Quality Assessment of Ambulatory ECG Using Wavelet Entropy of the HRV Signal

Christina Orphanidou and Ivana Drobnjak

***Abstract*—Data in recordings obtained from ambulatory patients using wearable sensors are often corrupted by motion artefact and are, in general, noisier than the data obtained from the nonmobile patients. Identifying and ignoring erroneous measurements from these data is very important, if wearable sensors are to be incorporated into clinical practice. In this paper, we propose a novel Signal Quality Index, intended to assess whether reliable heart rates can be obtained from a single channel of ECG collected from ambulatory patients, using wearable sensors. The proposed system is based on wavelet entropy measurements of the heart rate variability signal. The system was trained and tested on expert-labeled data from a particular wearable sensor and was also tested on labeled data from a different sensor. The sensitivities and specificities achieved were 94% and 98%, respectively, on data from the same sensor as the training set, and 91% and 97%, respectively, on data from a different sensor, indicating the potential of the system to generalize across different sensors. Because the system relies on a single channel of ECG, it has the potential for inclusion in applications using wearable sensors and in the most basic clinical environments.**

***Index Terms*—Electrocardiogram (ECG), heart rate (HR), heart rate variability (HRV), signal quality, wavelets, wearable sensors.**

I. Introduction

THERE is a widespread consensus that wearable sensors will be a key part of delivering healthcare in the future. However, for them to be successfully incorporated into clinical practice, the technology needs to advance to a reliable level. The issue of identifying unreliable data is particularly important, as data obtained from ambulatory patients, the patients more likely to benefit from the use of wearable sensors, are more likely to contain artefact than data obtained from bed-bound patients [1].

The electrocardiogram (ECG), routinely collected from hospital patients, is often contaminated with noise leading to unreliable vital sign measurements. Erroneous vital sign measurements may result in a large number of false alerts that can lead to the phenomenon of "alarm fatigue," whereby ward staff become desensitized to and ultimately ignore alerts from the monitoring systems [1], [2]. Furthermore, abnormal vital sign measurements, found to be significant predictors of adverse events and mortality in hospital patients [3], [4], may go unnoticed, thus compromising patient care and health outcomes.

In the past few years, a lot of research activity was directed toward the development of artefact detection (AD) algorithms for physiological signals, either based on a single signal (the ECG or the photoplethysmogram (PPG)), or combining information from several different signals, or multiple channels of the same signal. A comprehensive review of AD techniques in critical care units was recently published in [5]. The review highlights the complexity of the task: algorithms must be shown to generalize across units, manufacturers, and patient populations [2], [5].

Recently the proposed AD algorithms, reporting positive results, are based on the fusion of different features extracted from multiple signals, such as the ECG, the PPG, and the arterial blood pressure signal [6]. In the case of the ECG, the proposed quality assessment methods were based on features, such as metrics of agreement between two different QRS detectors [6], [7] and spectral density ratios between different frequency bands [7]. Especially when frequency-based features were used, the type of rhythm present was shown to be important (i.e., whether a particular type of arrhythmia is present) with algorithms needing to be tailored to the specific variety of arrhythmia [2]. Despite the impressive results reported by many of the proposed algorithms, the same issues persist: most proposed systems are tailored to a specific patient population, sensor, or manufacturer, and would require modification for validation and reuse with a different one. Algorithms requiring the presence of different signals or multiple channels of the same signal will not be usable in many clinical environments (e.g., in the context of m-health applications for the third world). Finally, often, the proposed algorithms are trained on nonclinical data with artificial noise. This is an important weakness since systems targeted for use by possibly anxious, unwell patients for extended periods of time need to be designed so as to reflect these characteristics in the algorithm specifications. Recently, a single channel Signal Quality Index (SQI) was proposed for the ECG based on QRS template matching [1], [8], trained and validated on real-world clinical data. While the proposed system showed promise, its performance was not consistent across all sensors tested.

In this paper, we propose a new algorithm for classifying segments of ECG as "acceptable" or "unacceptable" (for obtaining reliable heart rate (HR) measurements), which is based on spectral analysis of the heart rate variability (HRV) signal.

Manuscript received April 8, 2016; revised August 12, 2016 and August 31, 2016; accepted September 30, 2016. Date of publication October 5, 2016; date of current version September 1, 2017.

C. Orphanidou is with the KIOS Research Center, Department of Electrical and Computer Engineering, University of Cyprus, Nicosia 1678, Cyprus (e-mail: orphanid@ucy.ac.cy).

I. Drobnjak is with the Department of Computer Science, University College London, London WC1E 6BT U.K. (e-mail: I.Drobnjak@cs.ucl.ac.uk).

Digital Object Identifier 10.1109/JBHI.2016.2615316

HRV refers to the variation over time of the intervals between consecutive heartbeats. Since the heart rhythm is modulated by the autonomic nervous system (ANS), HRV is thought to reflect the activity of the sympathetic and parasympathetic branches of the ANS [9]. Analysis of HRV has been found to be clinically useful in evaluating a number of cardiovascular conditions and disorders [9], [10]. The beat-to-beat changes reflected in the HRV signal, occur at multiple frequencies and the signal is, in general, regarded as the sum of several physiologically relevant components occurring at different frequency subbands: the very low-frequency (VLF) component (frequencies below 0.03 Hz), which is modulated by the renin-angiotensin system; the low-frequency (LF) component (0.03–0.15 Hz), which is thought to be related to both sympathetic and parasympathetic activity of the heart; and the high-frequency (HF) component (0.15–0.4 Hz), which is mostly related to the parasympathetic system and has been found to contain the respiration frequency [9]–[12] (the upper and lower bounds of these spectral components are not strictly defined and may appear slightly different in some research articles).

