ECG Signal Quality During Arrhythmia and Its Application to False Alarm Reduction

Joachim Behar*, Julien Oster, Qiao Li, and Gari D. Clifford, *Senior Member, IEEE*

***Abstract*—An automated algorithm to assess electrocardiogram (ECG) quality for both normal and abnormal rhythms is presented for false arrhythmia alarm suppression of intensive care unit (ICU) monitors. A particular focus is given to the quality assessment of a wide variety of arrhythmias. Data from three databases were used: the Physionet Challenge 2011 dataset, the MIT-BIH arrhythmia database, and the MIMIC II database. The quality of more than 33 000 single-lead 10 s ECG segments were manually assessed and another 12 000 bad-quality single-lead ECG segments were generated using the Physionet noise stress test database. Signal quality indices (SQIs) were derived from the ECGs segments and used as the inputs to a support vector machine classifier with a Gaussian kernel. This classifier was trained to estimate the quality of an ECG segment. Classification accuracies of up to 99% on the training and test set were obtained for normal sinus rhythm and up to 95% for arrhythmias, although performance varied greatly depending on the type of rhythm. Additionally, the association between 4050 ICU alarms from the MIMIC II database and the signal quality, as evaluated by the classifier, was studied. Results suggest that the SQIs should be rhythm specific and that the classifier should be trained for each rhythm call independently. This would require a substantially increased set of labeled data in order to train an accurate algorithm.**

***Index Terms*—Electrocardiogram (ECG), intensive care unit (ICU), signal quality.**

I. Introduction

IT is very common for the electrocardiogram (ECG) to be contaminated by noise and artifacts. Worse yet, often this noise frequency content overlaps with the frequency band of interest (thus limiting denoising approaches in the frequency domain) or has morphology similar to the one of the ECG (thus limiting denoising approaches in the time domain) [1].

Identifying bad-quality ECG signals is of tremendous importance when considering using automated signal processing algorithms for screening or monitoring cardiac conditions. For example in the context of Holter ECG, which is often corrupted by a significant amount of motion artifact, or when considering telemonitoring systems.

In particular, in the intensive care unit (ICU) the ECG is often severely corrupted by noise, artifact, and missing data, which leads to high false alarm (FA) rates from ICU monitors. Alarms are triggered when parameters are not within a defined range whose default setting varies from one monitor to another. However, in practice, there can be wide variations of a given parameter without alteration of the physiological function [2]. The alarm will be true with respect to the monitoring system (e.g., heart rate (HR) crossed a certain threshold) but will not have any clinical significance. Studies performed in paediatric and critical care units reported that less than 10% of alarms are associated with therapeutic modification [3]–[5]. ICU FA rates up to 86% have been reported [3] and lead to an increase in the workload of intensive care staff [6] and eventually their desensitization [2]. False arrhythmia alarms are often caused by single-channel ECG artifacts and low-voltage signals [7].

This paper addresses the issue of studying the association between ECG signal quality and the alarms triggered by the ICU monitor. This work is intended to provide real-time information on the diagnostic quality of the ECG which will help the ICU monitoring system to decide whether or not to trigger the alarm. In this context, the focus is twofold: assessing the quality of single-lead ECG channels so that ECG derived HR could be computed more consistently from the ECG lead with the best quality. The second aim is to assess the ability of our signal quality indices (SQIs) system to identify bad-quality ECG records and, thus, not to trigger FA while keeping all true alarms (TAs). The two topics are distinguishable as in the first, we assess the ability of our system to distinguish between good- and bad-quality single-lead ECGs but without studying the direct association between ECG quality and alarm status. In the second scenario, we study the direct link between ECG signal quality and the decision of whether the alarm should be canceled. We use machine learning techniques for quality classification based on a pool of features corresponding to several SQIs. In the following, the term “signal” is used to designate a single-lead ECG.

II. Background

Different methods have been employed to reduce ICU FAs (interested readers are referred to Aboukhalil *et al.* for a good review [7]). These approaches can be divided into two categories: data fusion and filtering. In particular, in recent work Aboukhalil *et al.* [7] used morphological and timing information derived from the arterial blood pressure (ABP)

Manuscript received July 30, 2012; revised October 29, 2012 and December 11, 2012; accepted January 4, 2013. Date of publication January 15, 2013; date of current version May 15, 2013. The work of J. Behar was supported in part by the U.K. Engineering and Physical Sciences Research Council and in part by the Balliol French Anderson Scholarship Fund. The work of J. Oster was supported by the Royal Academy of Engineering under a Newton International Fellowship. The work of Q. Li was supported by the Mindray North America. *Asterisk indicates corresponding author.*

*J. Behar is with the Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, OX1 3PJ, U.K. (e-mail: joachim.behar@eng.ox.ac.uk).

