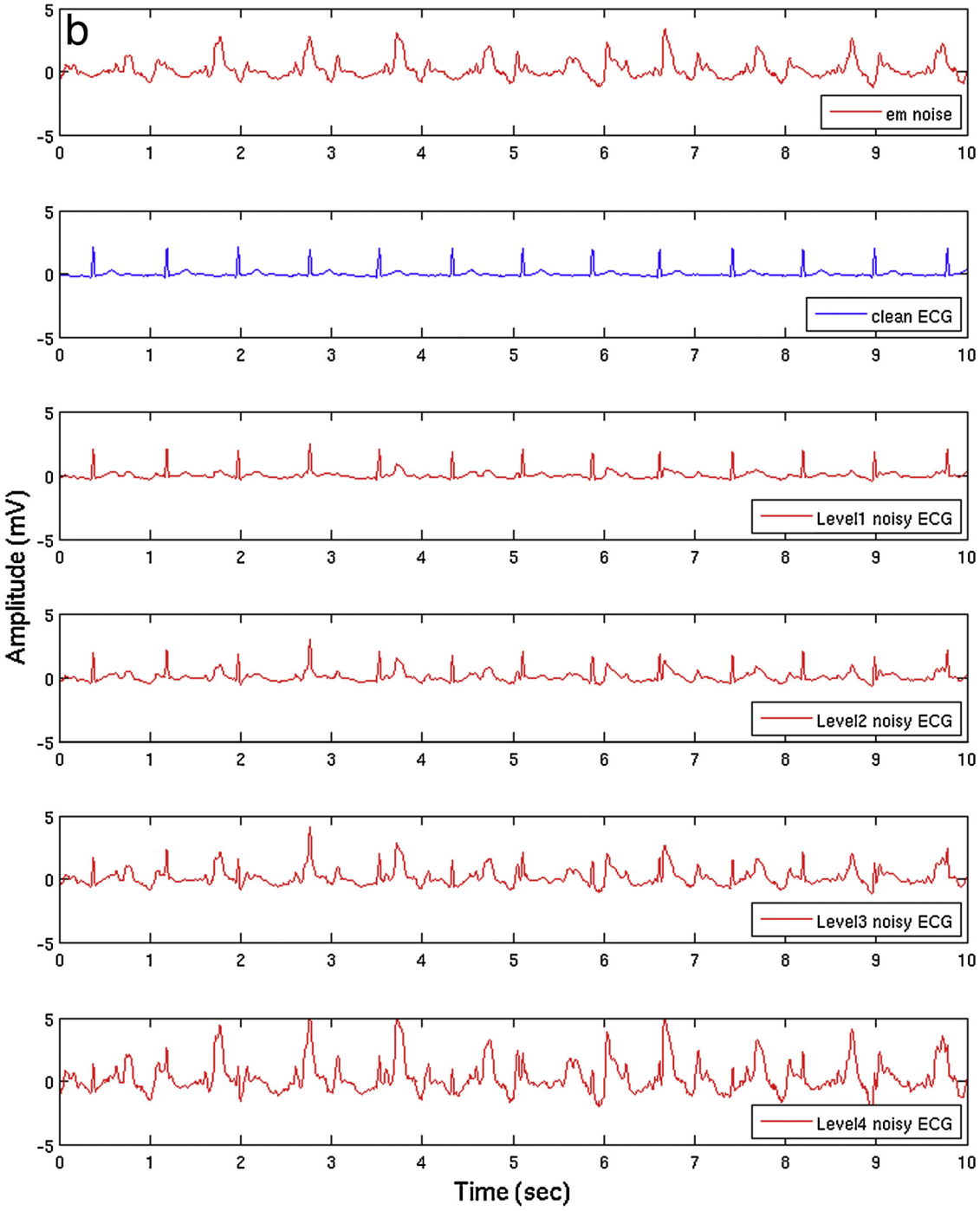


Fig. 2 – (*Continued*)

Table 2 – The number of ECG segments corresponding to each quality level of the real dataset after annotation.

Quality level	Real dataset
Level 0	22,441
Level 1	9298
Level 2	1450
Level 3	673
Level 4	117

2.3. *Pre-processing of ECGs*

Each channel of ECG was down-sampled to 125 Hz using an anti-aliasing filter which is a linear-phase lowpass FIR filter designed using the Parks-McClellan algorithm [23], with a 35 Hz passband and 45 Hz stopband cutoff frequency, and at least 48.25 dB attenuation in the stopband and less than 1.925 dB of ripple in the passband. QRS detection was performed on each channel individually using two open source QRS detectors (*eplimited* and *wqrs*) since *eplimited* is