Our proposal is to use the spectral characteristics of the HRV signal as a measure for assessing signal quality. The HRV signal derived from a "clean" segment of ECG is rich in physiologically relevant information with the literature indicating that the upper frequency limit of the highest band for HRV analysis is 0.4 Hz. Our hypothesis is, therefore, that the HRV signal derived from a "clean" segment of ECG should have most of its energy concentrated in the frequency bands below 0.4 Hz. In the presence of noise, errors in the detection of QRS peaks will result in a distorted HRV signal. The distorted HRV signal would probably have a disordered distribution of energy in the different frequency bands and would presumably contain increased energy in nonphysiologically relevant bands as well. Our rationale, therefore, is that the spectral content of HRV signals obtained from "clean" segments of ECG would differ from that of the distorted HRV signals obtained from "noisy" segments. Furthermore, since the HRV signal may be obtained in a standardized way from any ECG signal of sufficiently high sampling rate, using a universal QRS detector, it has the potential to be the basis of an SQI which can generalize across different sensors and manufacturer specifications. An important consideration for any application involving HRV spectral analysis is the window size required for any useful indices to be obtained. While the standard recommendation for obtaining reliable HRV measurements has been set at a minimum of 5 min [11], studies have shown that HF components can be satisfactorily analyzed in window sizes as small as 20 s [13]. To balance the requirement of real-time implementation and the need to identify meaningful differences between "clean" and "noisy" signals, we chose a window size of 30 s. While this window size is not sufficiently large for clearly identifying VLF and LF components in the HRV signal, it is large enough to identify differences between "clean" and "noisy" signals in the HF band.

For the spectral analysis of the HRV signal, we propose using wavelet entropy measurements. Wavelet entropy has been proposed in the past as a measure for diagnosing congestive heart failure [14] and for the prognosis of cardiovascular risk [15] from the HRV signal; however, to the best of our knowledge, it has not been used in the context of signal quality assessment via the HRV signal. We chose to use wavelet entropy measurements because the main idea fits well with our task at hand; if a system exhibits "disordered" behavior at a specific frequency subband, a high-entropy value will be obtained. Therefore, the *distorted* HRV signal obtained from a "noisy" segment of ECG will have high-entropy values in different subbands compared to the *true* HRV signal obtained from a "clean" segment of ECG which should have most of its energy contained in the physiologically relevant subbands. Since our aim is for the proposed SQI to work in the presence of different (and multiple) kinds of noise, the wavelet decomposition scheme offers the flexibility that all the signal frequency components may be examined and taken into consideration when building the classifier, such that different types of noise can be accounted for.

The relationship between the wavelet entropy measurements and occurrences of noise was learned using support vector machines (SVM) leading to a classifier which labels segments of ECG signal as "acceptable" (clean) or "unacceptable" (noisy).

Finally, in contrast to many other proposed systems, we used real clinical data for training and validating the system, obtained from ambulatory patients using wearable sensors, therefore containing realistic noise.

II. Methods

A. Database Used

For this study, ECG data were taken from a database which was collected as a part of feasibility study investigating the suitability of commercially available wearable sensors for clinical use [16]. The patient population consisted of adult patients recruited from the acute medical, acute surgical, and care-of-the-elderly wards at the John Radcliffe Hospital in Oxford who were able to move around unassisted. An ambulatory score ranging from 1 to 5 was recorded for each patient with one being "bed bound" and five being "able to mobilize independently." The range of ambulatory scores was 1–5 for the participating patients with a mean of 4.5 and a median of 5 [16]. For this study, we used data obtained using two different systems: the Equivital EQ-02 LifeMonitor (Hidalgo, Swavesey, U.K.) and the Dyna-Vision DVM012S (RS-TechMedic, Langedijk, The Netherlands). The Equivital EQ-02 was attached to a belt worn around the patient's chest. The ECG was sampled at 256 Hz with a 10-bit resolution. The Dyna-Vision DVM012S samples the ECG through at 100 Hz with 12-bit resolution using conventional ECG leads attached to adhesive wet gel electrodes.

ECG recordings were collected from 18 patients. Of these, 11 records (denoted $\mathrm{JR}_{\mathrm{D-ECG}}$) were captured using Dyna-Vision DVM012S monitors (123 h of recording) and seven (denoted $\mathrm{JR}_{\mathrm{E-ECG}}$) were captured using Equivital EQ-02 LifeMonitors (63 h of recording).

B. Development of SQI

1) Training Data and Labeling: To develop the SQI, we used a total of 1100 30-s segments of ECG. Seven-hundred segments were used for training the algorithm and 400 for testing it. The 700 segments used for training were taken from

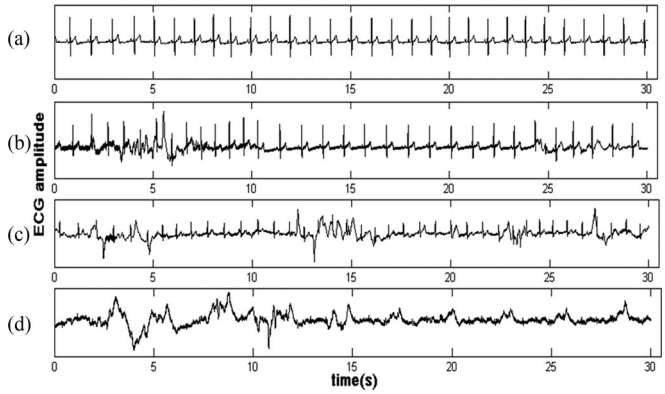

Fig. 1. Example ECG samples from the training set with associated labels: (a) "Acceptable." (b) "Acceptable" (borderline case). (c) "Unacceptable" (borderline case). (d) "Unacceptable".