J. Oster, Q. Li, and G. D. Clifford are with the Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, OX1 3PJ, U.K. (e-mail: julien.oster@eng.ox.ac.uk; qiao.li@eng.ox.ac.uk; gari@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2013.2240452

TABLE I
DATABASES AND RHYTHMS

Databases		Rythms
Extended CinC [8]	DB_1	NSR
MIT-BIH Arrhythmia [9]	DB_2	AFIB, SVTA, AFL, SBR, VT, VFL, A, V
MIMIC II [10]	DB_3	ASYS, SBR, TACHY, VT, VFIB

Acronyms: NSR: normal sinus rhythm, AFIB: atrial fibrillation, SVTA: supraventricular tachyarrhythmia, AFL: atrial flutter, SBR: sinus bradycardia, VT: ventricular tachycardia, VFL: ventricular flutter, A: atrial premature beat, V: premature ventricular contraction.

signal in order to classify alarms as true or false. A rule-based system was created for each type of arrhythmia. The algorithm was optimized on 2915 arrhythmia alarms extracted from 267 ICU patients and tested on 2471 arrhythmia alarms from another 180 patients. The authors reported 63.2% FA reduction and 1.4% TA suppression on the test set. However, since, their window of 17 s ABP waveform segment was extracted with 13 s prior the alarm and 4 s after the alarm, it is not a causal alarm suppression algorithm, which is unideal. Most work aimed at reducing the number of FA made use of additional waveforms (blood pressure, photoplethysmography, etc.) and to our knowledge there has been no thorough study of FA reduction using ECG quality assessment only.

III. METHODS

A. Databases

For this work, ECG records from three different databases were used: the Physionet/Computing in Cardiology (CinC) Challenge 2011 [8], the MIT-BIH arrhythmia [9], and the MIMIC II [10] databases denoted DB_1, DB_2, DB_3, respectively, (see Table I). DB_1 was used in order to train an ECG quality assessment model. DB_2 was used to test how the model performs on arrhythmic records as well as on records of a different modality (Holter ECG) and was then used in the training set to account for pathological signals. Finally, the resulting model was tested on ECG segments extracted from DB_3 in order to study how effectively ECG quality assessment can identify artifactual signals and to assess the association between ECG quality and FA in the ICU context.

The Physionet/CinC DB includes 1500 10 s recordings of standard 12-lead ECGs, i.e., 18 000 (12 × 1500) individual segments with full diagnostic bandwidth (0.05–100 Hz), sampled at 500 Hz with 16-bit resolution. The recordings were performed for a minimum of 10 s by nurses, technicians, and volunteers with varying amount of training in recording ECGs. After human annotation, less than one-third of the data were classified as poor quality. Therefore, classes were balanced by generating additional bad-quality data from the good-quality records by adding noise from the NSTDB [11]. The method for balancing the dataset is described in [12]. However, for this work, the baseline wander of the electrode motion and muscle artifact records [11] was removed before adding the noise to the ECGs. This is because in some cases most of the power of the added noise was contained in the baseline, thus, resulting in ECGs not systematically to be considered as bad quality. This resulted in 20 000 10 s ECGs signals for the training set and 10 000 for the test set with approximately half of the signals of good quality and half of bad quality (for both sets). These segments constituted the first dataset, referred to as the extended CinC DB (DB_1, see Table I). The training and test sets for this DB are denoted Set-a‡ and Set-b‡, respectively.

The second dataset, DB_2, includes 48 complete two-lead ECG records with arrhythmia reference annotations from the MIT-BIH arrhythmia DB. The records have a diagnostic bandwidth of 0.1–100 Hz with 12-bit resolution and were digitized at 360 Hz. Six arrhythmia types were analyzed (see Table I). Data were segmented for 10 s after the onset of each arrhythmia. In the event of an arrhythmia being less than 10 s the segment was discarded to avoid mixed rhythms. Atrial premature beats (A) and premature ventricular contraction beats (V) were located and 10 s segments, centered on each beat, while ensuring no overlap between segments, were extracted. Both available leads were used. In most records, the first channel is the modified limb lead II and the second channel is usually a precordial lead V1. This resulted in 9452 individual 10 s segments whose quality was manually evaluated. Within the 9452 signals, 269 were annotated as being of poor diagnostic quality and the remaining 9183 as good or medium quality. The number of good- and bad-quality signals from this dataset was not balanced because it was used as a first layer for evaluating the model on real arrhythmic data.

The third dataset includes a subset of over 4050 life-threatening HR-related arrhythmia alarms taken from DB_3. Five types of annotated arrhythmia were available (see Table I). ECGs were available at 125 Hz and at a resolution of 8 to 10 bits. A team of experts annotated the veracity of each alarm (at least twice, with a separate pass and adjudication for disagreements). Segments of 10 s were extracted from the records starting 10 s prior the alarm was triggered. The American Organization for the Advancement of Medical Instrumentation recommendations [13] specifies that a cardiac alarm should be triggered within 10 s after an arrhythmia is detected. This motivated the choice for working with 10 s segments preceding each alarm. The number of ECG leads per record was variable and all flat line leads were removed from all records. In addition, there were some intrinsic limitations using this DB because the raw signals were not accessible from the monitoring system. Though the ECG signals were originally sampled at a high sampling rate with 12-bit resolution, they were scaled and resampled prior to capture. This reduced resolution from 12 bits to 8 or 10 bits in most cases. Furthermore, only one sample out of four was recorded using a turning-point compressor, thus providing a nonlinearly filtered signal. This resulted in reduced time and amplitude resolution of the ECGs and consequently in a high proportion of ECG signals manually labeled as medium quality. Table II summarizes the distribution of alarms per arrhythmia type. Note that all abnormal rhythms but VT are relatively limited in number.