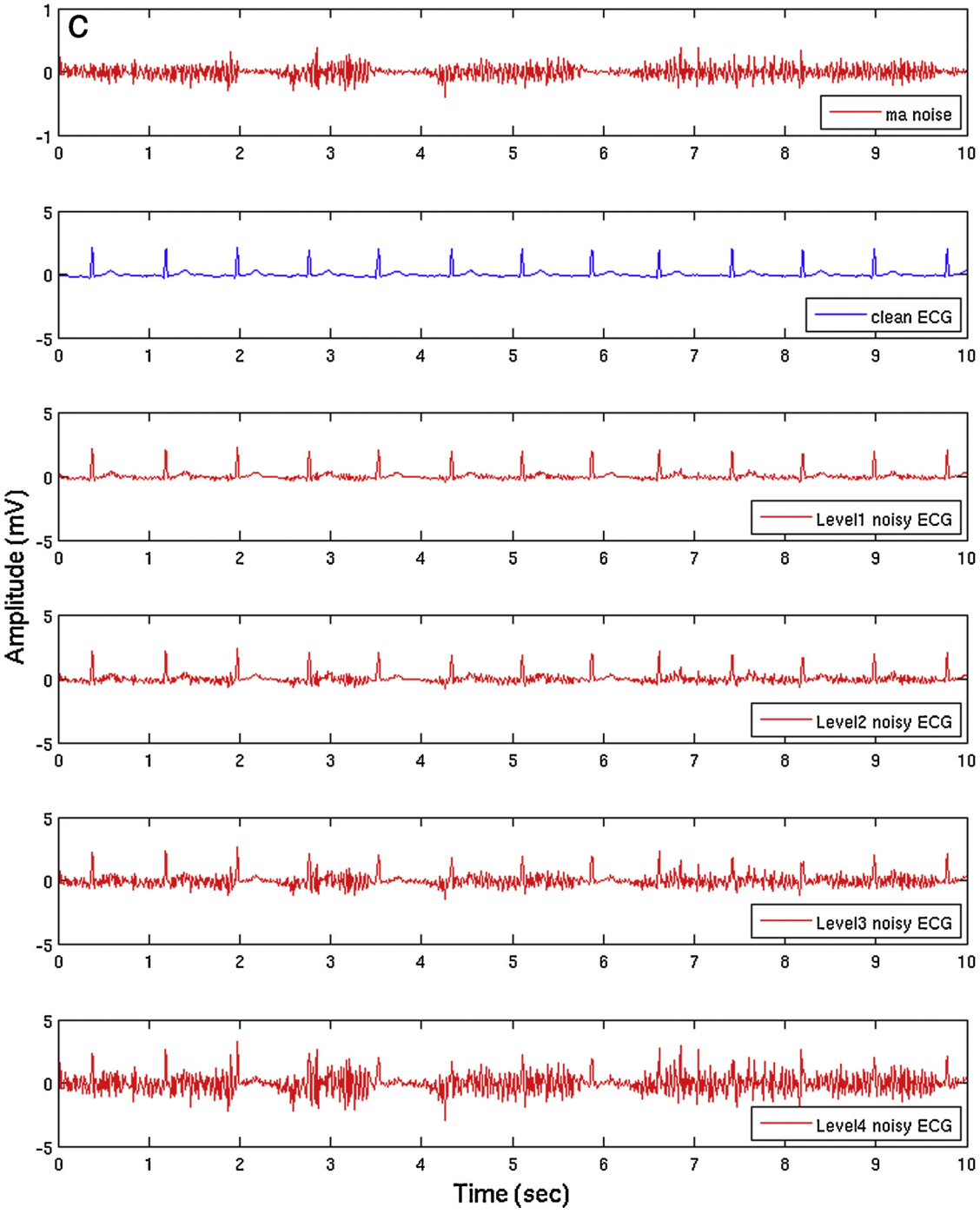


Fig. 2 – (*Continued*).

less sensitive to noise [4]. Note that *eplimited* is a QRS detector based on the Hamilton and Tompkins algorithm [24], whereas *wqrs* is based upon a length transform [25].

2.4. *ECG Signal quality indices*

Thirteen signal quality indices were extracted from the ECG waveform by expanding on earlier works and published papers.

1. *bSQI*: the percentage of beats detected by *wqrs* that were also detected by *eplimited* [2].
2. *sSQI*: the third moment (skewness) of the ECG signal [8].
3. *kSQI*: the fourth moment (kurtosis) of the ECG signal [2].
4. *pSQI*: the relative power in the QRS complex [2].