the JR_{E-ECG} database, and comprised of 100 segments randomly drawn from each one of the seven subjects' recordings. The test set comprised of 200 segments randomly chosen from the JR_{E-ECG} database and 200 randomly chosen from the JR_{D-ECG} (total of 400 segments). We deliberately chose to train the classifier on data from a single sensor, in order to compare its performance on data from the same versus a different sensor, and assess its potential for generalizing across different sensors and manufacturers. All 1100 ECG segments were annotated by a biomedical engineer, expert on ECG analysis, based on the following rule:

"An ECG segment is labeled as 'acceptable' if a human expert can confidently derive a reliable HR from it, by counting the number of R-peaks over a fixed time interval. Otherwise it is labeled as 'unacceptable."'

More specifically, the annotator was instructed to allow for a single noisy segment in a sample, provided it was smaller in length than approximately two beat periods. This was based on the fact that samples were 30 s long and the presence of a single noisy segment, not longer than two beat periods, would still result to a HR within an acceptable error range.

In total, 47% of the segments were labeled as "acceptable" and 53% as "unacceptable." Examples of "acceptable" and "unacceptable" ECG segments are shown in Fig. 1.

2) SQI Algorithm: Fig. 2 shows a flowchart of the proposed SQI classifier.

The first step of the proposed SQI algorithm is to perform QRS detection on an ECG sample using the Hamilton and Tompkins algorithm [17] and apply a simple *feasibility rule*: the HR extrapolated from the 30-s sample must fall within a physiologically probable range of 40–180 beats per minute (bpm). If this condition is not satisfied, the ECG sample is immediately classified as "unacceptable." Otherwise, the HRV signal is extracted and analyzed using discrete wavelet decomposition. The entropy of the wavelet coefficients at the different wavelet subbands is then calculated. Vectors containing the entropies measured at the different wavelet decomposition levels are then taken as the feature vectors of each sample (the feature extraction process is shown in Fig. 3). In a final step, the feature vector is fed into a SVM classifier, previously trained using the labeled training data, and the sample is classified as "acceptable" or "unacceptable." The different steps in creating the SQI are explained in detail in the next section.

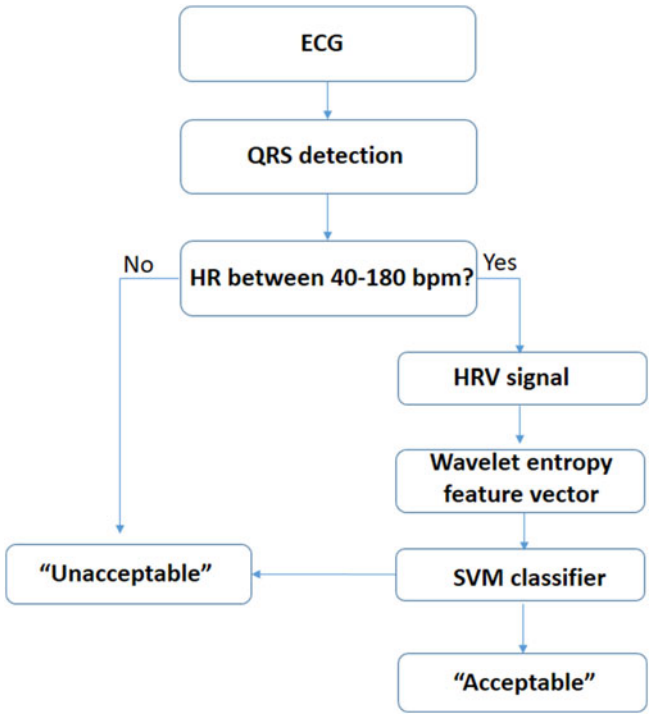

Fig. 2. Flowchart of the proposed SQI algorithm.

a) Extracting the HRV Signal: For obtaining the HRV signal, the R–R interval time series was created for each ECG sample by calculating the periods between consecutive R-peaks. The HRV signal was then obtained by applying spline interpolation at the recommended frequency of 4 Hz [11] to the R–R interval time series. Fig. 4 shows an example of a "clean" and a "noisy" ECG segment, their extracted HRV signals, and associated HRV spectra. As can be seen, the erroneous identification of additional R-peaks in the noisy sections of the signal result in the appearance of "dips" resulting in a distorted HRV signal which has power in frequencies higher than the ones commonly present in the HRV signal. (The appearance of additional erroneous R-peaks was the most common error we observed in the presence of noise).

b) Wavelet Transform and Wavelet Entropy: The wavelet transform describes signals in terms of coefficients and allows the representation of the temporal features of a signal at different resolutions. A signal $f(t)$ can be decomposed as [18]:

$$f(t) = \sum_j \sum_k d_{j,k} \psi_{j,k}(t) = \sum_j f_j(t) \tag{1}$$

where $j, k \in Z$ and $\psi(t)$ is the mother wavelet. To obtain the wavelet coefficients $d_{j,k}$ at different frequency bands, the mother wavelet is dilated and translated. The wavelet coefficients at each level j are then given by the inner product

$$d_{j,k} = \langle f(t),\ \psi_{j,k}(t) \rangle = \frac{1}{\sqrt{2^j}} \int f(t)\, \psi\left(2^{-j}t - k\right) dt \tag{2}$$

In practice, application of the discrete wavelet transform is done via successive application of a two-channel perfect recon-