B. Labeling ECG Signals

The ECG signals were manually annotated using the classification scheme suggested in [14]. In this paper, we only considered two classes: good (A and B) and bad quality (D and E). Segments of class C were not considered because most

TABLE II
Distribution of Annotated Alarms per Arrhythmia Type in the MIMIC II Database

Type	All alarms		True alarms		False alarms	
	Total alarms	% all alarms	N	% true	N	% false
ASYS	606	15%	46	7.6%	560	92.4%
SBR	356	8.8%	257	72.2%	99	27.8%
TACHY	841	20.8%	741	88.1%	100	11.9%
VT	2132	52.6%	1141	53.5%	991	46.5%
VFIB	115	2.8%	25	21.7%	90	78.3%
Overall	4050		2210		1840	

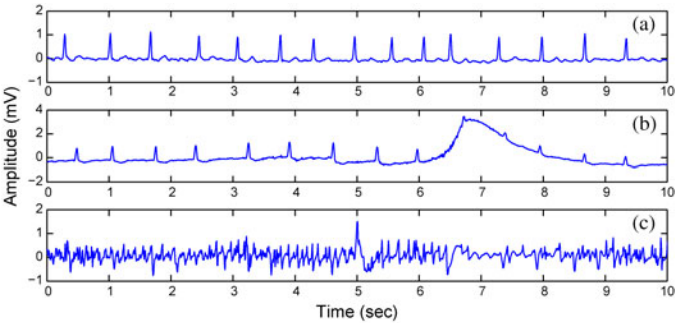

Fig. 1. Example of lead quality: (a) good quality, (b) medium quality, and (c) bad quality. ECG segments are taken from the MIT-BIH arrhythmia DB.

TABLE III
Distribution of Signal Quality Across All Databases

Database \Quality	# Bad	# Medium	# Good	Total #
Extended CinC (DB_1)	14968	1977	13055	30000
MIT-BIH arrhythmia (DB_2)	269	415	8768	9452
MIMIC II (DB_3)	1080	780	3774	5634

TABLE IV
Distribution of Signal Quality per Arrhythmia Type for the MIT-BIH Arrhythmia DB

Arrhythmia \Quality	# Bad	# Medium	# Good	Total #
AFIB	24	87	1261	1372
SVTA	0	2	12	14
AFL	3	8	81	92
SBR	1	24	333	358
VT	1	0	11	12
VFL	1	1	14	16
A	55	106	1453	1614
V	184	187	5603	5974
Overall	269	415	8768	9452

of the time they were borderline, either because the level of noise was not high enough to cause bad QRS detections or because the noise was very localized within the 10 s segment. Clinically speaking, these ECGs were often interpretable but there were transient artifacts or a moderate amount of noise present making them prone to label disagreement between annotators. Therefore, it should not be considered wrong for the algorithm to classify such segments as either good or bad quality. As a consequence leads of type C were excluded in order to avoid confusing the classifier during model training.

Fig. 1 shows examples of quality labels. We used 14 968 bad-quality (10 020 in Set-a‡ and 4948 in Set-b‡) and 13 055 good-quality signals (8745 in Set-a‡ and 4310 in Set-b‡) from DB_1 as well as 269 bad-quality and 8768 good-quality signals from DB_2. The distribution of per class signal quality is summarized in Table III. Table IV gives the distribution of signal quality per arrhythmia type for DB_2. Although we might want to modify the window size on which we assess the signal quality, we made the choice of working with 10 s segments of a given rhythm in order to derive consistent statistics (e.g., kurtosis, skewness) on which to build up our model. No mixed rhythms were retained for the DB_2 but segments with various numbers of A or V beats were considered.

TABLE V
Repartition of Record Quality per Arrhythmia Type for TAs in the MIMIC II DB

Arrhythmia \Quality	# Bad	# Good	Total #
ASYS	39	62	101
SBR	2	231	233
TACHY	46	835	881
VT	34	1370	1404
VFIB	7	34	41
All TA	128	2532	2660
FA (all arrhythmias)	952	1242	2194
Overall	1080	3774	4854

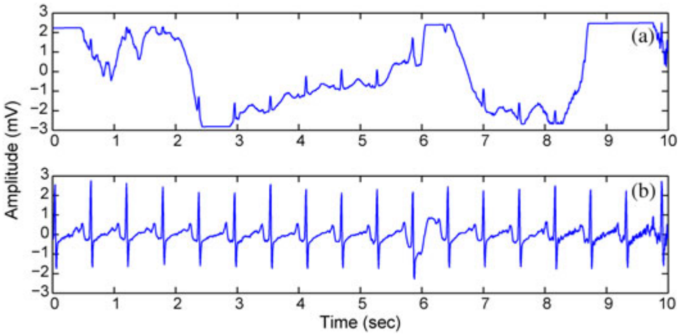

Fig. 2. Example of TA (TACHY) with (a) bad- and (b) good-quality ECGs. Taken from the MIMIC II DB.

