Assessing the signal quality of electrocardiograms from varied acquisition sources: A generic machine learning pipeline for model generation

Adnan Albaba [a,b,1], Neide Simões-Capela [a,b,1], Yuyang Wang [c], Richard C. Hendriks [c], Walter De Raedt [b,*], Chris Van Hoof [a,b]

[a] Electrical Engineering Department of University of Leuven, Belgium
[b] Connected Health Solutions Group at Imec-Leuven, Belgium
[c] Electrical Engineering Department of University of Delft, the Netherlands

ARTICLE INFO

ABSTRACT

Background and objective: Long-term electrocardiogram monitoring comes at the expense of signal quality. During unconstrained movements, the electrocardiogram is often corrupted by motion artefacts, which can lead to inaccurate physiological information. In this situation, automated quality assessment methods are useful to increase the reliability of the measurements. A generic machine learning pipeline that generates classification models for electrocardiogram quality assessment is presented in this article. The presented pipeline is tested on signals from varied acquisition sources, towards selecting segments that can be used for heart rate analysis in lifestyle applications.
Methods: Electrocardiogram recordings from traditional, wearable and ubiquitous devices, are segmented in 10 s windows and manually labeled by experienced researchers into two quality classes. To capture the electrocardiogram dynamics, a comprehensive set of 43 features is extracted from each segment, based on the time-domain signal, its Fast Fourier Transform, the Autocorrelation function and the Stationary Wavelet Transform. To select the most relevant features for each acquisition source we employ both a customized hybrid approach and the state-of-the-art Neighborhood Component Analysis method and compare them. Support Vector Machines (SVM), Decision Trees, K-Nearest-Neighbors and supervised ensemble methods are tested as possible binary classifiers.
Results: The results for the best performing models on traditional, wearable and ubiquitous electrocardiogram datasets are, respectively: balanced-accuracy: 89%, F1-score: 93% with the Fine Gaussian SVM model and 10 features; balanced-accuracy: 93%, F1-score: 93% with the Fine Gaussian SVM model and 11 features; balanced-accuracy: 95%, F1-score: 86%, with the Fine Gaussian SVM model and 8 features.
Conclusions: According to the results, our generic pipeline can generate classification models tailored to individual acquisition sources, provided that a standard Lead I or Lead II is available. Such models accurately establish whether the electrocardiogram quality is good or bad for heart rate analysis. Furthermore, removing bad quality segments decreases errors in heart rate calculation.

1. Introduction

The electrocardiography (ECG) captures the electrical activity of the heart through a set of electrodes placed on the body surface. The electrode placement relative to the heart (i.e. lead configuration) defines the expected morphology of the ECG waveform. Due to its stability and well-known properties (i.e. normal patterns and pathological deviations), the

ECG is widely used for health monitoring. The ECG has been collected in portable devices ever since the Holter system was developed in the 1940's [1,2] and allows for long-term ECG monitoring. This system is widely used in intensive care units, cardiology wards and follow-up after disease onset. Recent technological advancements provide the capability for collecting ECG signals during longer periods, up to months [3]. Moreover, non-contact ECG devices (using capacitive coupled
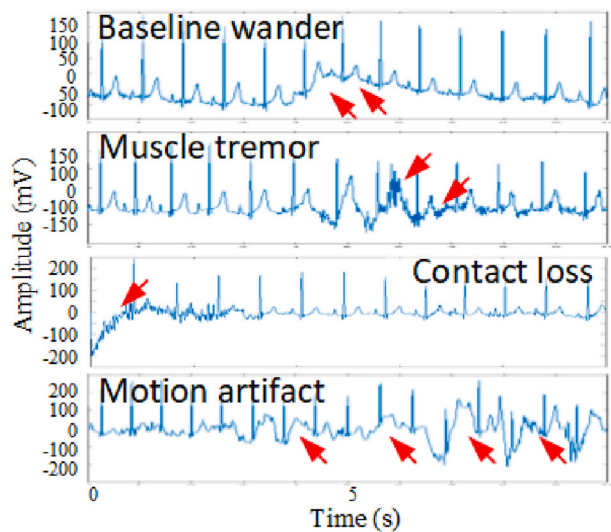

Fig. 1. Common ECG signal artefacts: baseline wander; muscle tremor; contact loss of the sensors with the skin and motion artefacts.

electrodes) and radar-based heart rate (HR) monitors are now used for ubiquitous sensing [4]. Together, these technologies enable early anomaly detection and life-style monitoring with little hassle for the user.

However, long-term wearable ECG monitoring comes at the expense of signal quality. ECG signals are usually contaminated with: 1) power line interference, 2) muscle activity (tremor), 3) baseline wander (respiration), and 4) motion-related artefacts (cf. Fig. 1). Such added noise results in deformed ECG morphology and increases the possibility of wrong medical conclusions if left untreated. Artefacts that are out of the frequency range of interest for ECG analysis, can be eliminated with adequate filtering. On the other hand, in-band noise, that overlaps with the ECG spectral components, can be difficult to address, especially when movements are unconstrained as in wearable and ubiquitous setups. While visual inspection can be performed to identify and remove corrupted segments, manual quality control is not scalable for long records. Therefore, automated algorithms are needed to aid for long-term ECG quality assessment.

The ECG quality requirements depend on the objective of the analysis. For instance, in the case of morphology analysis aimed at detecting malfunctions in specific parts of the heart, the waveform of the whole ECG cycle should be intact, and several leads might be required for better diagnosis. On the other hand, HR and HR variability (HRV)

Table 1
Summary of selected publications related to ECG signal quality assessment.