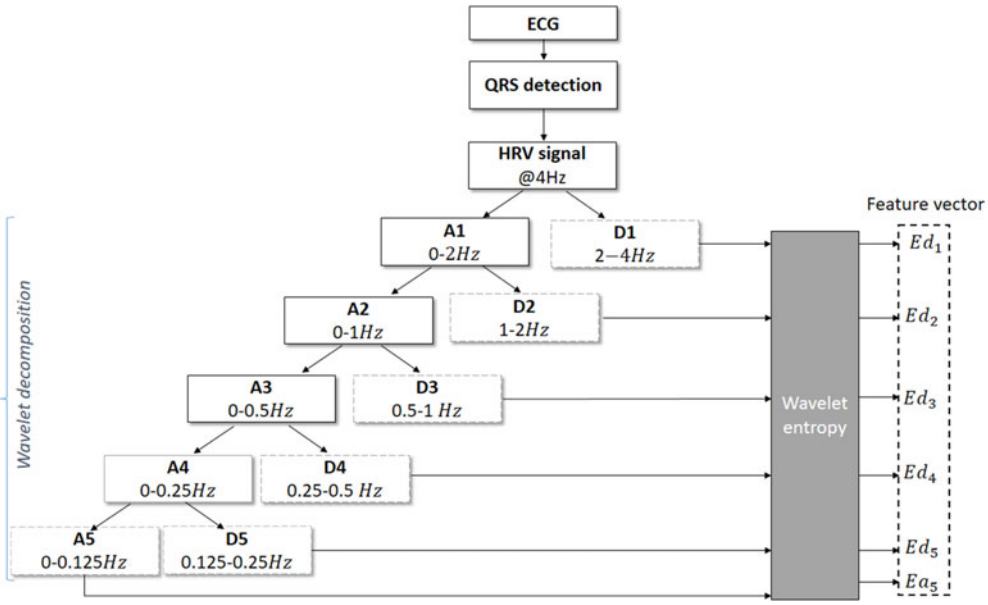

Fig. 3. Flowchart of the feature extraction algorithm.

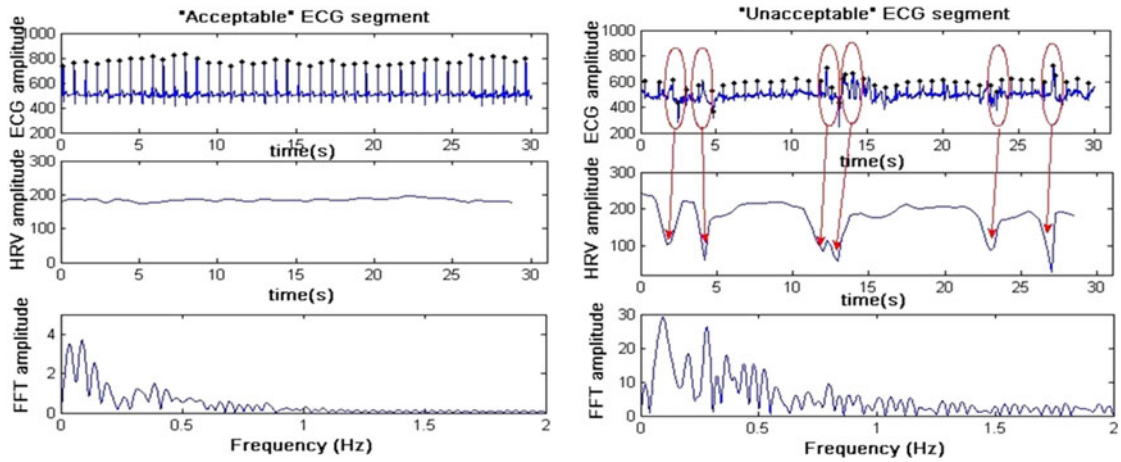

Fig. 4. "Acceptable" (left) and "unacceptable" (right) ECG samples and identified R-peaks (in black dots) (top) with extracted HRV signals (middle) and associated FFT spectrum (bottom) taken from the training set. As can be seen in the "unacceptable" case, the identification of false R-peaks caused by the presence of noise, results in "dips" which distort the HRV signal (indicated by arrows). These dips alter the spectrum of the extracted HRV signal. Please note that for the ease of interpretation, HRV spectrums were plotted in different scales. As expected, the spectrum of the distorted HRV signal on the right has energy in frequencies higher than the physiologically relevant upper limit of 0.4 Hz.

struction filter bank comprising of a low-pass and high-pass filter, followed by decimation by a factor of 2. The result of applying this filter bank is a set of *approximation* wavelet coefficients, resulting from the application of the low-pass filter, and a set of *detail* coefficients, resulting from the application of the high-pass filter, at different decomposition levels. The frequency bands associated with every decomposition level depend on the sampling frequency f_s of the signal and the number of decomposition levels depends on the characteristics of the studied signal. In our case, the extracted HRV signal is sampled at 4 Hz and the successive breakdown of the different frequency bands can be seen in Fig. 3 for a five-level wavelet decomposition. The rationale behind using a five-level wavelet decomposition was that for a window size of 30 s and under the assumption that at least four cycles are required for a reliable frequency-based estimate to be made, the minimum frequency that can be observed is 0.125 Hz (max period of 7.5 s); thus, any decomposition beyond five levels would not contain any valuable additional information. Once the wavelet coefficients are calculated at each level, the Shannon entropy can be calculated, giving a measure of the disorder at each decomposition level. Mathematically, the Shannon entropy of the details coefficients of f at level j can be expressed as [19]:

$$E_j = -\sum_k {d_{j,k}}^2 \log\left({d_{j,k}}^2\right). \tag{3}$$

Consequently, for the five-level wavelet decomposition, the result is a 6-D vector containing the wavelet entropy values measured in decomposition levels A_5, D_5, D_4, D_3, D_2, and D_1.