Furthermore, we manually annotated about half of the records from DB_3. This provided an additional 5634 annotated signals as detailed in Table III. Table V summarizes the repartition of good- and bad-signal quality per arrhythmia type (i.e., TA of a given type of arrhythmia that have been labeled as good or bad quality). Note that a non-negligible number of signals were manually annotated as bad quality although they were associated with a TA (see Table V). Conversely, about 43% of bad-quality signals were associated with FA signals. Indeed, FA and TA records might contain a mixture of good- and bad-quality single-lead signals. Fig. 2 shows an example of a two-lead TA record where one lead is of bad quality and the other is of good quality. Note that the 10 s segments were selected with respect to the time that the alarm was triggered and consequently they may contain mixed rhythms (the alarm being triggered within 10 s of detection of an arrhythmic rhythm).

C. *Preprocessing and SQIs*

Each channel of ECG was downsampled to 125 Hz when necessary and for computing efficiency. QRS detection was performed on each channel individually using two open source QRS detectors (*eplimited* [15] and *wqrs* [16], [17]). The *eplimited* algorithm is less sensitive to noise than *wqrs* [18] and consequently discrepances in their output is an indicator of noise (see bSQI and rSQI below). Seven SQIs introduced in previous

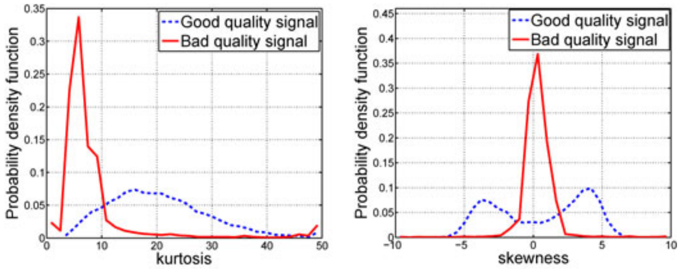

Fig. 3. Probability density function of (left) kurtosis and (right) skewness for 21 823 10 s signals from the CinC DB (13 055 good quality and 14 968 bad quality) before normalization.

works [12], [18], [19] were calculated for every single-lead separately as follows:

1) kSQI: The fourth moment (kurtosis) of the signal is defined as kSQI $= E\{X - \mu\}^4/\sigma^4$, where X is the signal vector considered as the random variable, μ is the mean of X, σ is the standard deviation of X, and $E\{X - \mu\ \}$ is the expected value of the quantity $X - \mu$. We expect a good ECG to be highly non-Gaussian since it is not very random.
2) sSQI: The third moment (skewness) of the signal defined as sSQI $= E\{X - \mu\}^3/\sigma^3$. We expect the ECG to be highly skewed due to the QRS complex.
3) pSQI: The relative power in the QRS complex: $\int_{5\,\mathrm{Hz}}^{15\,\mathrm{Hz}} P(f)\,\mathrm{d}f / \int_{5\,\mathrm{Hz}}^{40\,\mathrm{Hz}} P(f)\,\mathrm{d}f$. We expect most of the power to be in the 5–15-Hz band.
4) basSQI: The relative power in the baseline: $\int_{1\,\mathrm{Hz}}^{40\,\mathrm{Hz}} P(f)\,\mathrm{d}f / \int_{0\,\mathrm{Hz}}^{40\,\mathrm{Hz}} P(f)\,\mathrm{d}f$. A sudden "low frequency (≤ 1 Hz) bump" will result in low basSQI.
5) bSQI: The fraction of beats detected by *wqrs* that matched with beats detected by *eplimited*.
6) rSQI: The ratio of the number of beats detected by *eplimited* and *wqrs*.
7) pcaSQI: A ratio comprising of the sum of the eigenvalues associated with the five principal components over the sum of all eigenvalues obtained by principal component analysis applied to the time-aligned ECG cycles detected in the window by the *eplimited* algorithm, segmented at 100 ms either side of the R-peak.

Note that baseline wander was filtered prior to computing sSQI and kSQI using a second-order zero-phase high-pass filter with 0.7 Hz cut-off frequency. Power spectral density was evaluated using the Burg method [20] with 11 poles. Fig. 3 shows the probability density function of kurtosis and skewness obtained for 21 823 10 s signals with 13 055 good-quality signals and 14 968 bad-quality signals before normalization. The distributions suggest that the two statistical measures might provide information for distinguishing between good- and bad-quality signals because of the limited overlap between the distributions.

D. *Support Vector Machine Classifier*

We used a support vector machine (SVM) classifier (libSVM library [21], [22]) with a Gaussian (nonlinear) kernel defined by $k(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|^2\right)$, where γ controls the width of the Gaussian and plays a role in controlling the flexibility of the resulting classifier. $\mathbf{x}_n$ and $\mathbf{x}_m$ are two vectors expressed in the initial feature space. The SVM with a Gaussian kernel has two parameters: C and γ. Based on cross validation in [12], we used $C = 25$ and $\gamma = 1$.