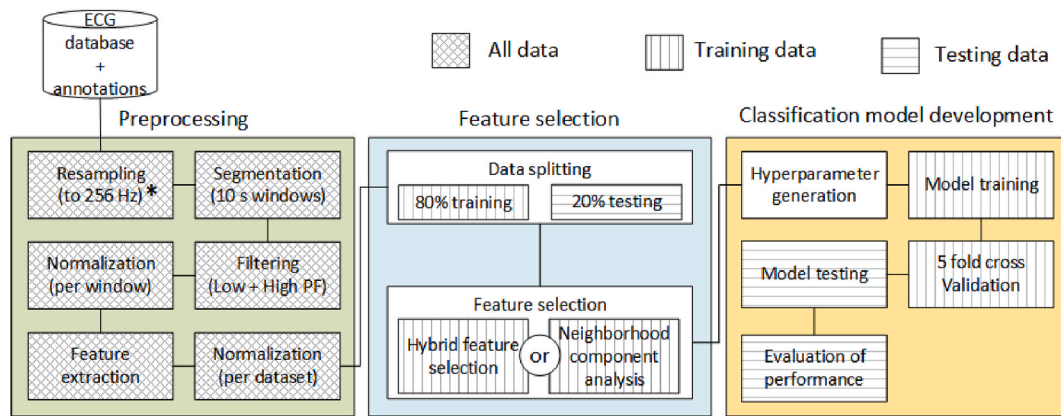| Year	Title	Datasets [name, description, acquisition (traditional or holter/wearable/ubiquitous)]	Processing methodology [approach (heuristic/ machine learning), description]	Results [best performance metrics]
2019	Artefact detection and quality assessment of ambulatory ECG signals [14]	• SWEET (wearable) • PhysioNet/CinC Challenge 2017 (wearable)	ML Uses the posterior probabilities of RUSBoost model for detecting artefacts in ECG segments.	Sensitivity = 96.6% Specificity = 84.8% AUC = 97.0% (Balanced Acc = 90.7%)
2019	Signal Quality Assessment and Lightweight QRS Detection for Wearable ECG SmartVest System [16]	• PhysioNet/CinC Challenge 2014 (wearable)	ML 7 signal quality indices are used along with an SVM classifier.	Accuracy = 96.4%
2018	SQI Quality Evaluation Mechanism of Single-Lead ECG Signal Based on Simple Heuristic Fusion and Fuzzy Comprehensive Evaluation [9]	• Physionet/Cinc Challenge 2011 (traditional) • Physionet/Cinc Challenge 2017 (wearable)	Heuristic Fuzzy comprehensive evaluation using four quality indexes: R peak detection match, QRS power spectrum distribution, kurtosis, and relative power of the baseline.	Accuracy = 94.7% Sensitivity = 90.3% Specificity = 93.0% (Balanced Acc = 91.7%)
2018	An automated ECG signal quality assessment method for unsupervised diagnostic systems [19]	• Allengers virgo EEG kit (traditional) • Bioradio wireless physiology monitor (wearable) • MIT-BIH arrhythmia (traditional)	ML Unsupervised learning model based on wavelets along with statistical features computed on low frequency and high frequency spectral bands.	Accuracy = 95.7% Sensitivity = 97.5% Specificity = 92.0% (Balanced Acc = 94.7%)
2017	Quality Assessment of Ambulatory ECG Using Wavelet Entropy of the HRV Signal [17]	• Recordings obtained from ambulatory patients (wearable)	ML Wavelet entropy of RR intervals along with an SVM classifier.	Accuracy = 95.0% Sensitivity = 92.5% Specificity = 97.5% (Balanced Acc = 95.0%)
2015	Electrocardiogram signal quality assessment using an artificially reconstructed target lead [13]	• Physionet/Cinc Challenge 2011 (traditional) • High resolution ECG from DAY hospital [20] (traditional)	Heuristic Energy-concavity index: energy and concavity of ECG. Correlation based quality index: correlation between ECG leads generated by a neural network.	Accuracy = 93.6% Sensitivity = 92.4% PPV = 94.8%
2015	Quality estimation of the ECG using cross-correlation among leads [18]	• Physionet/Cinc Challenge 2011 (traditional)	ML Cross-covariance matrix among different leads. Classification based on the eigenvalues.	Accuracy = 92.7% Specificity = 95.5% AUC = 91.0%
2015	Signal-Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring [10]	• Ambulant hospital database (wearable)	Heuristic Template matching: correlation between QRS complex and QRS template (based on averaging all detected QRS complexes on a window).	Sensitivity = 94.0% Specificity = 97.0% (Balanced Acc = 95.5%)
2012	Automatic motion and noise artefact detection in Holter ECG data using empirical mode decomposition and statistical approaches [11]	• 5-lead ECG recordings (traditional)	Heuristic Mean value, standard deviation, entropy of normalized first-order intrinsic mode function of Raw ECG.	Accuracy = 96.6% Specificity = 94.7% Sensitivity = 96.6% (Balanced Acc = 95.6%)
2012	QRS detection-based ECG quality assessment [12]	• Physionet/Cinc Challenge 2011 (traditional)	Heuristic Empty lead detection, spike detection, number of lead crossing points, and stability of the QRS detection.	Accuracy = 91.6%

Fig. 2. Proposed ML pipeline for the quality evaluation of ECG signals from diverse acquisition sources. * Resampling with anti-aliasing filter.

analysis mainly requires clean R-peaks and a single lead recording will usually be enough. Our work is aimed at tackling quality issues pertaining HR and HRV analysis for lifestyle and wellbeing applications.

Approaches to address the artefact problem in previous literature can be categorized into three main classes: 1) denoising the whole signal using various signal processing techniques [5], 2) identifying low quality segments and excluding them from further analysis, or 3) combining the previous two approaches in order to identify noisy ECG segments and denoise them [6]. In this work, we focus on the second approach.

The ECG signal quality problem has been broadly explored in the literature, especially after the Physionet/Computing in Cardiology (CinC) [7,8] challenge in 2011. Table 1 provides a summary of selected publications in the field of ECG signal quality assessment. All solutions concern binary quality assessment, i.e. good or bad quality. The proposed solutions either utilize statistical and signal processing techniques to characterize the signal, combined with heuristic hard thresholds [9–13], or use diverse signal descriptors (i.e. signal features) coupled with machine learning (ML) pipelines to provide a holistic signal quality score [14–18]. It can be noted that in most cases, only one ECG dataset was used for developing a solution. Moreover, none of these works explored ubiquitous ECG signals.

In this work, we study the possibility of a unified framework to cater for varied signal acquisition sources (cf. Fig. 2), by proposing a ML pipeline that can generate binary classification models capable of accurately classifying the signal quality of ECG segments acquired from traditional, wearable, and ubiquitous devices. The presented pipeline consists of a generic feature space to characterize the ECG quality, along with two alternative methods for feature selection and an array of classification algorithms that are trained on the selected features towards finding the best performing solution. All resulting models attained an acceptable performance, with balanced accuracies (BAcc)[2] [21] above 87%. MATLAB R2018a functions and toolboxes were used throughout the work.

The remaining sections of this paper are structured as follows: Section 2 presents the methods and summarizes the main steps of the proposed pipeline; Section 3 includes the results from the feature selection, training and validation of classification algorithms, and the test of each model on new data from three different datasets. A use case for the quality indicators is also presented. Results are discussed in Section 4, and conclusions are drawn upon Section 5.

---

[2] Balanced accuracy: a performance metric for unbalanced datasets, that is equal to (sensitivity + specificity)/2.

Table 2
ECG Datasets: protocol, acquisition setup, data contents and quality classes.

Dataset	Protocol	Acquisition setup	Original data contents	Quality classes
wDS	12 subjects were asked to perform everyday tasks under controlled and uncontrolled conditions.	Wearable ECG from Health Patch (imec-Leuven, Belgium); lead II; 256 Hz.	9693 segments, each with 10 s	Label 1: good quality; clear QRS complex. (4724 segments) Label 0: bad quality; unclear QRS complex due to noise. (4969 segments)
ccDS	109 subjects at diverse scenarios such like sitting on an office chair, car seat while driving, and lying over bed mattress.	Capacitive coupled ECG sensors (ccECG) from the work of Miller [22] UnoVis dataset [23]; 512 Hz.	9910 ECG segments, each with 15 s	Label 1: good quality; clear ccECG, which can be used be used for HR estimations. (1937 segments) Label 0: bad quality; no ccECG or ccECG with noise, which affects HR estimations. (7973 segments)
phyDS	Standard 12-lead ECG recordings, acquired by medical professionals with varied degrees of experience and under diverse acquisition conditions.	Standard 12-lead ECG from the open source database of 2011 Physionet challenge [8]; lead II; 500 Hz.	1307 of ECG segments, each with 10 s	Label 1: good quality; QRS complexes are identifiable. (1017 segments) Label 0: bad quality; QRS complexes are unidentifiable. (290 segments)

2. Methods

2.1. Datasets and annotations

Three different ECG datasets are employed in this work (cf. Table 2): Wearable ECG (wDS), Non-contact ECG (ccDS), and Physionet ECG (phyDS). To establish a quality ground truth, the ECG was manually labeled by several annotators with experience in ECG processing. To achieve a fine-grained quality assessment, the ECG time series was divided into small segments of 10 or 15 s, and each was individually labeled. Table 2 provides a summary of the ECG datasets used in this work. It should be noticed that phyDS includes pathological patterns (e.

g. atrial fibrillation, atrial fluttering), in this case abnormal beats that maintain a detectable R-peak are still labeled as good quality segments.