In addition to varying the levels of decomposition, we also tested different mother wavelets from the coiflet, daubechies,

and biorthogonal families in order to obtain the optimum one for the application at hand.

c) Feature Selection: Before feeding the feature vectors into the classifier, we performed standard two-sample t-tests on the training data in order to determine which features better differentiate between "acceptable" and "unacceptable" ECG segments. A two-sample t-test essentially tests the *null hypothesis* that two different sets of observations come from distributions of equal means [20]. By performing a t-test to the features extracted from the "acceptable" and "unacceptable" ECG segments, we calculate the p-value which is the probability of observing the given data assuming the null hypothesis is true. If the p-value is sufficiently small (typically smaller than 0.05 although different thresholds may be used for different applications), then the null hypothesis can be rejected, and we may conclude that the two groups of observations come from different distributions. While not a "hard" metric, the p-value can be taken as an *indicator* of feature separability since the smaller the value, the more likely it is that the two sets of observed data ("acceptable" and "unacceptable") come from different distributions.

d) Machine Learning Using SVM: The SVM algorithm is a powerful classifier which combines the simplicity of a linear process for separating high-dimensional feature data with the sometimes necessary complexity of nonlinear modeling of the input data in order to obtain the high-dimensional feature space [21].

Our application considers the commonly used two-class classifier formulation, in which N-dimensional patterns $\boldsymbol{x_i}$ and class labels y_i are trained in order to estimate a function $f : R^N \rightarrow \{\pm 1\}$ such that f will correctly classify new examples $(\boldsymbol{x}, y)$, that is $f(\boldsymbol{x}) = y$ for examples $(\boldsymbol{x}, y)$, which were generated from the same underlying probability distribution $P(\boldsymbol{x}, y)$ as the training data [21]. The SVM classifier is based on the class of hyperplanes

$$(w \cdot \boldsymbol{x}) + b = 0,\ w \in R^N, b \in R \tag{4}$$

where the decision function is given by

$$f(x) = \text{sign}\left((w \cdot x) + b\right). \tag{5}$$

The optimal hyperplane, defined as the one with the maximal margin of separation between the two classes, can be uniquely constructed by solving a constrained optimization problem whose solution $\boldsymbol{w}$ has an expansion $w = \sum_i v_i \boldsymbol{x_i}$ in terms of training patterns that lie on the margin, the so-called *support vectors.*

Because (4) and (5) depend only on dot products between patterns, it is possible to map the training data nonlinearly into a higher dimensional feature space F, via a map Φ, and construct the optimal separating hyperplane in F. This is accomplished by substituting $\Phi(\boldsymbol{x_i})$ for each pattern $\boldsymbol{x_i}$ by simple *kernels* k such that

$$k(\boldsymbol{x}, \boldsymbol{x_i}) := ((\Phi(\boldsymbol{x}) \cdot\ \Phi(\boldsymbol{x_i})). \tag{6}$$

The decision boundary then becomes

$$f(x) = \text{sign}\left(\sum_{i=1}^{l} v_i \cdot k(\boldsymbol{x}, \boldsymbol{x_i}) + b\right) \tag{7}$$

where the parameters v_i are computed as the solution of a quadratic programming problem.

For our application, we investigated the performance of quadratic, polynomial, and radial basis function (RBF) kernels, and obtained the best performance using a RBF kernel, which is given by

$$k(x, y) = \exp\left(-\|x - y\|^2\ \right)/2\sigma^2 \tag{8}$$

where σ is a scaling factor [21].

e) Training and Evaluation of the Proposed Approach: We ranked the features based on the p-value and presented them to the SVM classifier for training and testing, using six different combinations of features: model A included only the highest ranked feature, model B the two highest ranked features, model C the three highest ranked features, and so forth, until model F which included all wavelet entropy measurements. The best model was chosen as the one with the best classification accuracy compared to the manual annotations. Accuracy was determined by calculating the percentage of correctly classified samples (true positives plus true negatives) with respect to the total number of samples.

The true positive rate (sensitivity) and true negative rate (specificity) were also calculated in order to get a measure of the type of misclassifications occurring. The model with the maximum accuracy was then used to classify the two test sets. In assessing the performance of the proposed system on the test data, we calculated the sensitivity, specificity, and accuracy with respect to the manual annotations first using only the feasibility rule and then using the full system (feasibility rule and SVM classification), in order to put into context the contribution of the various steps of the proposed approach.

f) Measurement of HRs: In order to illustrate the effect of using the SQI on the reliability of HRs obtained from wearable sensors, we ran the SQI on the entire $\text{JR}_{\text{E-ECG}}$ and $\text{JR}_{\text{D-ECG}}$ databases, and calculated the HRs in beats per minute (bpm) in successive 30-s windows using

$$\text{HR} = \frac{60 \times f_s}{\text{RR}_{\text{median}}} \tag{9}$$

where f_s is the sampling rate and $\text{RR}_{\text{median}}$ is the median R–R interval in the 30 s segment. We then calculated the coefficient of variation of the HRs obtained from segments classified as "acceptable" or "unacceptable" for the entire database. The coefficient of variation is a measure of the variability of a measurement relative to its mean and is given by

$$r = \frac{\sigma}{\mu} \tag{10}$$

where σ is the standard deviation and μ is the sample mean.