$$pSQI = \frac{\int_5^{15} P(f)df}{\int_5^{40} P(f)df} \quad (1)$$

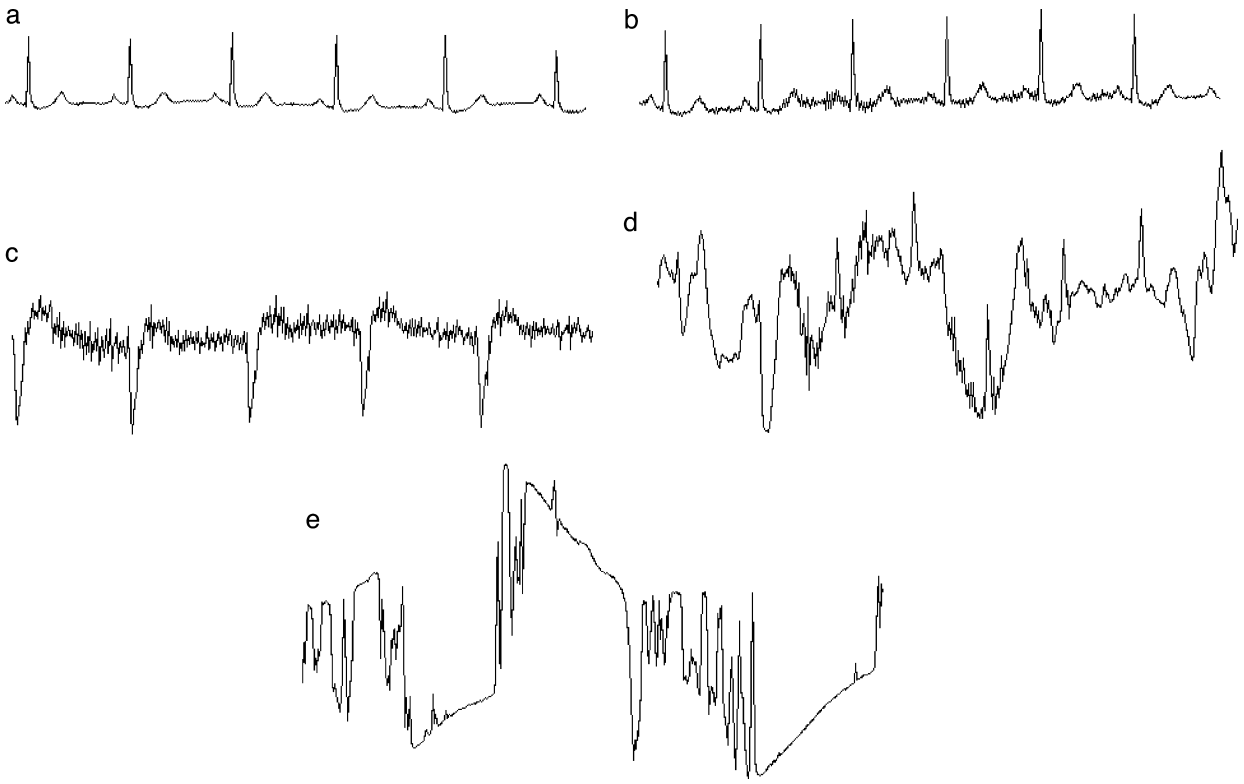


Fig. 3 – An illustration of signal quality at each level of the real dataset. (a) Level 0 (Record 101, channel 1, from 0:01 to 0:05). (b) Level 1 (Record 101, channel 1, from 0:31 to 0:35). (c) Level 2 (Record 108, channel 2, from 26:14 to 26:18). (d) Level 3 (Record 105, channel 1, from 28:13 to 28:17). (e) Level 4 (Record 103, channel 2, from 22:21 to 22:25).

5. *basSQI*: the relative power in the baseline [8].

$$basSQI = 1 - \frac{\int_0^1 P(f)df}{\int_0^{40} P(f)df} \tag{2}$$

6. *bsSQI*: baseline wander check in time domain.

$$bsSQI = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Ra_i}{Ba_i}\right) \tag{3}$$

where Ra_i is the peak-to-peak amplitude of the ECG waveform around each QRS complex (from $R-0.07$s to $R+0.08$s), R is the fiducial marker of QRS complex of *eplimited*, $i=1, 2, 3, \ldots, N$ is the detected QRS complex in the analysis window.
Ba_i is the peak-to-peak amplitude of baseline (filtered by a 1 Hz lowpass filter, $H(z)=0.0503/(1-0.9497z^{-1})$) around each QRS complex(from $R-1$s to $R+1$s).

7. *eSQI*: the relative energy in the QRS complex.

$$eSQI = \frac{\sum_i Er_i}{Ea} \tag{4}$$

where $Er_i = \sum x^2$ is the energy of each QRS segment (from $R-0.07$s to $R+0.08$s), $i=1, 2, 3, \ldots$ is the detected QRS complex in the analysis window, Ea is the energy of the analysis window.

8. *hfSQI*: the relative amplitude of high frequency noise.

$$hfSQI = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Ra_i}{H_i}\right) \tag{5}$$

where the ECG signal (x) was multiplied by integer coefficients high pass filter, with the difference equation $y(j)=x(j)-2x(j-1)+x(j-2)$. Then the filtered signal (y) was summed for every six points $s(j)=|y(j)|+|y(j-1)|+\cdots+|y(j-5)|$. Ra_i is the peak-to-peak amplitude of each QRS complex. H_i is the mean of $s(j)$ before each QRS complex (from $R-0.28$s to $R-0.05$s).

9. *purSQI*: signal purity of ECG [26].

$$purSQI = \frac{(\overline{\omega_2}(k))^2}{\overline{\omega_0}(k)\overline{\omega_4}(k))} \tag{6}$$

where $\bar{\omega}_n = \int_{-\pi}^{\pi} \omega^n P(e^{j\omega})d\omega$
$P(e^{j\omega})$ is the power spectrum of the ECG in the analysis window, and $\omega = 2\pi f$.

10. *rsdSQI*: the relative standard deviation (STD) of QRS complex.

$$hfSQI = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma r_i}{\sigma a_i * 2} \tag{7}$$

where σr_i is the STD of each QRS (from $R-0.07$s to $R+0.08$s), σa_i is the STD around each QRS (from $R-0.2$s to $R+0.2$s).

11. *entSQI*: the sample entropy of the ECG waveform [27].

$$entSQI = En(m, r, N) = -\ln\left[\frac{A^m(r)}{B^m(r)}\right] \quad (8)$$

where N is the length of the ECG waveform, m is the preset length of repeat templates, r is a preset tolerance. For a time series of N points, $\{u(j): 1 \le j \le N\}$ forms the $N-m+1$ vectors $x_m(i)$ for $\{i \mid 1 \le i \le N-m+1\}$, where $x_m(i) = \{u(i+k): 0 \le k \le m-1\}$ is the vector of m data points from $u(i)$ to $u(i+m-1)$. $B^m(r) = (N-m)^{-1}\sum_{i=1}^{N-m} B_i^m(r)$ and $A^m(r) = (N-m)^{-1}\sum_{i=1}^{N-m} A_i^m(r)$. We defined $B_i^m(r)$ as $(N-m-1)^{-1}$ times the number of vectors $x_m(j)$ within r of $x_m(i)$, where j ranges from 1 to $N-m (j \ne i)$. Similarly, we define $A_i^m(r)$ as $(N-m-1)^{-1}$ times the number of vectors $x_{m+1}(j)$ within r of $x_{m+1}(i)$, where j ranges from 1 to N-m $(j \ne i)$.