2.2. Preprocessing

The preprocessing step consists of resampling, segmentation, filtering, and normalization of the ECG signals. A frequency of 256 Hz is considered an acceptable acquisition frequency when targeting HR/HRV applications [24,25] and is also the lowest sampling frequency among datasets. In the case of ccDS and phyDS, data was down sampled using the MATLAB function *decimate*, which incorporates an anti-aliasing filtering step based on a Chebyshev type 1 low pass filter of order 8. The segmentation used for quality annotation is maintained. Though, to exclude the influence of segment size from the analysis, in the case of ccDS, the last 5 s of each segment are discarded. Each resulting 10-s segment is independently processed in the following steps. The filtering procedure consists of a zero-phase 3rd order Butterworth band-pass filter, with 0.5 Hz and 40 Hz cut-off frequencies (MATLAB function *zerophase*). This allows to detrend (low pass) and denoise (high pass) the ECG waveform, while preserving the information in the frequency band of interest. Lastly, the amplitudes of the ECG signals are normalized in the minimum-maximum range within each segment according to Eq. (1), where y_{norm} represents the normalized segment, y represents the original segment, and t is the number of samples.

$$y_{norm(t)} = \frac{y(t) - \min(y)}{\max(y) - \min(y)} \ , \ t \in \{0, .., 10\} \ s \tag{1}$$

2.3. Training, validation and testing data splits

Each of the original datasets is randomly split (cf. Fig. 2) into training/validation (80% of the segments) and testing (20% of the segments) subsets. The training/validation subset is employed in finding the features that are most relevant for the classification problem, and in training and validating a set of models using 5-fold cross-validation [26]. The best performing classification models are then tested on the testing subset (not used in previous steps), to attain the final unbiased performance metrics.

2.4. Performance evaluation

The performance of the classification models, is reported in terms of statistical metrics commonly used for binary classification problems, namely: accuracy (Acc), sensitivity (Sen), specificity (Spe), precision (Prec) and F1-score (F1) [27]. Since two of the datasets (i.e. ccDS and phyDS) are unbalanced in terms of the number of good and bad segments, the BAcc metric [21], expressed in Eq. (2), is employed for choosing the best performing models.

$$BAcc = \frac{Sen + Spe}{2} \tag{2}$$

The BAcc includes information about the sensitivity and specificity on a single value, hence, providing a better understanding of the classification errors than Acc.

2.5. Multidimensional feature space

A diverse set of descriptive features is put together to describe the signal quality of ECG segments. The ECG characteristics are investigated using the properties of the time-domain signal, the frequency-domain signal from the Discrete Fourier Transform (DFT), the multiresolution frequency-domain signal from the Stationary Wavelet Transform (SWT), and the time-series behavior as depicted by the Autocorrelation Function (ACF). After extraction, the features are normalized in the min-max range within the whole dataset, to reduce the computation time in the following steps.
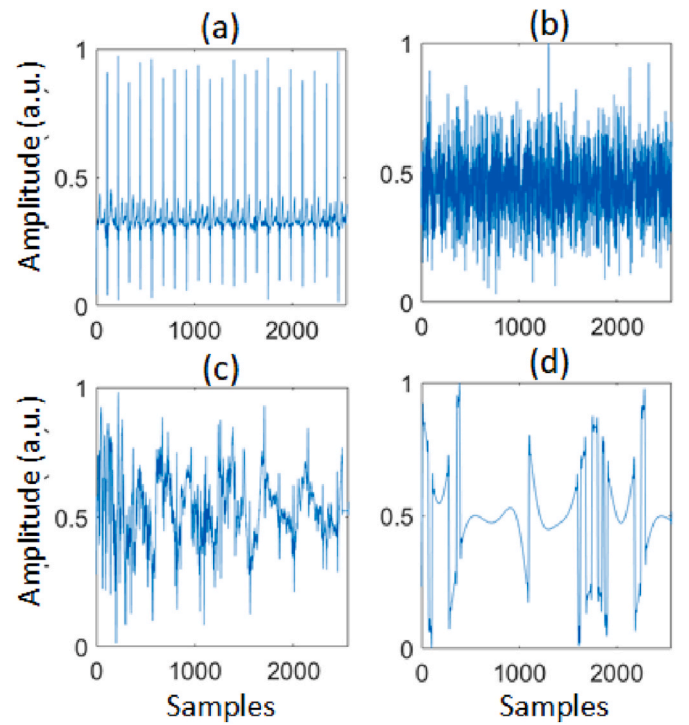

Fig. 3. ECG segments (10 s) in the time domain. a) good quality segment from wDS; b, c, and d) bad quality segments from wDS, phyDS, and ccDS, respectively.
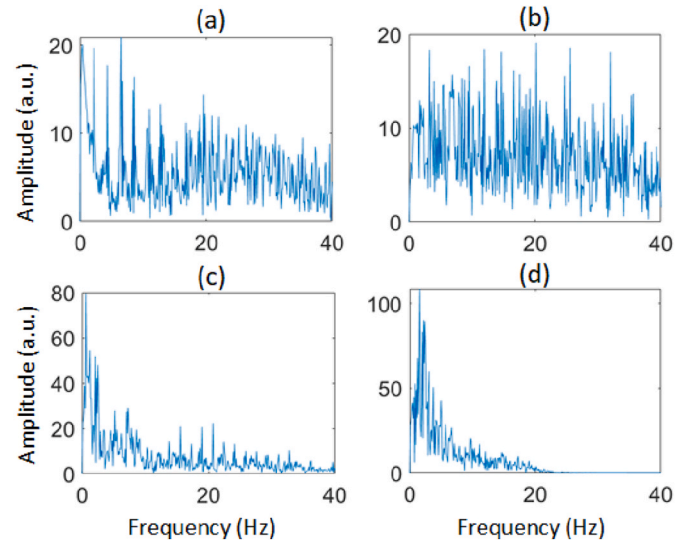

Fig. 4. ECG segments after applying the FFT. These frequency-domain depictions correspond to the time-domain segments in Fig. 3 a) spectrum of a good quality segment from wDS; b, c, and d) spectrum of bad quality segments from wDS, phyDS, and ccDS, respectively.

Time domain signal analysis: The morphological differences between a good and a bad quality ECG segment are depicted in Fig. 3. It is noticeable that different ECG acquisition setups present different artefacts. To characterize these behaviours, 6 features are derived from the ECG segments: 1) mean (mean_raw); 2) standard deviation (std_raw); 3) kurtosis (kurt_raw); 4) skewness (skew_raw); 5) median absolute deviation (mad_raw.); 6) approximate entropy (apen_raw).