III. Results

A. Feature Selection

Table I shows the six features ranked by the p-value. All features had a p-value less than 0.05 indicating that they all have strong discriminant power. In fact, most features had a p-value of almost 0. Interestingly, level D2 (1–2 Hz) had the lowest p-value since the HRV signals obtained from "acceptable" ECG seg-

TABLE I
WAVELET ENTROPY FEATURES RANKED BY THEIR *p*-VALUES

Wavelet Level	Frequency Band	*p*-value
D_2	1–2 Hz	<0.001
D_5	0.125–0.25 Hz	<0.001
D_4	0.25–0.5 Hz	<0.001
D_3	0.5–1 Hz	<0.001
D_1	2–4 Hz	<0.001
A_5	0–0.125 Hz	0.007

TABLE II
CLASSIFICATION PERFORMANCE OF TRAINING SET FOR DIFFERENT COMBINATIONS OF WAVELET ENTROPY FEATURES

Combination	Features	Accuracy (%)	Sensitivity (%)	Specificity (%)
A	D_2	95.1	91.4	99.1
B	D_2, D_5	95.4	91.7	99.4
C	$\boldsymbol{D_2, D_4, D_5}$	**96.0**	**92.3**	**99.1**
D	D_2, D_3, D_4, D_5	96.0	92.3	99.1
E	D_1, D_2, D_3, D_4, D_5	96.0	92.3	99.1
F	$D_1, D_2, D_3, D_4, D_5, A_5$	96.0	92.3	99.1

ments have hardly any energy in that frequency band, whereas the presence of HF noise in the "unacceptable" segments altered the spectra of the resulting HRV signals causing high-entropy values in the D2 band. The least discrimination power was observed in the approximation coefficients A5 (0–0.125 Hz). As explained in Section II-B, because of the relatively short duration of ECG segments used, no meaningful spectral information can be expected for frequencies below 0.125 Hz. As a result, any differences in the entropy values of the approximation coefficients at level A_5 do not have any physiological justification, but are mostly random.

B. Training and Model Selection

Because of the excellent discriminant power of our feature vector, we initially checked the performance of a simple linear classifier on the training data, using all six features, using the well-known linear discriminant analysis [22]. The accuracy obtained on the training data was 86% with a sensitivity of 99% and a specificity of 75%. While the results showed a fairly good linear separability of the data, the high number of false positives indicated that a more complex system was needed. We then proceeded to test the proposed SVM system with different combinations of features in order to obtain the best model. Table II shows the accuracy, sensitivity, and specificity values, with respect to the manual annotations, obtained on the training set using the different combinations of features using the "db12" wavelet basis. As is evident, the inclusion of additional features after model C caused no improvement in the classification accuracy. In fact, just using a single measure of entropy (model A) gave extremely good results and the improvement of adding additional features was marginal. Because of the small dimensionality of our feature space, it is possible to choose the marginally better model C as the optimum model of the classifier. We found little to no difference in the performance of the classifier when using different wavelet bases. We, therefore, used a "db12" basis and feature combination C to classify the two test sets.

TABLE III
PERFORMANCE OF CLASSIFIER ON TEST SETS USING ONLY FEASIBILITY RULE OR THE FULL SYSTEM

Database used	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
$JR_{E\text{-}ECG}$	Feasibility	75	77	74
	Full System	**96**	**94**	**98**
$JR_{D\text{-}ECG}$	Feasibility	66	59	73
	Full System	**94**	**91**	**97**

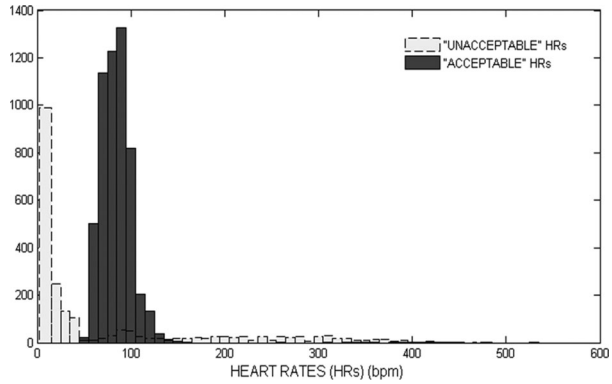

Fig 5. Histogram of HR values obtained from "acceptable" ECG segments (in dark gray) and "unacceptable" ECG segments (in dashed outline) over the entire database, as classified by the proposed SQI.

C. Classification Performance of SQI on Test Sets

Table III shows the accuracy, sensitivity, and specificity values, with respect to the manual annotations, on the two test sets. On the first test set, drawn from the JR_{E-ECG} database, i.e., the same database as the training data, using only the feasibility rule, the accuracy was 75%, and the sensitivity and specificity were 77% and 74%, respectively. Using the full system, the accuracy was 96% and the sensitivity and specificity were 94% and 98%, respectively.

On the second test set, drawn from the JR_{D-ECG} database, using only the feasibility rule, the accuracy was 66% and the sensitivity and specificity were 59% and 73%, respectively. Using the full system, the accuracy was 94% and the sensitivity and specificity were 91% and 97%, respectively. As is evident, the performance of the SQI on data obtained using the same sensor as the training set was slightly better; however, the performance of the system on data from a different sensor was also satisfactory.