12. *hfMSQI*: high frequency mask of ECG waveform [5].
The ECG signal was first notch filtered at 50/60 Hz to remove line noise. The signal was then high-pass filtered using a 5th order elliptic filter with a cut-off of 40 Hz, −80 dB gain in the stopband and 0.5 dB ripple in the passband. Note that in the pre-processing step the low pass filter has cut out the frequencies that are greater than 45 Hz, the high-pass filter here keeps the information within the high frequency band 40–45 Hz. However, it can still be used for high frequency muscle artifact and electrode-tissue contact noise detection. The high-pass signal was squared and then lowpass filtered using a 0.05 s normalized Hamming window FIR filter. Then the instantaneous power estimate (p) was square-rooted and used as a high frequency mask to compare with a fixed threshold. However, a ratio of p between the sum of QRS range (p_i) and the sum of full analysis window (p_a) is more comparable than the instantaneous value (p) used by Redmond in [5].

$$hfMSQI = \frac{\sum_i p_i}{p_a} \quad (9)$$

where p_i ranges from $R-0.07$s to $R+0.08$s.

13. *PiCASQI*: periodic component analysis (PiCA) periodicity measure of the ECG waveform [28].
If we define the covariance of the signal $x(t)$ as:

$$C_x(\tau) = E_t\{x(t+\tau)x(t)\} \quad (10)$$

where $E_t\{\bullet\}$ indicates averaging over t,τ is a constant time-lag. In order to apply this to the ECG signal, we replace τ with a variable τ_t that is calculated from beat-to-beat ECG. Therefore, in each ECG cycle, the sample at time-instant t is compared with the sample $t+\tau_t$, which is the sample with the same phase value in the succeeding ECG beat. Then,

$$PiCASQI = \left|\frac{C_x(\tau_t)}{C_x(0)}\right| = \left|\frac{E_t\{x(t+\tau_t)x(t)\}}{E_t\{x(t)^2\}}\right| \quad (11)$$

Table 3 – Distribution of N subjects by rater and response category (adapted from [31]).

Rater A	Rater B				Total
	1	2	...	q	
1	n_{11}	n_{12}	...	n_{1q}	n_{1+}
2	n_{21}	n_{22}	...	n_{2q}	n_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
q	n_{q1}	n_{q2}	...	n_{qq}	n_{q+}
Total	n_{+1}	n_{+2}	...	n_{+q}	N

2.5. Machine learning for classifying ECG signal quality

To classify the signal quality of the ECG, we used a SVM classifier (Lib-SVM library) [29,30] with a Gaussian radial basis function (RBF) kernel defined by: $K(x_n, x_m) = \exp(\gamma\|x_n - x_m\|^2)$, where γ controls the width of the Gaussian and plays a role in controlling the flexibility of the resulting classifier. x_n and x_m are two vectors expressed in the initial feature space. Lib-SVM decouples the multiclass classification problem to several two-class problems and a voting strategy is used: each binary classification is considered to be a voting where votes can be cast for all data points. In the end, a point is designated to be in a class with the maximum number of votes.

Every possible combination of the SQIs was fed into the SVM model as input to train the model using the training set from the simulated dataset with 5s window lengths beginning from the individual SQI, then pairs, triplets and so on. The models were then evaluated on the test set of the simulated dataset. The SQI combination which provided the best performance was selected and used to classify the real dataset without retraining. Cross validation was also performed on the simulated dataset and real dataset separately.

2.6. Performance evaluation

Classification accuracy (Ac) was used to evaluate the performance of the algorithm. Table 3 presents the confusion matrix, where rater A stands for the algorithm output and rater B stands for the annotation label. The Ac is defined as follows:

$$Ac = \sum_{k=1}^{q} \frac{n_{kk}}{N} \quad (12)$$

where q is the number of categories, N is the total number of segments.

However, as neighboring classes may have substantial overlap, we also used a metric which assumes that an individual type classified into neighbor types is acceptable (e.g. level 2 is classified as level 1 or level 3). This 'acceptable single class overlap accuracy' (OAc) was also used to denote this process and further evaluate performance.

$$OAc = \sum_{k=1}^{q} \frac{n_{kk} + n_{(k+1)k} + n_{(k-1)k}}{N} \quad (13)$$

Note there are no $n_{(q+1)q}$ and n_{01} in (13).

Furthermore, we report the Cohen's Kappa and Gwet's AC1 [31] for performance evaluation.

$$Kappa = \frac{p_a - p_e}{1 - p_e}, \quad \text{where} \quad p_a = \sum_{k=1}^{q} p_{kk}, p_e = \sum_{k=1}^{q} pk + p + k \quad (14)$$

$$ACI = \frac{p_a - p_e}{1 - p_e}, \quad \text{where} \quad p_a = \frac{1}{1 - p_M} \sum_{k=1}^{q} p_{kk},$$

$$p_e = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k (1 - \pi_k) \quad (15)$$

where $p_{kl} = n_{kl}/N$ represents the percentage of subjects classified into category k by the algorithm and into category l by the annotated label; p_{k+} and p_{+k} represent the percentage of subjects assigned to category k by the algorithm and annotated label respectively; p_m is the relative number of subjects rated by a single rater and $\pi_k = (p_{k+} + p_{+k})/2$.

3. Results

3.1. Classification model

The performance of the individual SQIs was determined using the training and test datasets of the simulated data. Table 4 shows that *hfSQI* got the best acceptable overlap accuracy on training and test datasets of the simulated data. The results for the best SQIs combinations are summarized in Table 5. The best OAc of 98.84% on the training set and 98.60% on the test set was obtained by selecting 10 SQIs, namely the combinations of *hfSQI, kSQI, PiCASQI, sSQI, basSQI, bsSQI, entSQI, pSQI, purSQI* and *hfMSQI*. When selecting four or more SQIs, the OAc remained above 98% on the training and test sets. However, the Kappa and AC1 continued to increase from 65% to 75% along with the increasing number of SQIs, which showed the increase in agreement between the algorithm and the label along with the increasing number of selected SQIs. Table 6 provides the confusion matrix on the test dataset when the selected 10-SQI model was used.