Discrete Fourier Transform (DFT) analysis [28]**:** The DFT converts the signal form the time domain to the frequency domain, revealing the spectrum of its frequency components. The mean HR can be derived
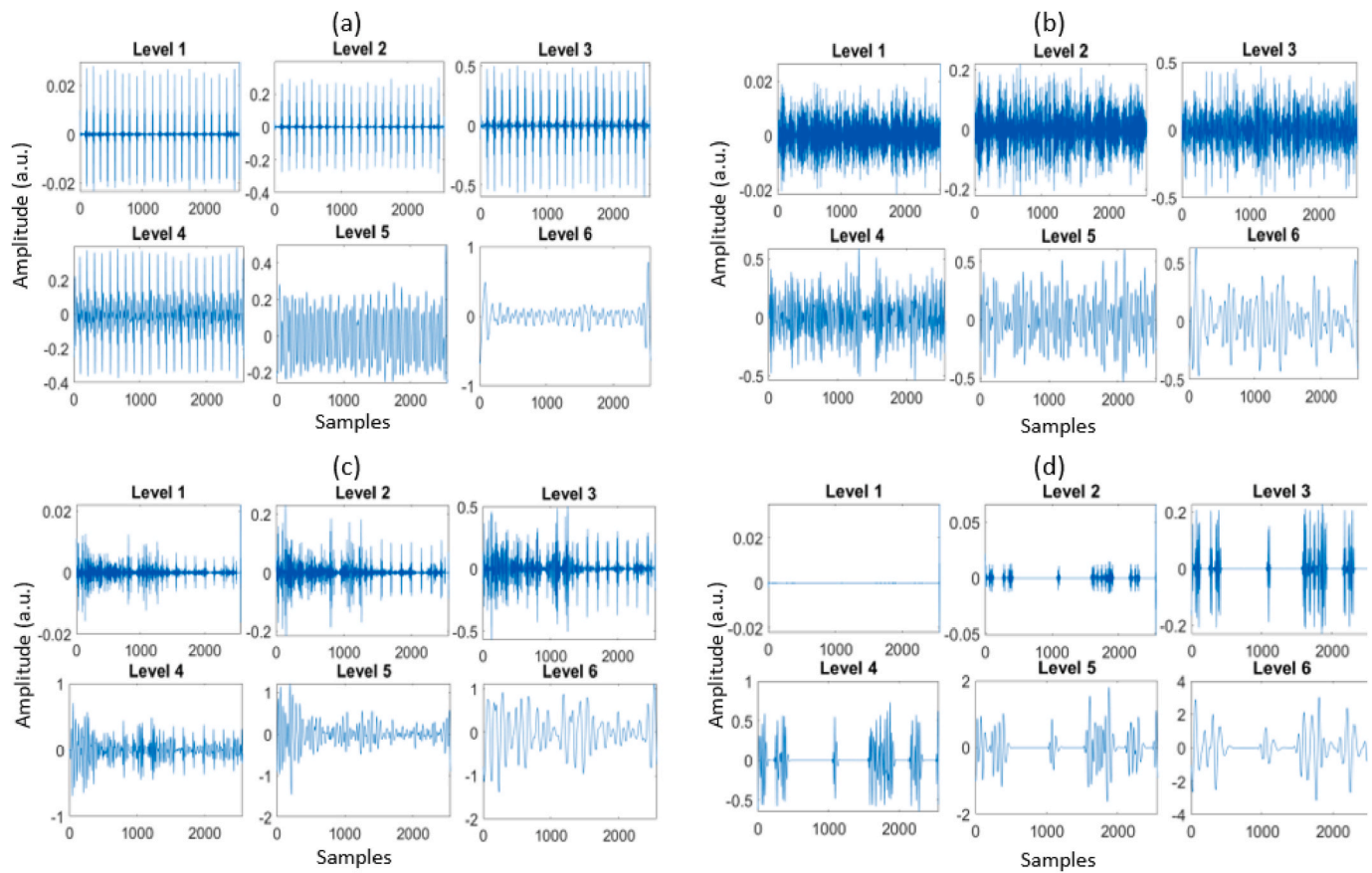
Fig. 5. SWT decomposition showing the wavelet detail coefficients up to level 6. These four decompositions correspond to the time-domain segments in Fig. 3 a) 6 levels of decomposition of a good quality segment from wDS; b, c, and d) 6 levels of decomposition of a bad quality segment from wDS, phyDS, and ccDS, respectively.

from the frequency of the highest peak in the spectrum of a clean ECG.

In this work, the Fast Fourier Transform (FFT) algorithm [29] is applied on each pre-processed ECG segment using the MATLAB function *fft*. The output amplitude is normalized and converted to a single sided spectrum and truncated in the range 0–40 Hz. As illustrated in Fig. 4, when artefacts are introduced the spectrum of the ECG segment is affected in diverse ways. To quantify the different spectral components, 6 features are extracted: 1) mean (mean_fft); 2) maximum value (max_fft); 3) standard deviation (std_fft); 4) kurtosis (kurt_fft); 5) skewness (skew_fft); 6) approximate entropy (ent_fft).

Stationary Wavelet Transform (SWT) analysis [30–32]: the SWT can be pictured as a filter bank, where the input signal is passed repeatedly through a low pass filter (LPF) (i.e. the scaling filter) and a high pass filter (HPF) (i.e. the wavelet filter), with each filtering step constituting a level of decomposition. The filter impulse response is modulated by the mother wavelet function, that has a specific shape. It is worth mentioning that in the SWT, the wavelet function and the number of decomposition levels are application dependent. The output of the HPF (i.e. the detail coefficient) is in the range [$Fmax/2$, $Fmax$], where $Fmax$ is the maximum frequency of interest (i.e. the Nyquist frequency), and the output of the LPF is (i.e. the approximation coefficient) is in the range [0, $Fmax/2$]. Since half of the signal frequencies are removed, according to Nyquist theorem, half of the samples are redundant and can be discarded by down sampling the signal by a factor of 2. However, in the SWT algorithm, no samples are discarded [33], hence it is shift invariant. At the next decomposition level, the previously obtained approximation coefficients can be further decomposed by applying the LPF and the HPF again, each having half the cut-off frequency. The main advantage of the SWT over DFT is that it captures, in the same output, both the frequency and time domain
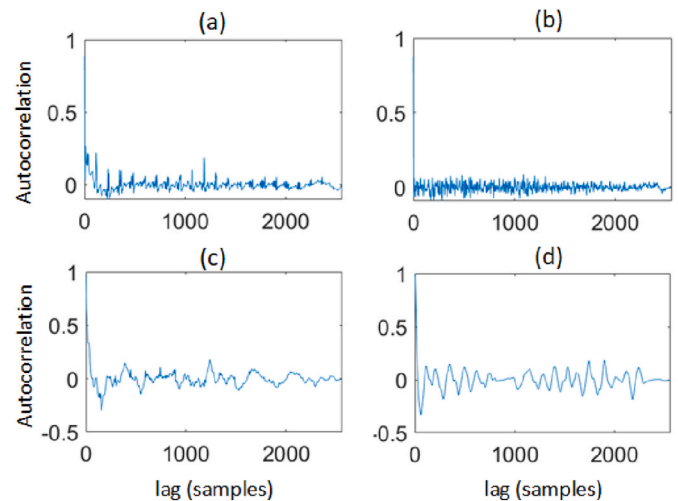

Fig. 6. ACF of the ECG segments with a maximum lag of 2560 samples. These ACF depictions correspond to the time-domain segments in Fig. 3 a) ACF of a good quality segment from wDS; b, c, and d) ACF of bad quality segments from wDS, phyDS, and ccDS, respectively.

information. On a clean ECG, each level of decomposition will depict the ECG waves (PQRST) on that frequency range.

The *Daubechies* wavelets are recurrent in ECG processing literature because of their structural similarity to the QRS complex [34,35] and its ability to capture changes in signals [36]. Therefore, the *Daubechies* wavelet of order 6 is used in this work. The SWT is performed using the

MATLAB function *swt* (with 'db6' wavelet and 6 levels decomposition), covering a frequency range from 2 to 128 Hz. Fig. 5 demonstrates how different acquisition conditions result in different SWT outputs at different frequency ranges. To describe each wavelet level, 4 features are extracted: 1) mean of the absolute value (mean_swt_level); 2) standard deviation (std_swt_level); 3) median absolute deviation (mad_swt_level); 4) approximate entropy (apen_swt_level).

Autocorrelation Function (ACF) analysis [37]: The ACF measures the self-similarity of a signal, and is used in finding periodic patterns. This is done by calculating the correlation between the original signal and a shifted copy of itself for a set of time lags. The ACF of a normal ECG will present a correlation degradation as the time lag increase, related to the small, but continuous tuning of the cardiac activity overtime.

Previous ECG literature used a maximum time lag of 250 ms to measure the beat-to-beat autocorrelation [14]. In this work, we set the maximum time lag to be the segment length (MATLAB function *acf* [38], with a maximum time lag of 2560 samples), see Fig. 6. This enables the evaluation of important ACF characteristic in an averaged form. To capture relevant information, 7 features are derived from the ACF: 1) amplitude (Fpmax) and 2) location (Fploc) of the first local maximum; 3) amplitude (fm_amp) and 4) location (fm_loc) of the first local minimum; 5) abscissa of the first zero-crossing (Fzloc); 6) zero-crossing rate (Zcr); 7) standard deviation of zero-point to zero-point intervals (Zxistd).

2.6. Feature selection

In this work, a customized hybrid method and the Neighborhood Component Analysis (NCA) are employed in selecting the most relevant features, and their results are compared. The subsets of features resulting from each feature selection method, are used to train/validate and test the predictive models.