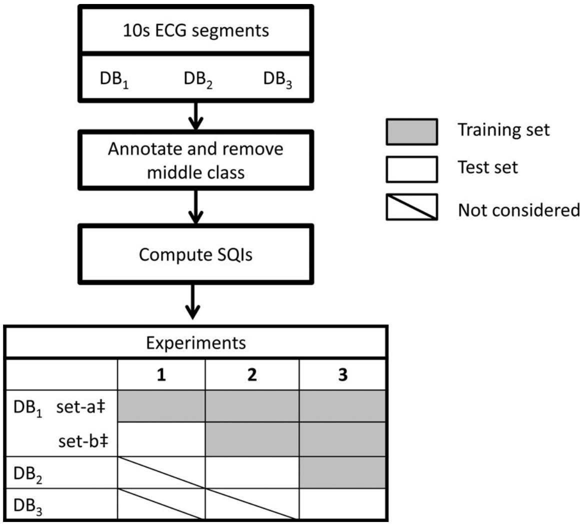

Fig. 4. Flowchart of experiments. DB$_1$: extended CinC DB, DB$_2$: MIT-BIH arrhythmia DB, DB$_3$: MIMIC II DB. Set-a‡: training set of DB$_1$ and Set-b‡: test set of DB$_1$. Experiments 1, 2, and 3 are described in Section III. The databases are described in Tables I–V.

The choice of SVM over alternative machine learning methods is justified by the property that computation of the model parameters is a convex optimization problem and by consequence local solutions are also global optimum [23]. Thus, given a set of input features derived from a dataset and given C and γ parameters, the SVM will always converge to the same solution. Thus, SVMs are stable nonlinear classifiers for repeated development. Moreover, in [19], we used a neural network in addition to the SVM and showed that the two classifier gave extremely similar results, as would be expected if both were optimized correctly [24].

All SQIs except sSQI and kSQI take their value in the range $[0, 1]$. Therefore, sSQI and kSQI were normalized by subtracting the mean and dividing by the standard deviation, with both of which computed on the training set.

E. *Statistical Measures*

Sensitivity (Se) measures the proportion of poor quality signals that have been correctly identified as such. Specificity (Sp) measures the proportion of good quality signals that have been correctly identified as acceptable, and accuracy (Ac) corresponds to the proportion of signals that have correctly been classified. These statistical measures are calculated from the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) with Se = TP/(TP+FN), Sp = TN/(FP+TN) and Ac = (TP+TN)/(TP+TN+FP+FN).

F. *Quality Assessment*

The classification quality assessment was conducted in three phases. Fig. 4 shows a flowchart summarizing how the experiments were conducted.

TABLE VI
SINGLE-LEAD CLASSIFICATION USING SVM ON INDIVIDUAL SQIs AND ON SET-A‡

	bSQI	kSQI	sSQI	pSQI	basSQI	qSQI	pcaSQI
Ac	0.969	0.879	0.892	0.752	0.626	0.766	0.950
Se	0.983	0.911	0.944	0.806	0.679	0.805	0.945
Sp	0.953	0.842	0.831	0.690	0.564	0.721	0.956

TABLE VII
SINGLE-LEAD CLASSIFICATION PERFORMANCE USING SVM WITH COMBINATIONS OF SQIs

	Ac train Set-a‡	Ac test Set-b‡
bSQI, pcaSQI	0.974	0.971
bSQI, pcaSQI, qSQI	0.982	0.980
bSQI, pcaSQI, kSQI, sSQI	0.986	0.981
bSQI, pcaSQI, kSQI, basSQI, qSQI	0.988	0.983
bSQI, pcaSQI, pSQI, sSQI, qSQI, basSQI	0.991	0.985
bSQI, pcaSQI, kSQI, sSQI, qSQI, basSQI, pSQI	0.993	0.984

Classification is performed on DB_1.

TABLE VIII
SINGLE-LEAD CLASSIFICATION USING SVM

	Quality			Statistics		
	# Bad	# Good	Total #	Ac	Se	Sp
AFIB	24	1261	1285	0.967	0.958	0.968
SVTA	0	12	12	1	0	1
AFL	3	81	84	0.833	1	0.827
SBR	1	333	334	0.961	1	0.961
VT	1	11	12	0.833	1	0.818
VFL	1	14	15	0.467	1	0.429
A	55	1453	1508	0.972	0.927	0.974
V	269	5603	5787	0.936	0.826	0.940
Overall	269	8768	9037	0.946	0.863	0.948

Classifier is trained on DB_1 and DB_2.

1) The classifier model was built on DB_1 and the performance of the individual SQIs was evaluated followed by the evaluation of all other possible combinations of SQIs (pairs, triplets, etc.). This first test allowed the study of the contribution of each SQI on the classification performance.
2) The classifier was trained on all good- and bad-quality data from DB_1 and the performance of the trained classifier was assessed on DB_2. DB_1 and DB_2 were then merged and used as the new training set. Five fold cross validation was performed on this set in order to assess the performance of the predictive model. By performing cross validation in phase 2 we also intrinsically validate the predictive model built upon phase 1 since all the data from experiment one are also considered in experiment two and they represent approximately three quarters of the overall data used for cross validation.
3) Classification performance evaluation on DB_3; the classifier was trained on the combination of DB_1 and DB_2 and performance was tested on DB_3.

G. *Association Between Quality and Alarms Status*

Finally, we studied the association between quality assessed by the classifier and alarm status (TA, FA) for DB_3. We computed the probability estimate of the quality for all signals of the 4050 MIMIC II records and retained the maximum quality estimate (i.e., the quality probability of the lead with the best quality) in order to limit the number of FP. If the best single-lead signal quality of a given record was inferior to a given threshold then we classified it as an FA and if higher than the threshold we classified it as TA.