To evaluate the performance of the selected 10-SQI model, we performed the fivefold cross validation on the simulated dataset. The training and test datasets were combined and randomly partitioned into five equal sized subsamples (folds). Of the five folds, a single fold was retained as the validation data for testing the model, and the remaining four folds were used as training data. The cross-validation process was then repeated five times, with each of the five folds used exactly once as the validation data. We then repeated the fivefold cross validation process ten times and averaged the results which are presented in Table 7. The average OAc on the validation fold was 98.79 ± 0.13%, which was almost the same as that on the training folds (98.79 ± 0.04), indicating the robustness of the model's performance.

Table 4 – Performance of each individual SQI on the simulated dataset.

SQI	Training dataset (%)				Test dataset (%)			
	Ac	OAc	Kappa	AC1	Ac	OAc	Kappa	AC1
bSQI	44.06	82.96	27.28	31.43	43.02	80.76	25.92	30.14
sSQI	39.87	75.92	21.86	26.13	39.62	76.51	21.84	26.05
kSQI	39.68	77.47	22.10	25.92	39.65	77.86	21.79	25.59
pSQI	37.74	71.71	19.06	23.31	37.45	70.45	18.69	22.92
basSQI	33.41	67.79	13.44	19.88	34.07	68.12	14.29	20.74
bsSQI	35.00	77.18	15.85	19.85	35.56	78.14	16.53	20.49
eSQI	34.15	80.00	14.55	19.74	34.24	80.23	14.62	19.94
hfSQI	43.67	**89.94**	27.60	30.48	**45.71**	**90.06**	**30.36**	**32.96**
purSQI	37.85	76.29	19.20	23.42	37.74	76.66	19.06	23.28
rsdSQI	**47.67**	89.62	**31.97**	**35.40**	45.55	88.06	29.21	32.80
entSQI	37.66	65.18	20.53	23.58	35.96	63.46	18.45	21.44
hjMSQI	30.43	63.62	10.25	15.15	30.43	63.66	10.20	15.17
PiCASQI	41.35	83.00	23.88	27.67	42.55	84.45	25.42	29.10

Table 5 – Performance for combinations of varying numbers of SQIs on the simulated dataset.

# of features	Combinations of SQIs	Training dataset (%)				Test dataset (%)			
		Ac	OAc	Kappa	AC1	Ac	OAc	Kappa	AC1
1	*hfSQI*	43.67	89.94	27.60	30.48	45.71	90.06	30.36	32.96
2	*hfSQI,purSQI*	57.18	95.97	44.93	46.87	58.54	95.75	46.79	48.55
3	*hfSQI,kSQI,PiCASQI*	68.20	97.89	59.19	60.52	67.53	97.69	58.38	59.67
4	*hfSQI,kSQI,PiCASQI,sSQI*	72.66	98.38	64.92	66.04	72.59	98.41	64.86	65.95
5	*hfSQI,kSQI,PiCASQI,sSQI,basSQI*	74.82	98.58	66.76	67.81	73.20	98.45	65.67.	66.71
6	*hfSQI,kSQI,PiCASQI,sSQI,basSQI,bsSQI*	77.03	98.62	70.53	71.47	76.93	98.59	70.43	71.35
7	*hfSQI,kSQI,PiCASQI,sSQI,basSQI,purSQI,entSQI*	78.17	98.73	72.04	72.88	77.91	98.45	71.72	72.55
8	*hfSQI,kSQI,PiCASQI,sSQI,basSQI,bsSQI,entSQI,pSQI*	79.48	98.82	73.71	74.51	79.54	98.56	73.81	74.59
9	*hfSQI,kSQI,PiCASQI,sSQI,basSQI,bsSQI,entSQI,pSQI,purSQI*	79.99	98.82	74.36	75.14	80.06	98.59	74.46	75.22
10	*hfSQI,kSQI,PiCASQI,sSQI,basSQI,bsSQI,entSQI,pSQI,purSQI,hfMSQI*	79.93	**98.84**	74.28	75.06	80.26	**98.60**	74.72	75.47
11	*hfSQI,kSQI,PiCASQI,sSQI,basSQI,bsSQI,entSQI,pSQI,purSQI,hfMSQI,bSQI*	80.58	98.77	75.12	75.87	80.54	98.56	75.08	75.83
12	*hfSQI,kSQI,PiCASQI,sSQI,basSQI,bsSQI,entSQI,pSQI,purSQI,hfMSQI,bSQI,eSQI*	80.87	98.79	75.48	76.23	80.39	98.55	74.88	75.63
13	All	**80.93**	98.67	75.56	76.31	80.38	98.41	74.87	75.62

Table 6 – Classification confusion matrix for the 10-SQI model (see Table 5 on the simulated test dataset).