2.6.1. Hybrid method

In this approach three methods with different characteristics are used consecutively to select the best feature subset. The first two are filter methods (i.e. features that are not relevant according to some metric are filtered out of the subset), and the last one is a wrapper method (i.e. a classifier is trained with different subsets of features, and the best performing subset is chosen). Such combination minimizes the selection bias, presented when using each of the methods alone.

Wilcoxon signed-rank test (WSRt): a non-parametric test that statistically verifies whether two samples come from different populations. The test tries to disprove the null hypothesis (H_0) that the differences (X) between pairs of observations from two related samples follow a symmetric distribution around zero (i.e. come from populations with similar median), at a given significance level defined as the p-value (i.e. the probability of the occurrence of H_0). When H_0 holds, Eq. (3) is true, and one can conclude that the two samples are not significantly different.

$$H_0 : P(X > 0) = P(0 > X) \tag{3}$$

For each feature, the MATLAB function *signrank* is used to test a sample of observations corresponding to bad quality segments, against a sample of observations corresponding to good quality segments. Each pair of observations comes from the same dataset, and the sample is randomly chosen from within the observations in the training set. A significance threshold of p = 0.01 is used for rejecting H_0. When H_0 holds, good and bad quality samples are not statistically different, and the feature under evaluation is excluded.

Correlation-based feature selection (CFS) [39]: an optimal feature set contains entries that are highly correlated with the target class (i.e. relevant) and uncorrelated with each other (i.e. non-redundant). To find the optimal feature subset, CFS calculates the merit (S_n) of a series of subsets consisting of *n* features according to Eq. (4), where *n* is an integer varying from 1 to *N* (total number of features after WSRt), $\overline{r_{ff}}$ is the mean

value of the feature-to-feature Pearson's correlation coefficients, and $\overline{r_{cf}}$ is the mean value of the feature-to-class Pearson's correlation coefficients within each n-feature subset [39].

$$S_n = \frac{n \times \overline{r_{cf}}}{\sqrt{n + n(n-1)\times \overline{r_{ff}}}} \tag{4}$$

In this work, greedy forward selection is used as the search strategy to find the best subset of features. Starting from a null subset, each possible feature is iteratively added to the subset. The feature resulting in the highest S_n is kept, and the procedure is repeated with the remaining features. The merits S_n are then ranked, and the subset with the highest merit is selected.

Recursive feature elimination with Support **Vector Machines (SVM-RFE)** [40] **and cross-validation:** finds the best subset of features by considering the performance of a linear SVM,[3] when iterating over an array of candidate feature subsets.

A backward search strategy is implemented, that starting with the whole set of features (N') produces candidate subsets of *n* features ($n \in \{1, …N'\}$, where N' is the total number of features elected by CFS). Each n-feature subset is obtained by randomly and iteratively eliminating one feature from the n+1 subset. The SVM is trained and validated using a 5-fold cross-validation scheme, on all possible n-feature subsets. The best performing subset is the one that minimizes the loss function (α) in Eq. (5), and will be the basis for the next backward search iteration.

$$\alpha = \left(1 - \frac{Sen + Spe}{2}\right) \tag{5}$$

Considering all best performing n-feature candidate subsets, the overall best feature set is the one that yields the highest BAcc, obtained according to Eq. (6).

$$BAcc = 100 - (\alpha \times 100) \tag{6}$$

The MATLAB functions *sequentialfs* and *cvpartition* are used to implement the SVM-REF method.

2.6.2. Neighborhood Component Analysis

NCA is a non-parametric approach used for dimensionality reduction [41]. It attributes a measure of importance (i.e. a weight) to each feature, after maximizing the expected accuracy of a leave-one-out classification,[4] based on a stochastic nearest-neighbor method [41]. The closer the weight is to zero, the less important the corresponding feature is. A regularization parameter (λ) is applied to scale the feature weights and assist in the detection of relevant features. This value is selected as the one producing the least classification error, after a 5-fold cross validation is ran for each λ within the empirical interval of $[0 : 1.05 : 20]/length(trainingdataset)$. Since the algorithm is not deterministic, to ensure robust results the NCA procedure is repeated 100 times with the same data input. The weight of each feature is then taken as the median value over the 100 repetitions. The MATLAB function *fscnca* is used here for implementing the NCA-based feature selection.

2.7. Classification: algorithms, training and validation

A set of supervised ML algorithms is employed for binary quality classification on each feature subset. The hyperparameters of the models are tuned over a finite range of possibilities, and each model is weighted on the training data. Three elementary classifiers were exploited in this work, plus ensemble methods:

---

[3] SVM (cf. chapter II.G) is a deterministic classifier, widely used in machine learning related to its robustness and simplicity.

[4] Leave-one-out classification: attributes a class to a single observation based on the consensus of the observations within its neighborhood.

Table 3
Models and hyperparameters for SVM and KNN.

Model	Hyperparameter	Model	Hyperparameter
Linear SVM	BoxConstraint: 1/ KernalScale: Auto KernalFunction: Linear PolynomialOrder: NA	Fine KNN	Distance: Euclidean DistanceWeight: Equal Exponent: NA/ NumNeighbors: 1
Quadratic SVM	BoxConstraint: 1/ KernalScale: Auto KernalFunction: Polynomial PolynomialOrder: 2	Medium KNN	Distance: Euclidean DistanceWeight: Equal Exponent: NA/ NumNeighbors: 10
Cubic SVM	BoxConstraint: 1/ KernalScale: Auto KernalFunction: Polynomial PolynomialOrder: 3	Coarse KNN	Distance: Euclidean DistanceWeight: Equal Exponent: NA/ NumNeighbors: 100
Fine Gaussian SVM	BoxConstraint: 1/ KernalScale: 2 KernalFunction: Gaussian PolynomialOrder: NA	Cosine KNN	Distance: Cosine DistanceWeight: Equal Exponent: NA/ NumNeighbors: 10
Medium Gaussian SVM	BoxConstraint: 1/ KernalScale: 7.9 KernalFunction: Gaussian PolynomialOrder: NA	Cubic KNN	Distance: Minkowski DistanceWeight: Equal Exponent: 3/ NumNeighbors: 10
Coarse Gaussian SVM	BoxConstraint: 1/ KernalScale: 31 KernalFunction: Gaussian PolynomialOrder: NA	Weighted KNN	Distance: Euclidean DistanceWeight: SquaredInverse Exponent: NA/ NumNeighbors: 10

- k-Nearest-Neighbors (KNN): uses the majority vote of *k* closest training observations in order to assign another observation to a class [42]. The hyperparameter *k* is bounded to the number of classes in the problem. The MATLAB function *fitcknn* is used to train the KNN models. Table 3 lists the hyperparameter configuration in the KNN models.
- Support Vector Machine (SVM): finds the optimal hyperplane, serving as the decision boundary to separate different classes of observations [43]. The hyperparameters include the kernel function, scale, order, offset, and the box constraint. The MATLAB function *fitcsvm* is used to train the SVM models. Table 3 lists the hyperparameter configuration in the SVM models.
- Decision Trees [44]: a divide-and-conquer model that uses a tree-like scheme. The leaf nodes represent the constraints on the feature values, leading to an end node that attributes the constrained observation to a class. The hyperparameters include the split criterion and the maximum number of splits. The MATLAB function *fitctree* is used to train Fine, Medium, and Coarse decision trees models, with Gini's diversity index as that split criterion, and a maximum number of splits equal 100, 20, and 4, respectively.
- Ensemble learning methods [45]: combine several weak learning methods into one ensemble predictor, yielding a higher performance predictor. The MATLAB function *fitcensemble* was used to train Boosted, Bagged, and RUSBoosted trees as well as Subspace KNN models, all of which using number of learning cycles equal 30.