D. Reliability of HR Measurements

Fig. 5 shows a histogram of the HR values obtained from "acceptable" and "unacceptable" segments of ECG, over the entire database, as classified by the proposed SQI. As is evident, the dispersion of "unacceptable" HR values is much greater than that of the "acceptable" HR values. The coefficient of variation of all the measured HR values was 0.74. For the HR values obtained from "unacceptable" segments, it was 1.42, while for the ones obtained from "acceptable" segments, it was 0.16. The

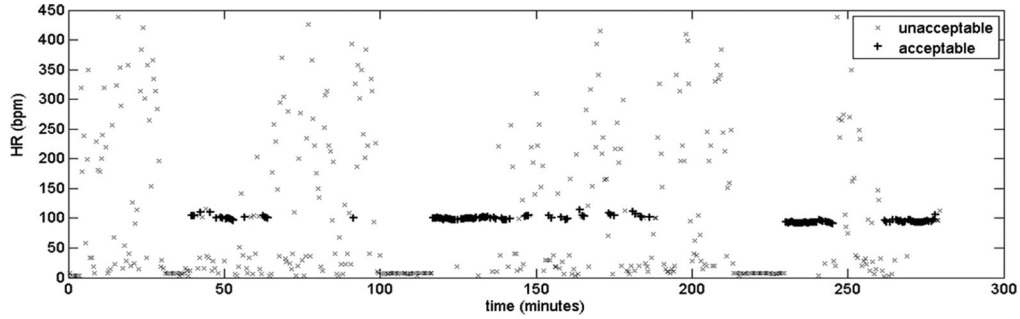

Fig. 6. HR values obtained from a 290-min continuous ambulatory ECG record. Measurements classified as "acceptable" are in black plus signs and ones classified as "unacceptable" are in gray crosses. It is evident that application of the SQI significantly decreases the dispersion of HR values.

effect of application of the SQI can be more easily seen by looking at the HR values obtained from a typical noisy continuous ECG recording. Fig. 6 shows a 290-min noisy continuous ECG record taken from the $\mathrm{JR}_{\mathrm{E-ECG}}$ database. HR values from ECG segments classified as "acceptable" are shown in black plus signs and HR values from segments classified as "unacceptable" are shown in gray crosses. It is evident that application of the SQI increases the reliability of the HR measurements since the amount of dispersion is reduced. For a healthy person or a patient in a stable condition, HR measurements are not anticipated to vary greatly in time. While it could be argued that HR values outside the physiologically viable range would be rejected by any monitoring system, a substantial proportion of HR values within the physiologically viable range would still be likely to cause a false alert (this can be attested by the performance metrics presented in Table III which show that using only the feasibility rule, we still have a large number of false negatives which would result in false alerts). Application of the SQI would likely suppress most of these false alerts.

IV. Discussion

In this paper, we presented an SQI for the ECG, intended to provide real-time assessment of the suitability of ECG signals for deriving reliable HR values. Our approach attempted to address some of the persisting issues of existing algorithms and systems: many are either tailored to a specific patient population, sensor, or manufacturer or require the presence of multiple channels of information, which is unrealistic in many clinical environments. Furthermore, most algorithms are trained on nonclinical data, often corrupted by artificial noise. Our proposed algorithm is designed for a single channel of ECG and is novel in that it is based on the analysis of the frequency content of the HRV signal, a signal rich in physiologically relevant information [12].

Our hypothesis was that the distribution of energy across the spectrum of the derived HRV signal would be different when comparing "clean" and "noisy" ECG segments. The reason is that errors in QRS detection, caused by the presence of noise, result in a distorted HRV signal with altered frequency content. Spectral analysis of the HRV was done by using discrete wavelet decomposition followed by the measurement of the entropy at each decomposition level, a measure of the distortion of the specific frequency subband. We then ranked the different frequency subbands based on the discriminant power of their wavelet entropy measurements, as indicated by the result of a two-sample t-test, and tested different combinations of features in order to find the optimal model of our classifier. Interestingly, the frequency subband with the strongest discriminant power was the 1–2 Hz frequency band (D2) which does not contain any physiologically meaningful information. Addition of the 0.125–0.25 Hz (D5) and 0.25–0.5 Hz (D4) frequency bands, containing the physiologically meaningful information of the HRV signal, caused only a marginal improvement to the performance of the classifier. An explanation for this could be the fact that the distortion of the HRV signal in the presence of HF noise significantly increases the wavelet entropy in the 1–2 Hz frequency band for "noisy" ECG samples compared to "clean" ones and that alone has a high enough discriminant power to differentiate most samples. Despite the fact that the 0.125–0.25 and 0.25–0.5 Hz frequency bands contain the physiologically meaningful information, both "clean" and "noisy" samples have strong wavelet entropy in those levels and their discriminant power is less significant.

The system was trained on data from ambulatory hospital patients, using wearable sensors. To investigate the generalization properties of the system, we evaluated the SQI on data obtained using the same sensor as the training data and on data obtained using a different sensor. Sensitivity and specificity of 94% and 98%, respectively, were obtained on data from the same sensor as the training data and 91% and 97%, respectively, on data from a different sensor. Finally, we investigated the effect of the SQI on the reliability of HR values obtained from ambulatory data by calculating HR values from continuous records of ambulatory ECG and calculating the coefficient of variation of the HRs with and without using the SQI. Application of the SQI reduced the coefficient of variation from 0.74 to 0.16, which is a more realistic value for subjects who are not in physiological distress during monitoring.