Model output	Annotation label				
	Level 0	Level 1	Level 2	Level 3	Level 4
Level 0	681	58	1	0	0
Level 1	66	1955	348	9	4
Level 2	12	220	1616	167	9
Level 3	1	28	280	1760	327
Level 4	3	28	44	353	1949

3.2. *Performance on real dataset*

The selected 10-SQI model was evaluated on the real dataset without retraining. The results showed an Ac of 57.26%, an OAc of 94.23%, a Kappa of 30.62% and an AC1 of 49.83% on the unseen real dataset. The confusion matrix for the real dataset is shown in Table 8 and the detailed results on each record of the real dataset are shown in Table 9. From Table 9, it can be seen that good OAc was achievable when the dominant rhythm was sinus rhythm. However, the accuracy was lower when arrhythmias were present, such as atrial flutter (AFL) and atrial fibrillation (AF) (record 201, 203 and 217), ventricular flutter (VFL) (record 207) and ventricular trigeminy (record 208 and 214). This reveals that the irregular arrhythmia waveforms, such as large AFL/AF waves, VFL waves and frequent ventricular beats, will impact correct signal quality classification. Since the model was trained on the simulated dataset which mostly consisted of normal sinus rhythm, the characteristics of arrhythmias were not included in the training of the model.

In order to test the performance when signals with arrhythmias were used for training, we performed the fivefold cross validation on the real dataset and repeated ten times. The average results are shown in Table 10. An Ac of $88.07 \pm 0.32\%$, OAc of $99.34 \pm 0.07\%$, Kappa of $74.65 \pm 0.67\%$ and AC1 of $86.48 \pm 0.37\%$ on the validation fold were obtained, which are substantially better than the results using the model trained on the simulated dataset (which included no arrhythmia data).

4. Discussion and conclusion

In general, different applications and situations require different levels of ECG quality [14]. The two-class mapping for ECG signal quality (acceptable or unacceptable) was the most common in the previous studies because of its simplicity [5,8–14]. But the signal quality scores to which these studies were relevant were not informative for interpretation. Annotators were only asked to indicate how easy it was to interpret the signal in general. The difficulty of interpretation is highly dependent on what information one intends to extract from the signal [15]. As Redmond pointed out in [15], determining heart rate was much easier than identifying waveform morphology changes, and the signal quality levels required to reliably perform each task would vary significantly.

One solution of this issue is to use a continuous signal quality score. In our previous studies [2,4], continuous signal quality metrics were extracted for the usages of false alarm reduction and robust heart rate estimation and thresholds were selected based on specific applications. Although a

Table 7 – Results of fivefold cross validation using the 10-SQI model on the simulated dataset.

Simulated dataset	Ac (%)	OAc (%)	Kappa (%)	AC1 (%)
Training (on 4 folds)	80.34 ± 0.12	98.79 ± 0.04	74.81 ± 0.16	75.57 ± 0.15
Validation (on 1 retained fold)	80.10 ± 0.54	98.79 ± 0.13	74.51 ± 0.68	75.28 ± 0.67

Table 8 – Classification confusion matrix using the 10-SQI model on the real dataset.

Model output	Annotation label				
	Level 0	Level 1	Level 2	Level 3	Level 4
Level 0	12,761	860	0	2	4
Level 1	8273	5512	120	19	21
Level 2	1252	2493	815	190	9
Level 3	106	398	448	314	30
Level 4	49	35	67	148	53

continuous score allows ECG qualities be divided into different levels, in the end specific thresholds must be set to distinguish the usability of the ECG in a given context. As Johannesen states [14], "In general, the quality required to accept or reject a recording depends on how the recording will be used. We can envision several different situations where the thresholds for a recording to be 'acceptable' differ: Detecting T-wave alternans or QRS late potentials might require a higher signal quality than that needed for regular rhythm analysis…" In this study, although some of the signal metrics are continuous, these metrics have to be fused to provide an integrated score through an SVM-based mapping. The SVM algorithm can indeed provide a continuous output by training the model with a two-type classification training set, but the continuous value is hard to validate since it is difficult to create a test dataset with continuous signal quality annotations.

The alternative solution is to use more specific signal quality levels. Redmond et al. classified the ECG quality into three categories based on the determinability of heart rate [15]. Quesnel et al. used five SNR levels to gate alarms for myocardial ischemia by indicating ST deviations in ambulatory ECG [18]. A five-level quality index was also used as initial grades for annotation by volunteer annotators in the PCinC Challenge [6]. The annotators were asked to provide an overall assessment of each selected 12-lead ECG by assigning one of five possible letter grades to it: A (an outstanding recording with no visible noise or artifact; such an ECG may be difficult to interpret for intrinsic reasons, but not technical ones); B (a good recording with transient artifact or low level noise that does not interfere with interpretation; all leads recorded well); C (an adequate recording that can be interpreted with confidence despite visible and obvious flaws, but no missing signals); D (a poor recording that may be interpretable with difficulty, or an otherwise good recording with one or more missing signals); or F (an unacceptably poor recording that cannot be interpreted with confidence because of significant technical flaws). Each grade represented the observer's assessment of the entire ECG

Table 9 – Results per record in real dataset using the 10-SQI model.