To understand if the trained classification models were generalizable across datasets, the one achieving the overall highest performance was

applied on the testing subsets of the two remaining datasets.

2.8. Use case

The final objective of employing our quality indicator is to decrease the errors when calculating HR. To test the usability of the proposed algorithm towards such end, three 10-min ECG segments, with variable signal quality, were extracted from each dataset (wDS, ccDS, and phyDS). For each 10-s window in the three segments, the mean HR estimate was calculated based on R-peaks found by the Pan-Tompkins QRS detection algorithm [46], while the mean HR ground truth was calculated based on manually detected R-peaks. HR estimates (E) were compared to the ground truth HR (T), based on the root mean squared error (RMSE), normalized by the number of 10-s ECG windows (n), as depicted in Eq. (7). For each of the three datasets, the RMSE was first calculated on the whole set of windows, and afterwards on a reduced set of windows, after discarding those classified as low-quality by the best performing algorithm. Both values are compared to understand if the HR error decreases after discarding low quality sections from the calculation.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(T_i - E_i)^2}{n}} \quad (7)$$

3. Results

3.1. Feature selection

The hybrid method and the NCA were used independently towards feature selection. The results for the feature selection based on the hybrid method are presented in Table 4. The computation times for the hybrid method were 4.18 min, 4.81 min, and 0.38 min, respectively for the wDS, ccDS, and phyDS.

For the NCA-based feature selection, all features with median weights higher than 0.5 were selected. The results of NCA on the three datasets are presented in Fig. 7. The computation times for the NCA-based feature selection were 156.69 min, 157.25 min, and 21.07 min, respectively for wDS, ccDS, and phyD.

3.2. Classification

For the three ECG dataset, each of the 19 classification models considered in this work were trained on three subsets of features: those resulting hybrid feature selection, the NCA-based subset, and the full set of 43 features. The results concerning the best performing models, as established on the testing partition of each dataset, are graphically illustrated in Fig. 8 and numerically summarized in Table 5.

The model with highest overall performance among those developed on a feature subset (Fine Gaussian SVM with the ccDS-NCA feature subset) was applied to features extracted on wDS and phsDS testing partitions, resulting on 0.49 and 0.48 BAcc, respectively.

3.3. Use case

When considering the HR estimates for all 10-s windows on each 10-min segment from wDS, ccDS, and phyDSm, the resulting RMSEs are

Table 4
Results of the hybrid feature selection for the wDS, ccDS, and phyDS.

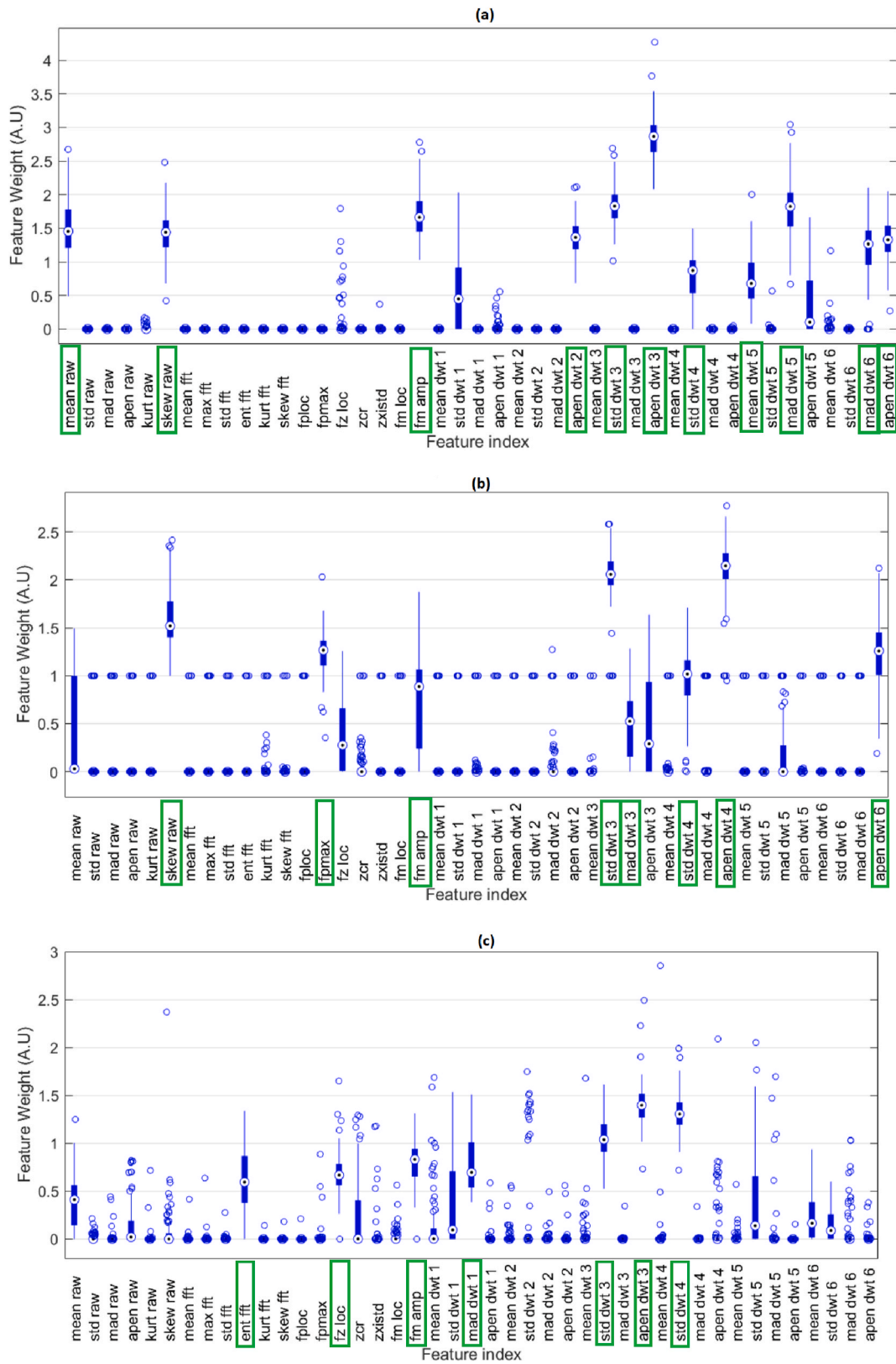| Dataset	Wearable ECG (wDS)			Non-contact ECG (ccDS)			Physionet ECG (phyDS)		
Method	WRSt	CSF	SVM-REF	WRSt	CSF	SVM-REF	WRSt	CSF	SVM-REF
# of selected features	39	11	7	40	12	7	31	12	10
Final subset	apen_swt_1; apen_swt_2; std_swt_3; std_swt_5; apen_swt_5; Fploc; fz_loc			mean_fft; max_fft; kurt_fft; skew_fft; mad_swt_3; mad_swt_5; fz_loc			mean_raw; apen_raw; std_fft; skew_fft; ent_fft; mad_swt_1; mean_swt_2; Fpmax; fm_amp; fz_loc		

Fig. 7. NCA feature weights across 100 iterations for each dataset: a) wDS, b) ccDS and c) phyDS. The green boxes indicate the selected features. Features described as DWT respect in fact to SWT.
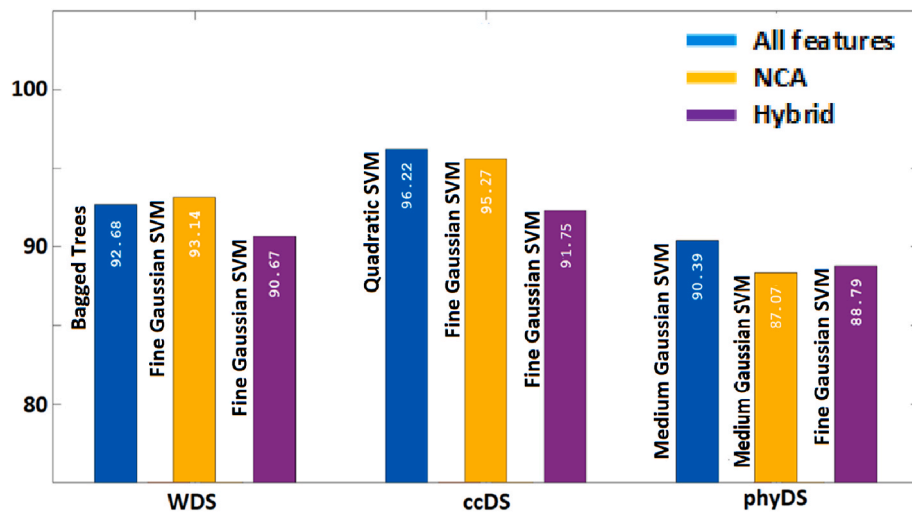
Fig. 8. Balanced accuracy scores of the best performing models on the testing partition for each ECG dataset using: all features, NCA, and Hybrid method feature subsets.