IV. RESULTS

A. *Classification Model*

For each of the three experiments, the results are now detailed:

1) Table VI shows that bSQI and pcaSQI best distinguish between records of good and bad quality on training Set-a‡. The SVM was then trained with all possible combinations of SQIs. The results for the best pair, triplet, etc., of SQIs combinations are summarized in Table VII. The best accuracy on the training set (Set-a‡) was obtained when considering all SQIs (Ac = 99.3%). However, the best accuracy (Ac = 98.5%) on the test Set-b‡ was obtained using only the best six SQIs. Although using all SQIs provided a negligible drop in test performance (to 98.4%) compared to using the best combination of six SQIs, we decided to keep all the SQIs for the later experiments, since the McNemar's test showed no statistically significant difference (with $\chi^2 = 1.32$ and $p = 0.25$).
2) Table VIII shows the results obtained on DB_2 per arrhythmia while training on DB_1. Note that the statistical measures (Ac, Se, Sp) are not always significant due to the very small number of signals for a given type of arrhythmia. We obtained 94.6% overall accuracy for the DB_2. Repeating fivefold cross validation 150 times on the combination of DB_1 and DB_2 resulted in a mean accuracy of $\mu = 0.991$ with standard deviation $\sigma = 2.86 \times 10^{-4}$ and 95% confidence interval (CI), $\mathrm{CI} = [0.990; 0.991]$ on the training set and $\mu = 0.986, \sigma = 1.20 \times 10^{-3}, \mathrm{CI} = [0.983; 0.988]$ on the validation set. Note that cross validation was repeated 150 times to ensure a stable value of accuracy and we found additional repetitions were not required as the variance estimates had converged.
3) Training the classifier on the combination of DB_1 and DB_2 and testing on DB_3 labeled signals resulted in $\mathrm{Ac} = 0.893$, $\mathrm{Se} = 0.744$, and $\mathrm{Sp} = 0.935$ on the test set. Fig. 5 shows the receiver operating characteristic (ROC) curve obtained on this dataset. Table IX reports the results per arrhythmia type.

B. *Association Between Quality and Alarms Status*

Fig. 5 shows the corresponding ROC curve. In the context of ICU FA suppression, the FP rate should be close to 0 (i.e., no TA suppressed) since missing a TA might lead to adverse consequences. Fig. 6 shows the association between FA and bad-quality ECG as assessed by the classifier for each of the five arrhythmia types. The number of alarms and the proportion of TA/FA is variable for each type of arrhythmia (see Table II) with about 2132 VT alarms (53.5% TA and 46.5% FA) and 115 VFIB (1.1% TA and 78.3% FA). As a consequence no final conclusion can be drawn for most of the pathological rhythms. However, in the case of VT, we note a 13% FA suppression for

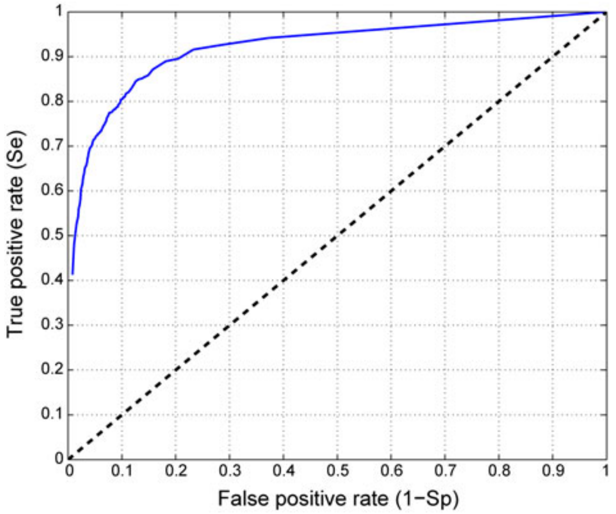

Fig. 5. ROC curve for the FA suppression algorithm on the data listed in Table IX (the MIMIC II database). Dotted line: random unbiased “coin toss” between true and false class.

TABLE IX
LEAD-QUALITY ASSESSMENT PER ARRHYTHMIA TYPE ON MIMIC II DB

	Quality			Statistics		
	# Bad	# Good	Total #	Ac	Se	Sp
TA-ASYS	39	62	101	0.683	0.974	0.500
TA-SBR	2	231	233	0.708	0	0.714
TA-TACHY	46	835	881	0.943	0.804	0.951
TA-VT	34	1370	1404	0.961	0.677	0.968
TA-VFIB	7	34	41	0.902	0.571	0.971
All TA	128	2532	2660	0.902		
All FA	952	1242	2194	0.857	0.737	0.949
Overall	1080	3774	4854	0.892	0.744	0.935

0.4% TA suppression which suggests that the quality assessed by the classifier is associated with the alarm status. Conversely, FA suppression for ASYS is associated with a high rate of TA suppression (the blue crosses are under the line FA = TA in Fig. 6).