Record	Dominant Rhythm	Ac (%)	OAc (%)	Record	Dominant Rhythm	Ac (%)	OAc (%)
100	N	47.92	99.58	201	N, AFIB, T	43.39	84.70
101	N	73.25	99.72	202	N, AFIB	56.98	97.49
102	P	71.05	100.00	203	AFL, AFIB	23.45	67.40
103	N	84.16	97.31	205	N	56.55	96.24
104	P	38.35	99.29	207	N, VFL, SVTA, B, IVR	26.14	84.00
105	N	63.27	90.52	208	N, T	36.57	61.14
106	N, B	36.98	82.45	209	N, SVTA	55.15	99.56
107	P	83.03	97.77	210	AFIB	59.38	95.77
108	N	52.89	93.79	212	N	51.37	98.27
109	N	96.65	99.02	213	N, B	44.65	93.88
111	N	80.06	98.60	214	N, T	62.27	77.45
112	N	45.48	99.86	215	N	65.07	96.38
113	N	67.33	98.59	217	P, AFIB	71.67	83.57
114	N	64.80	93.16	219	N, AFIB	67.37	100.00
115	N	87.71	99.44	220	N	50.84	100.00
116	N	87.39	99.16	221	AFIB, T	60.40	92.83
117	N	2.64	99.72	222	N, AFL, NOD, AB	41.24	90.40
118	N	53.10	99.44	223	N, B, VT, T	66.20	97.18
119	N, B, T	48.06	99.44	228	N, B	49.20	88.50
121	N	32.06	98.45	230	N, PREX	76.51	99.44
122	N	73.23	99.86	231	N, BII	59.86	99.72
123	N	50.63	99.86	232	SBR	67.51	99.58
124	N	82.96	99.72	233	N, B	44.17	87.91
200	N, B	46.84	84.77	234	N, SVTA	38.85	98.60

Note: N, Normal Sinus Rhythm; AB, Atrial Bigeminy; AFL, Atrial flutter; AFIB, Atrial Fibrillation; B, Ventricular Bigeminy; BII, 2nd Degree heart block; IVR, Idioventricular Rhythm; NOD, Nodal Rhythm; P, Paced Rhythm; PREX, Pre-excitation (WPW); SBR, Sinus Bradycardia; SVTA, Supra ventricular Tachyarrhythmia; T, Ventricular Trigeminy; VFL, Ventricular Flutter; VT, Ventricular Tachycardia.

Table 10 – Results of fivefold cross validation using the 10-SQI model on the real dataset.

Real dataset	Ac (%)	OAc (%)	Kappa (%)	AC1 (%)
Training (on 4 folds)	88.17 ± 0.08	99.34 ± 0.02	74.88 ± 0.18	86.60 ± 0.09
Validation (on 1 retained fold)	88.07 ± 0.32	99.34 ± 0.07	74.65 ± 0.67	86.48 ± 0.37

record (10 s and 12 channels), as an overall measure of quality. In this study, we graded the signal quality into five categories according to clinical usage, such as whether or not the noise interferes with interpretation or recognition of P, T waves and QRS complexes.

We created a simulated dataset for algorithm development by selecting clean segments of the ECG in the 2011 PCinC Challenge database, and adding three types of realistic ECG noise at different SNR levels from the NSTDB. Adding realistic ECG noise to clean ECG data in order to create noisy datasets has become a standard method since Moody introduced this method in 1984 [21] and has since been generally accepted by researchers [3,4,8,18,32]. The most commonly used noise dataset is the NSTDB Database. However, the additive noise model has the limitation that it differs to some extent with the real world ECG records where there can be non-linear distortion and missing channels. This difference partly explains why the performance (OAc) drops from 98.79% on the simulated dataset to 94.23% on the unseen real dataset. We reviewed the annotations in the noisy categories in MITDB and found that none of Level 1, 0.3% of Level 2, 3.2% of Level 3, and 45.8% of Level 4 are non-linear distortion and non-additive noise. That is, the percentage of non-linear and non-additive noise increases along with the increase in noise level.

In a previous study, we found that the signal quality classification algorithm experienced a reduction in performance on an arrhythmia database when it was not explicitly retrained using signals containing arrhythmia episodes [8,16]. The result presented here displays a similar performance drop, providing additional evidence that this is to be expected. The reason for the poor results on arrhythmic data is that the classification model was trained on the simulated dataset which mostly consisted of normal sinus rhythm. Moreover, the Challenge data were acquired from healthy subjects, and thus, they did not include a wide range of pathological cardiac conditions. Since the AFL or AF waves show characteristic flutter waves or disorganized electrical activity between two QRS complexes and the VFL waves show a sinusoidal waveform without clear definition of the QRS and T waves, the extracted signal quality metrics, such as kurtosis or the metrics involved in QRS detection, will differ from the metrics extracted from sinus rhythm. As the training set did not include the irregular arrhythmia waveforms, the model has no ability to learn how to classify these types of arrhythmia data into clean or noisy categories and then is likely to cause a classification error. In order to test the performance when arrhythmia data were used for training, we performed the fivefold cross validation on the real dataset. This resulted in an Ac of 88.07 ± 0.32% and OAc of 99.34 ± 0.07% on the validation folds, which are obviously better than the results using the model trained on the simulated dataset, wherein no arrhythmia data are present. However, as noted by Clifford et al. in [8] and Li et al. in [33] it may be more judicious to apply rhythm detectors first, and then select an appropriate SQI combination pertinent to the suspected rhythm, or develop a rhythm classifier which includes signal quality metrics.

In summary, in this study we have presented a five-level ECG signal quality classification algorithm using a machine learning approach. By validating on a simulated noisy dataset and real world data (the MIT-BIH arrhythmia database), the proposed algorithm was shown to provide a high classification accuracy. By classifying the signal quality of the ECG into specific levels, a common reference for different clinical applications may be achieved. However, we note that the subjective descriptions of quantized data quality levels have significant overlap and inter-observer variability is often high, which hampers the design of a perfect classifier.

Acknowledgment

The authors gratefully acknowledge funding for this research from Mindray DS USA.