Table 5

Evaluation metrics for the best performing models on the testing partition of each ECG dataset. The differences between each pair of models (All-NCA, All-hybrid, NCA-hybrid) developed on each dataset are always statistically significant (p«0.01) .[a]

Dataset	Feature set	Model	Cross-Validation BAcc [mean, std]	BAcc	Sen	Spe	Prec	F1	Acc	Computation time [s]
wDS	All features	Bagged Trees	0.93, 0.01	0.93	0.93	0.92	0.92	0.92	0.93	6.18
	NCA	Fine Gaussian SVM	0.93, 0.01	0.93	0.95	0.91	0.91	0.93	0.93	4.09
	Hybrid	Fine Gaussian SVM	0.91, 0.01	0.91	0.93	0.88	0.88	0.91	0.91	4.04
ccDS	All features	Quadratic SVM	0.95, 0.01	0.96	0.99	0.94	0.79	0.88	0.95	11.50
	NCA	Fine Gaussian SVM	0.95, 0.01	0.95	0.98	0.93	0.76	0.86	0.94	3.31
	Hybrid	Fine Gaussian SVM	0.92, 0.01	0.92	0.95	0.88	0.66	0.78	0.90	4.02
phyDS	All features	Medium Gaussian SVM	0.89, 0.02	0.90	0.91	0.90	0.97	0.94	0.93	0.20
	NCA	Medium Gaussian SVM	0.88, 0.01	0.87	0.90	0.84	0.95	0.92	0.92	0.13
	Hybrid	Fine Gaussian SVM	0.87, 0.02	0.89	0.90	0.88	0.96	0.93	0.90	0.16

[a] Based on ROC comparisons using DeLonge test for paired designs, with the r-package: pROC.

1.88, 3.53, and 5.54 beats per minute (bpm), respectively. When considering only HR estimates for windows with *good quality* ratings, the resulting RMSEs were 0.69, 1.54, and 1.87 bpm, respectively for wDS, ccDS, and phyDS.

4. Discussion

The proposed feature space (43 features) covers a wide range of ECG dynamics. However, utilizing the full feature space is not always necessary to achieve acceptable results. Feature selection is therefore applied to find the best trade-off. The minimal impact of feature elimination on the classification performance is well illustrated in Table 4: and Fig. 7. When it comes to the hybrid method, the features are selected in a complete data driven manner and the method is deterministic. In this case, the user may still decide to rely on the feature subsets resulting from the WRSt, CFS or SVM-RFE steps. On the other hand, the NCA algorithm produces a ranking of the feature weights, allowing for flexible cut-off threshold setting. However, the NCA requires a high computation time (up to 157 min), compared to 5 min in the case of the hybrid method. This high computation time is due to the 100 repetitions of the procedure, to account for the nondeterministic behaviour of NCA algorithm. Thus, different users are free to opt for the method that better fits their requirements.

All feature families (i.e. raw, FFT, SWT, and ACF) were represented in the final feature subsets, but only 28 features were included at least once. Within the same selection method (i.e. hybrid or NCA), the final feature subsets appear to be inconsistent across datasets, which is expected due to the differences in ECG morphology and artefacts across different datasets. The feature *fz_loc* was among the features selected by

the hybrid method for all datasets. Concerning NCA, *std_swt_3*, *std_swt_4*, and *fm_amp* were among the features selected on all datasets. The selected features, within the same dataset, were also inconsistent, which can be traced back to the different analytical nature of the selection methods. Overall, the SWT-based features appeared to be important in all ECG datasets, but this is especially noticeable and consistent across feature selection methods in the case of wDS. The remaining two datasets may be more challenging to model, related to inclusion of pathological patterns (phyDS) and impactful artefacts related to loss of contact (ccDS), resulting in higher variability related to the features that may characterize them.

The best performing classifiers presented BAccs above 87% in the training/validation and testing phases (cf. Table 5). Out the 9 best performing models, 8 belong to the SVM family, with different hyperparameters. In the case of ccDS, the classification models presented highest BAccs (92–96%) but low precision (66–79%), which may be attributed to the highly unbalanced classes, with large number of bad segments. The models with most stable performances respected the wDS case (BAcc 91–93%, Prec/Spe>88%). The phyDS, being an unbalanced dataset, with higher number of good segments, and including ectopic beats resulted in lowest BAcc scores (87–90%).

A model trained on one dataset is not generalizable to the other datasets, as depicted by the poor performance (BAcc<50%) of the overall best performing model on the remaining data sets (wDS and physDS).

The use case shows that discarding low quality windows flagged by the quality indicators, substantially decreases the RMSE when calculating HR based on the input from an automated QRS detection method.

The physionet/Cinc Challenge 2011 data (i.e. phyDS) was employed

here in order to establish a benchmark with previous literature. When comparing the performance of our models on phyDS (Acc = 90–93%) to the results presented by Ref. [13,14,18] (Acc = 91.6–93.6%, cf. Table 1), we verify that our models have comparable performance. The remaining methods presented in Table 1 cannot be directly compared to our work as they use different datasets, though, they provide a guideline for the expected quality assessment performance in wearable data. Thus, we can conclude that our results are within the expected range for wearable data.

4.1. Limitations

The proposed pipeline, at its current state, is limited by the following factors:

- Sampling frequency dependency: some of the extracted features are dependent on the ECG sampling frequency. This was overcome by resampling the input ECG signals to the lowest sampling frequency across datasets (256 Hz).
- Normalization of the extracted features: as shown in Fig. 2, all extracted features undergo a normalization step before the feature selection and the classification steps over the whole dataset. This is important to enhance computation time in the coming steps. However, this limits the application of the proposed pipeline to offline applications.
- Binary-class vs. multi-class: the current work separates the low-quality ECG segments from those of high quality, which may leave out some nuances. Especially relevant if the objective was to estimate mean HR, in which case the quality requirements would be lower, or if the objective was morphology analysis, in which case the requirements would be higher. For the current objective of estimating instantaneous HR and HRV, more classes would not provide added value.
- ECG lead configuration: only ECG Lead I and Lead II signals were explored in this work.
- Pathological patterns: although ECG beats with pathological morphologies were presented in the some of the phyDS ECG segments, no sensitivity analysis towards ectopic beats was performed, which limits the use of the proposed solution to lifestyle applications (e.g. activity and wellbeing tracking).

4.2. Future work

A thorough evaluation of the capabilities of each feature to detect specific kinds of artefacts, could enrich the understanding of the ECG and the effects of noise at a more fundamental level. Testing the proposed pipeline using different medical signals (e.g. body impedance) will be performed in a future research.

5. Conclusion

A generic pipeline for generating automatic ECG signal quality indicators, targeting lifestyle applications, was presented and evaluated. Unlike other state-of-the-art solutions presented in previous literature, we propose a generic framework that can be used on diverse ECG datasets to train classification models in order to assess ECG quality. Moreover, with this framework we managed to obtain a high performing classifier for non-contact ECG, something that was not addressed in previous work. The models attain a high performance when tested on data from the same acquisition source. Though, they are not generalizable across different acquisition sources.