V. DISCUSSION AND CONCLUSION

While training on the combination of DB_1 and DB_2 we noted that by taking into account only the best performing metric (bSQI), we obtained $Ac = 0.925$, $Se = 0.959$, and $Sp = 0.901$ whereas considering all the SQIs resulted in $Ac = 0.990$, $Se = 0.985$, and $Sp = 0.994$. McNemar’s test was performed in order to test the difference in classification when using bSQI alone and the seven SQIs. We obtained $\chi^2 = 995.3$, $p \ll 0.01$. This demonstrates the added value of using a machine learning framework that combines information provided by multiple measures instead of assessing the quality using one SQI with a single parameter threshold.

Despite the promising results on healthy ECGs, the accuracies for sinus bradycardia and asystole, when testing on the MIMIC II data, are very low. Outcomes for asystole are not surprising as our system is likely to classify this type of arrhythmia as bad quality (Sp not being closer to 0 might be explained by the 10 s window involving mixed rhythm). Thus, asystole should be considered as a particular case. Further analysis of the outcomes for sinus bradycardia showed that the statistically based SQIs were the main factors resulting in failing to better classify sinus bradycardia. Removing pcaSQI increased accuracy for this arrhythmia type by 10%. Furthermore, removing kSQI and sSQI increased accuracy another 18% resulting in accuracy very close to $Ac = 100\%$. This strongly suggests two primary considerations for classifying arrhythmias; First, as certain arrhythmias may behave very similar to noisy data according to certain SQIs, arrhythmia specific models are necessary. For example, sinus bradycardia is not expected to be classified well as the SQI behavior is distinctly different in sinus bradycardia than for other rhythms. Second, enough training data for the rhythms of interest would allow SQIs based on statistics, such as kSQI and sSQI, to properly capture the statistical properties of the different types of arrhythmias. The classifier is then able to learn the distinction between good- and bad-quality data using the captured statistical informations.

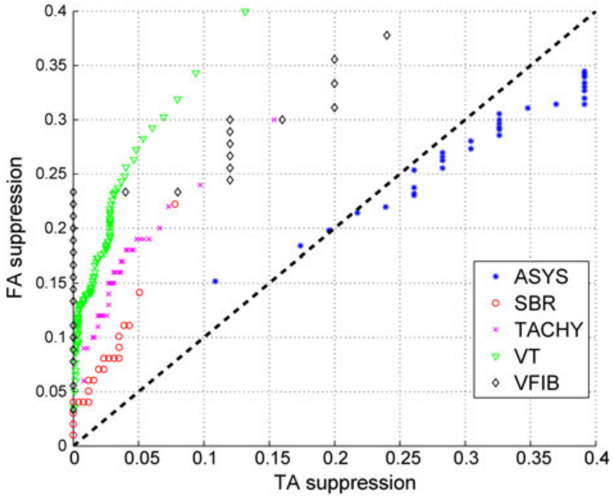

Fig. 6. ROC curves for the FA suppression algorithms per alarm type. The VT FA suppression algorithm (▽) displayed the best performance followed by the VFIB FA suppression algorithm (◇) followed by the TACHY FA suppression algorithm (x) followed by the SBR FA suppression algorithm (○). The asystole FA suppression alarm (+) performs not better than a random coin toss (dotted line).

Our observations underline the intrinsic limitation of the 2011 CinC challenge [8]. The challenge data are further limited as they were acquired from healthy subjects, and thus, they did not include a wide range of pathological cardiac conditions. Based on our results, we motivate an upgraded framework for constructing a machine learning-based model for ECG quality assessment. The set of SQIs used should be rhythm dependant and the model applied (a model is defined as a trained classifier for a given set of features) should be switched according to a rhythm call. Fig. 7 shows the logic of such a model; the monitoring system identifies a rhythm and the corresponding model is selected to assess the signal quality (Model_k in the figure) and outputs a probability estimate of the signal quality. Building such a system involves the following steps:

1) For each type of abnormal rhythm, select 10 s segments (no mixed rhythms) from all lead configurations.
2) Experts annotate the quality of the signals and only keep good and bad quality ones in order to build the model. These labels constitute our gold standard.

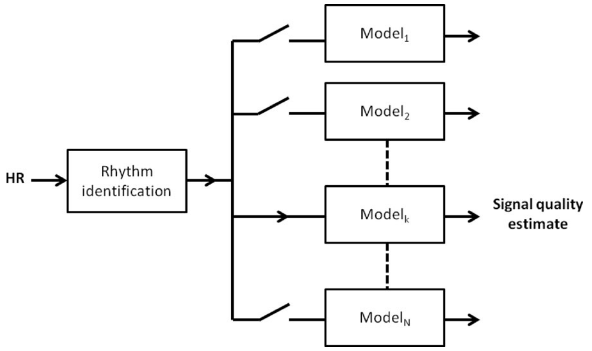

Fig. 7. Model representation: the classifier model is chosen with respect to the identified rhythm and outputs a signal quality estimate.

3) Balance the classes using the approach introduced in [12] (this is because the dataset is likely to have a high prevalence for good-quality signals in a given dataset) and divide into training and test sets.
4) For a given type of rhythm, train the classifier on the training set. This step includes selecting the set of SQIs to use for a given rhythm, which for low number of SQIs can be performed by an exhaustive search.
5) Test validity of model on the test set of data and report the classifier performances.