REFERENCES

[1] G.D. Clifford, F. Azuaje, P.E. McSharry, Advanced Methods and Tools for ECG Data Analysis, Artech House, Norwood, MA, 2006, pp. 55–70.
[2] Q. Li, G.D. Clifford, Signal quality and data fusion for false alarm reduction in the intensive care unit, J. Electrocardiol. 45 (2012) 596–603.
[3] G.B. Moody, R.G. Mark, QRS morphology representation and noise estimation using the Karhunen–Loève transform, Comput. Cardiol. 16 (1989) 269–272.
[4] Q. Li, R.G. Mark, G.D. Clifford, Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter, Physiol. Measure. 29 (2008) 15–32.
[5] S.J. Redmond, N.H. Lovell, J. Basilakis, B.G. Celler, ECG quality measures in telecare monitoring, in: Proceeding of 30th Annual International IEEE EMBS Conference, 2008, pp. 2869–2872.
[6] I. Silva, G.B. Moody, L. Celi, Improving the quality of ECGs collected using mobile phones: the PhysioNet/Computing in Cardiology Challenge 2011, Comput. Cardiol. 38 (2011) 273–276.
[7] G.D. Clifford, G.B. Moody, Signal quality in cardiorespiratory monitoring, Physiol. Measure. 33 (2012) E01, http://dx.doi.org/10.1088/0967-3334/33/9/E01.
[8] G.D. Clifford, J. Behar, Q. Li, I. Rezek, Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments, Physiol. Measure. 33 (2012) 1419–1433.
[9] D. Hayn, B. Jammerbund, G. Schreier, QRS detection based ECG quality assessment, Physiol. Measure. 33 (2012) 1449–1461.
[10] H. Xia, G. Garcia, J. Bains, D. Wortham, X. Zhao, Matrix of regularity for improving the quality of ECGs, Physiol. Measure. 33 (2012) 1535–1548.

[11] Y. Chen, H. Yang, Self-organized neural network for the quality control of 12-lead ECG signals, Physiol. Measure. 33 (2012) 1399–1418.
[12] L.Y. Di Marco, W. Duan, M. Bojarnejad, D. Zheng, S. King, A. Murray, P. Langley, Evaluation of an algorithm based on single-condition decision rules for binary classification of 12-lead ambulatory ECG recording quality, Physiol. Measure. 33 (2012) 1435–1448.
[13] I. Jekova, V. Krasteva, I. Christov, R. Abacherli, Threshold-based system for noise detection in multilead ECG recordings, Physiol. Measure. 33 (2012) 1463–1477.
[14] J. Johannesen, L. Galeotti, Automatic ECG quality scoring methodology: mimicking human annotators, Physiol. Measure. 33 (2012) 1479–1489.
[15] S.J. Redmond, Y. Xie, D. Chang, J. Basilakis, N.H. Lovell, Electrocardiogram signal quality measures for unsupervised telehealth environments, Physiol. Measure. 33 (2012) 1517–1533.
[16] J. Behar, J. Oster, Q. Li, G.D. Clifford, ECG signal quality during arrhythmia and its application to false alarm reduction, IEEE Trans. Biomed. Eng. 60 (2013) 1660–1666.
[17] M. Vaglio, L. Isola, G. Gates, F. Badilini, Use of ECG quality metrics in clinical trials, Comput. Cardiol. 37 (2010) 505–508.
[18] P.X. Quesnel, A.D.C. Chan, H. Yang, Real-time biosignal quality analysis of ambulatory ECG for detection of myocardial ischemia, in: Proceedings of IEEE International Symposium on Medical Measurements and Applications, 2013, http://dx.doi.org/10.1109/MeMeA.2013.6549694.
[19] G.B. Moody, R.G. Mark, The impact of the MIT-BIH Arrhythmia Database, IEEE Eng. Med. Biol. 20 (2001) 45–50.
[20] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P. Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet. Components of a new research resource for complex physiologic signals, Circulation 101 (2000) e215–e220.
[21] G.B. Moody, W.E. Muldrow, R.G. Mark, A noise stress test for arrhythmia detectors, Comput. Cardiol. 11 (1984) 381–384.
[22] http://www.physionet.org/physiotools/wag/wave-1.htm
[23] L.R. Rabiner, J.H. McClellan, T.W. Parks, FIR digital filter design techniques using weighted Chebyshev approximation, Proc. IEEE 63 (1975) 595–610.
[24] P.S. Hamilton, W.J. Tompkins, Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database, IEEE Trans. Biomed. Eng. 33 (1986) 1157–1165.
[25] W. Zong, G.B. Moody, D. Jiang, A robust open-source algorithm to detect onset and duration of QRS complexes, Comput. Cardiol. 30 (2003) 737–740.
[26] S. Nemati, A. Malhotra, G.D. Clifford, Data fusion for improved respiration rate estimation, EURASIP J. Adv. Signal Process. 2010 (2010) 926305.
[27] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. – Heart Circ. Physiol. 278 (2000) H2039–H2049.
[28] R. Sameni, C. Jutten, M.B. Shamsollahi, Multichannel electrocardiogram decomposition using periodic component analysis, IEEE Trans. Biomed. Eng. 55 (2008) 1935–1940.
[29] C.C. Chang, C.J. Lin, LIBSVM. A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (27) (2011) 27.
[30] C.C. Chang, C.J. Lin, LibSVM – a library for support vector machines, http://www.csie.ntu.edu.tw/cjlin/libsvm/
[31] K.L. Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters, 3rd ed., Advanced Analytics, LLC, P.O. Box 2696, Gaithersburg, MD 20886-2696, USA, 2012, March.
[32] V.X. Afonso, W.J. Tompkins, T.Q. Nguyen, K. Michler, S. Luo, Comparing stress ECG enhancement algorithms, IEEE Eng. Med. Biol. 15 (1996) 37–44.
[33] Q. Li, C. Rajagopalan, G.D. Clifford, Ventricular fibrillation and tachycardia classification using a machine learning approach, IEEE Trans. Biomed. Eng. 61 (2014) 1607–1613.