Declaration of competing interest

None Declared.

References

[1] J. Gawłowska, J. Wranicz, J. "Norman, "Jeff'' holter (1914-1983), Cardiol. J. 16 (4) (2009) 386–387.

[2] H.F. Macinnis, The clinical application of radioelectrocardiography, Can. Med. Assoc. J. 70 (5) (1954) 574–576.

[3] R. Trobec, I. Tomašić, A. Rashkovska, M. Depolli, V. Avbelj, Commercial ECG systems, in: *SpringerBriefs in Applied Sciences and Technology*, No. 9783319593388, Springer Verlag, 2018, pp. 101–114.

[4] I.D. Castro, M. Mercuri, A. Patel, R. Puers, C. Van Hoof, T. Torfs, Physiological driver monitoring using capacitively coupled and radar sensors, Appl. Sci. 9 (19) (Sep. 2019) 3994.

[5] Sameni, A review of fetal ECG signal processing issues and promising directions, Open Pacing Electrophysiol. Ther. J. 3 (4) (2010).

[6] S. Ramasamy, A. Balan, Wearable sensors for ECG measurement: a review, Sens. Rev. 38 (4) (17-Sep-2018) 412–419. Emerald Group Publishing Ltd.

[7] CinC [Online]. Available: http://www.cinc.org/. (Accessed 10 June 2020).

[8] A.L. Goldberger, et al., PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (Jun. 2000).

[9] Z. Zhao, Y. Zhang, SQI quality evaluation mechanism of single-lead ECG signal based on simple heuristic fusion and fuzzy comprehensive evaluation, Front. Physiol. 9 (Jun. 2018). JUN.

[10] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Vallance, L. Tarassenko, Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring, IEEE J. Biomed. Health Inf. 19 (3) (May 2015) 832–838.

[11] J. Lee, D.D. McManus, S. Merchant, K.H. Chon, Automatic motion and noise artifact detection in holter ECG data using empirical mode decomposition and statistical approaches, IEEE Trans. Biomed. Eng. 59 (6) (2012) 1499–1506.

[12] D. Hayn, B. Jammerbund, G. Schreier, QRS detection based ECG quality assessment, Physiol. Meas. 33 (9) (Sep. 2012) 1449–1461.

[13] H. Naseri, M.R. Homaeinezhad, Electrocardiogram signal quality assessment using an artificially reconstructed target lead, Comput. Methods Biomech. Biomed. Eng. 18 (10) (Jul. 2015) 1126–1141.

[14] J. Moeyersons, et al., Artefact detection and quality assessment of ambulatory ECG signals, Comput. Methods Progr. Biomed. 182 (Dec. 2019) 105050.

[15] U. Satija, B. Ramkumar, M. Sabarimalai Manikandan, A review of signal processing techniques for electrocardiogram signal quality assessment, IEEE Rev. Biomed. Eng. 11 (28-Feb-2018) 36–52. Institute of Electrical and Electronics Engineers.

[16] C. Liu, et al., Signal quality assessment and lightweight qrs detection for wearable ECG smartvest system, IEEE Internet Things J. 6 (2) (Apr. 2019) 1363–1374.

[17] C. Orphanidou, I. Drobnjak, Quality assessment of ambulatory ECG using wavelet entropy of the HRV signal, IEEE J. Biomed. Health Inf. 21 (5) (Sep. 2017) 1216–1223.

[18] E. Morgado, et al., Quality estimation of the electrocardiogram using cross-correlation among leads, Biomed. Eng. Online 14 (1) (Jun. 2015).

[19] U. Satija, B. Ramkumar, M.S. Manikandan, An automated ECG signal quality assessment method for unsupervised diagnostic systems, Biocybern. Biomed. Eng. 38 (1) (Jan. 2018) 54–70.

[20] A. Ghaffari, M.R. Homaeinezhad, M. Akraminia, M. Atarod, M. Daevaeiha, A robust wavelet-based multi-lead electrocardiogram delineation algorithm, Med. Eng. Phys. 31 (10) (Dec. 2009) 1219–1227.

[21] K.H. Brodersen, C. Soon Ong, K.E. Stephan, J.M. Buhmann, The Balanced Accuracy and its Posterior Distribution, 2010.

[22] I. Castro Miller, Robust Capacitively-Coupled Electrocardiography and Bioimpedance for Unobtrusive Health Monitoring, 2019.

[23] T. Wartzek, M. Czaplik, C.H. Antink, B. Eilebrecht, R. Walocha, S. Leonhardt, " UnoViS, The MedIT public unobtrusive vital signs database, Health Inf. Sci. Syst. 3 (1) (Dec. 2015) 1–9.

[24] M. Malik, Heart rate variability: standards of measurement, physiological interpretation, and clinical use, Circulation 93 (5) (1996) 1043–1065.

[25] R.J. Ellis, B. Zhu, J. Koenig, J.F. Thayer, Y. Wang, A careful look at ECG sampling frequency and R-peak interpolation on short-term measures of heart rate variability, Physiol. Meas. 36 (9) (2015) 1827–1852.

[26] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier Inc., 2016.

[27] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (Jun. 2006) 861–874.

[28] P. Duhamel, M. Vetterli, FAST FOURIER TRANSFORMS: A TUTORIAL REVIEW AND A STATE OF THE ART, 1990.

[29] M. Frigo, S.G. Johnson, FFTW: an adaptive software architecture for the FFT, in: *ICASSP*, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 3, 1998, pp. 1381–1384.

[30] O. Rioul, M. Vetterli, Wavelets and signal processing, IEEE Signal Process. Mag. 8 (4) (1991) 14–38.

[31] J.C. Pesquet, H. Krim, H. Carfantan, Time-invariant orthonormal wavelet representations, IEEE Trans. Signal Process. 44 (8) (1996) 1964–1970.

[32] G.P. Nason, B.W. Silverman, The Stationary Wavelet Transform and Some Statistical Applications, Springer, New York, NY, 1995, pp. 281–299.

[33] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.

[34] B.N. Singh, A.K. Tiwari, Optimal selection of wavelet basis function applied to ECG signal denoising, Digit. Signal Process. A Rev. J. 16 (3) (2006) 275–287.

[35] K.S. Khobragade, R.B. Deshmukh, ECG analysis using wavelet transforms, IETE J. Res. 43 (6) (1997) 423–432.

[36] H.M. Rai, A. Trivedi, S. Shukla, ECG signal processing for abnormalities detection using multi-resolution wavelet transform and Artificial Neural Network classifier, Meas. J. Int. Meas. Confed. 46 (9) (2013) 3238–3246.

[37] D.J. Bartholomew, Time series analysis forecasting and control, J. Oper. Res. Soc. 22 (2) (1971) 199–201.

[38] C. Price, Autocorrelation function (ACF) - file exchange - MATLAB central [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/30540-autocorrelation-function-acf, 2020. (Accessed 20 April 2020).

[39] M.A. Hall, Correlation-based Feature Selection for Machine Learning, 1999.

[40] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.

[41] W. Yang, K. Wang, W. Zuo, Neighborhood component feature selection for high-dimensional data, J. Comput. 7 (1) (2012) 162–168.

[42] N.S. Altman, An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, 1992.

[43] B. Schölkopf, SVMs - a practical consequence of learning theory, IEEE Intell. Syst. Their Appl. 13 (4) (Jul. 1998) 18–21.

[44] J.R. Quinlan, Simplifying decision trees, Int. J. Man Mach. Stud. 27 (3) (Sep. 1987) 221–234.

[45] R. Maclin, D. Opitz, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. 11 (Jun. 2011) 169–198.

[46] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. BME-32 (3) (1985) 230–236.