The proposed SQI is intrinsically linked to the QRS detector used both because of the application of the feasibility rule and in the wavelet entropy analysis after. Errors in the detection of R-peaks (either identifying erroneous R-peaks or missing actual R-peaks) result in a distorted HRV signal with altered frequency content and it is these exact alterations which the classifier is searching for in order to perform the classification. As a result, the proposed system relies on the use of the Hamilton and Tompkins algorithm [17]. However, an attractive property of our

approach is that since it is based only on the extraction of the HRV signal, provided the Hamilton and Tompkins algorithm is used, it can be extracted in the same way from any ECG signal, promising a technique with good generalization properties. A limitation of using the HRV signal is the requirement of a relatively high sampling rate of the ECG signal. The standard recommendation is for the ECG to be sampled at 250–500 Hz although, a minimum of 100 Hz would also be acceptable provided that an algorithm of interpolation (e.g., parabolic) (such as the Hamilton and Tompkins algorithm we used [17]) is used to detect the R-peak [11]. While these standards are defined for optimizing the physiological interpretation of the HRV characteristics, which is outside the scope of our proposed approach, given that our algorithm is based on comparing the distribution of energy across different frequency bands of the HRV signal, the minimum requirement of 100 Hz would need to be satisfied. The method is, thus, limited to systems with a minimum sampling rate of 100 Hz. Additionally to balance the limitation of real-time implementation, we picked segments of 30-s duration which means that LF and VLF frequency bands were not considered. A longer duration of signal would likely improve the performance of the classifier by identifying further differences between clean and noisy signals in lower frequency bands (with strong physiological content) but would compromise the possibility of usage of the SQI as part of real-time monitoring systems.

The binary classification scheme we propose was shown to significantly improve the reliability of HR measurements. An undesirable result, however, is that for extended periods of time, no reliable HR measurement is obtained (this can be clearly observed on Fig. 6). While this is a reflection of the quality of the data, and may be temporarily resolved via a sample-and-hold scheme, it may be the case that in certain clinical scenarios, it would be preferable to obtain "moderately erroneous" HR values at a higher frequency than no HR values at all for extended periods of time. This could be implemented by a more flexible fuzzy classification scheme which would assign different reliability indices to HR measurements and could be adapted by the users depending on their specific needs.

V. Conclusion

The system described in this paper has been shown to correctly classify ECG segments from ambulatory sensors as "acceptable" and "unacceptable" and, as a consequence, significantly increase the reliability of HR measurements obtained. It has shown potential for generalizing to different sensors and systems. Finally, in contrast to many recently proposed systems requiring the presence of multiple signals or multiple channels of the same signal, our system requires a single channel of ECG making it very promising for inclusion in applications using wearable sensors and in the most basic clinical environments.

Acknowledgment

The authors would like to thank T. Bonnici and D. Vallance for performing the data collection for the database used in this paper.

References

[1] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Valance, and L. Tarassenko, "Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 3, pp. 832–838, May 2015.

[2] A. E. W. Johnson, M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care," *Proc. IEEE*, vol. 104, no. 2, pp. 444–466, Feb. 2016.

[3] M. Buist, S. Bernard, T. V. Nguyen, G. Moore, and J. Anderson, "Association between clinically abnormal observations and subsequent in-hospital mortality: A prospective study," *Resuscitation*, vol. 62, pp. 137–141, 2004.

[4] L.W. Andersen *et al.* "The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest," *Resuscitation*, vol. 98, pp. 112–117, 2016.

[5] S. Nizami, J. R. Green, and C. McGregor, "Implementation of artefact detection in critical care: A methodological review," *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 127–142, Jan. 2013.

[6] Q. Li and G. Clifford, "Signal quality and data fusion for false alarm reduction in the intensive care unit," *J. Electrocardiol.*, vol. 45, pp. 596–603, 2012.

[7] 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, Jan. 2008.

[8] C. Orphanidou, T. Bonnici, D. Vallance, A. Darrell, P. Charlton, and L. Tarassenko, "A method for assessing the reliability of heart rates obtained from ambulatory ECG," in *Proc. IEEE Int. Conf. Bioinformat. Bioeng.*, 2012, pp. 193–196.

[9] G. E. Billman, "Heart rate variability—A historical perspective," *Front. Physiol.*, vol. 2, 2011, Art. no. 86.

[10] S. J Pieper and S. C. Hammill, "Heart rate variability: Technique and investigational applications in cardiovascular medicine," *Mayo Clin. Proc.*, vol. 70, no. 10, pp. 955–64, 1995.

[11] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, "Heart rate variability Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, pp. 354–381, 1996.

[12] C. A. Rickards, K. L. Ryan, and V. A. Convertino, "Characterization of common measures of heart period variability in healthy human subjects: Implications for patient monitoring," *J. Clin. Monit. Comput.*, vol. 24, pp. 61–70, 2010.

[13] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2007, pp. 4656–4659.

[14] Y. İşler and M. Kuntalp, "Combining classical HRV indices with wavelet entropy measures improves to performance in diagnosing congestive heart failure," *Comput. Biol. Med.*, vol. 37, no. 10, pp. 1502–1510, 2007.

[15] J. F Ramirez-Villegas, E. Lam-Espinosa, D. F. Ramirez- Moreno, P. C. Calvo-Echeverry, and W. Agredo-Rodriguez, "Heart rate variability dynamics for the prognosis of cardiovascular risk," *PLoS One*, vol. 6, no. 2, p. e17060, 2011.

[16] T. Bonnici, C. Orphanidou, D. Vallance, A. Darrell, and L. Tarassenko, "Testing of wearable monitors in a real-world hospital environment: What lessons can be learnt?," in *Proc. 9th Int. Conf. Wearable Implantable Body Sens. Netw.*, 2012, pp. 79–84.

[17] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Trans. Biomed. Eng.*, vol. 33, no. 12, pp. 1157–1165, Dec. 1986.

[18] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[19] C. E. Shannon, "A mathematical theory of communication," *Bell. Syst. Technol. J.*, vol. 27, pp. 379–423, Jul. 1948.

[20] H. A. David and J. L. Gunnink, "The paired t test under artificial pairing," *Amer. Statistician*, vol. 51, no. 1, pp. 9–12, 1997.

[21] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[22] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

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