The main limitation in building such a model is the large and diverse amount of labeled data that is required from a large and varied number of patients and rhythms (to account for the population variability). It should be noted that different preprocessing steps were applied to each database before being made available to the authors and the public. This may alter performance of the quality metrics. Moreover, many of the arrhythmias are underrepresented even in the large databases used, and so further data and annotation is required.

Better results are also expected by improving the gold standard labels. Indeed, this study is based on signal quality annotations from one annotator that attributed a quality label to the ECG signals. However, these annotations are likely to suffer from an expert's individual bias (even with strict guidelines and training).

With regards to the association between ECG quality and alarm status, it is important to note that the channel(s) used by the monitoring system to trigger the alarms were unknown to us. This is the main limitation of the study of the correlation between signal quality and alarm status as we have intrinsically assumed that all channels' signals should be of bad quality in order to consider the alarm to be an FA (which is obviously not true if the alarm is triggered by the recording(s) of only one or a subset of the ECG channels).

Nevertheless, the approach described here demonstrates good results for ectopic beats, tachycardia rhythms, atrial fibrillation, and on sinus rhythm. It can be used for deriving a better HR estimate by selecting the lead with the best quality but in its current state it cannot to be used alone as a decision algorithm for FA suppression in an ICU context, since the tolerance for suppressing a TA should tend toward zero. However, with a larger representation of arrhythmic rhythms and a consequent retraining of the model the approach described here is expected to be useful in the ICU.

References

[1] G. D. Clifford, *ECG Statistics, Noise, Artifacts, and Missing Data, Engineering in Medicine and Biology*, 1st ed. Norwood, MA, USA: Artech House, Oct. 2006.
[2] M. C. Chambrin, "Alarms in the intensive care unit: How can the number of false alarms be reduced?" *Crit. Care*, vol. 5, pp. 184–188, 2001.
[3] S. T. Lawless, "Crying wolf: False alarms in a pediatric intensive care unit," *Crit. Care Med.*, vol. 22, pp. 981–985, 1994.
[4] C. L. Tsien and J. C. Fackler, "Poor prognosis for existing monitors in the intensive care unit," *Crit. Care Med.*, vol. 25, pp. 614–619, 1997.
[5] M. C. Chambrin, P. Ravaux, D. Calvelo, A. Jaborska, C. Chopin, and B. Boniface, "Multicentric study of monitoring alarms in the adult intensive care unit (ICU): A descriptive analysis," *Int. Care. Med.*, vol. 25, pp. 1360–1366, 1999.
[6] J. Allen and A. Murray, "Assessing ECG signal quality on a coronary care unit," *Physiol. Meas.*, vol. 17, pp. 249–58, 1996.
[7] A. Aboukhalil, L. Nielsen, M. Saeed, R. G. Mark, and G. D. Clifford, "Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform," *J. Biomed. Inf.*, vol. 41, pp. 442–451, 2008.
[8] I. Silva, G. B. Moody, and L. Celi, "Improving the quality of ECGs collected using mobile phones: The physionet/computing in cardiology challenge 2011," *Comput. Cardiol.*, vol. 38, pp. 273–276, 2011.
[9] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorf, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley., "Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, pp. 215–220, 2000.
[10] M. Saeed, M. Villarroel, A. T. Reisner, G. D. Clifford, L. W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access ICU database," *Crit. Care Med.*, vol. 39, no. 5, pp. 952–960, 2011.
[11] G. B. Moody, W. E. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Comput. Cardiol.*, vol. 11, pp. 381–384, 1984.
[12] G. D. Clifford, J. Behar, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments," *Phys. Meas.*, vol. 33, pp. 1419–1433, Sep. 2012.
[13] "Cardiac monitors, heart rate meters, and alarms, american national standard (Revision of ANSI/AAMI EC13:2002)," Assoc. for the Advancement of Med. Instrum., Arlington, VA, Tech. Rep., 2002.
[14] G. D. Clifford, D. Lopez, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments," *Comput. Cardiol.*, vol. 38, pp. 285–288, 2011.
[15] P. S. Hamilton. (2002). *Open source ECG analysis software documentation*. [Online]. Available: http://www.eplimited.com/osea13.pdf
[16] W. Zong. (2010). *Single-lead QRS detector based on length transform*. [Online]. Available: http://www.physionet.org/physiotools/wfdb/app/wqrs.c
[17] W. Zong, G. B. Moody, and D. Jiang, "A robust open-source algorithm to detect onset and duration of QRS complexes," *Comput. Cardiol.*, vol. 30, pp. 737–740, 2003.
[18] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter," *Physiol. Meas.*, vol. 29, no. 1, pp. 15–32, 2008.
[19] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "A single channel ECG quality metric," *Comput. Cardiol.*, vol. 39, pp. 381–384, 2012.
[20] J. P. Burg, "Maximum entropy spectral analysis," presented at the 37th Meet. Soc. Exploration Geophys., Oklahoma City, OK., vol. 37, 31 Oct. 1967.
[21] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Sys. Tech.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[22] C. C. Chang and C. J. Lin, *LIBSVM—A library for support vector machines*. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
[23] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.
[24] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Comput.*, vol. 11, no. 2, pp. 305–345, 1999.

Authors' photographs and biographies not available at the time of publication.