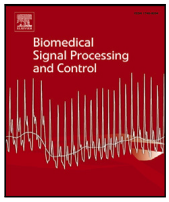
Biomedical Signal Processing and Control 91 (2024) 105920

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Single-lead electrocardiogram quality assessment in the context of paroxysmal atrial fibrillation through phase space plots

Álvaro Huerta [a,*], Arturo Martinez-Rodrigo [a], Vicente Bertomeu-González [b], Óscar Ayo-Martin [c], José J. Rieta [d], Raúl Alcaraz [a]

[a] Research Group in Electronic, Biomedical and Telecommunication Engineering, University of Castilla-La Mancha, Cuenca, Spain
[b] Clinical Medicine Department, Miguel Hernandez University, Elche, Spain
[c] Department of Neurology, University Hospital of Albacete, Albacete, Spain
[d] BioMIT.org, Electronic Engineering Department, Universitat Politecnica de Valencia, Valencia, Spain

ARTICLE INFO

Keywords:
Signal quality assessment
Paroxysmal atrial fibrillation
Phase space portraits
Machine learning classifiers
Deep learning algorithms

ABSTRACT

Current wearable electrocardiogram (ECG) recording systems have great potential to revolutionize early diagnosis of paroxysmal atrial fibrillation (AF). They are able to continuously acquire an ECG signal for long weeks and then increase the probability of detecting first brief, intermittent signs of the arrhythmia. However, the recorded signal is often broadly corrupted by noise and artifacts, and accurate assessment of its quality to avoid automated misdiagnosis and false alarms of AF is still an unsolved challenge. In this context, the present work is pioneer in exploring the usefulness of transforming the single-lead ECG signal into two common phase space (PS) representations, such as the Poincaré plot and the first order difference graph, for evaluation of its quality. Several machine and deep learning models fed with features and images derived from these PS portraits reported a better performance than well-known previous methods, even when they were trained and validated on two separate databases. Indeed, in binary classification of high- and low-quality ECG excerpts, the generated PS-based algorithms reported a discriminant power greater than 85%, misclassifying less than 20% of high-quality AF episodes and non-normal rhythms as noisy excerpts. Moreover, because both PS reconstructions do not require any mathematical transformation, these algorithms also spent much less time in classifying each ECG excerpt in validation and testing stages than previous methods. As a consequence, ECG transformation to both PS portraits enables novel, simple, effective, and computational low-cost techniques, based both on machine and deep learning classifiers, for ECG quality assessment.

1. Introduction

Cardiovascular disorders (CVDs) are the most common non-communicable diseases worldwide, being responsible for about 18 million annual deaths (i.e., more than 31% of all deaths globally) [1,2]. A prevalent cause of these diseases is the presence of cardiac arrhythmias, which are mainly featured by diverse conditions of irregular heart rhythm [3]. Among these cardiac alterations, atrial fibrillation (AF) is the most frequent, provoking high morbidity and triggering ischemic stroke [4]. Despite the huge recent technological advances in the field of healthcare, AF prevalence continues to grow in the developed countries. Indeed, this arrhythmia is today considered one of the most important epidemics in the 21st century and involves a major economic burden to the society [4].

So far, the mechanisms triggering and supporting AF are not fully known, leading current therapies to be suboptimal [5]. Moreover, the arrhythmia presents a recursive nature, which fosters its perpetuation. Although nascent arrhythmic episodes are often sporadic and of a short length, their frequency and duration increase over the time [6]. Hence, the early detection of these paroxysmal events plays a key role to manage successfully AF treatment, reduce its perpetuation, and limit its prevalence [6]. However, current clinical guidelines demand AF documentation on a sufficient quality electrocardiogram (ECG) recording before administration of any treatment [7]. Although current ambulatory systems are able to acquire an ECG signal with several days in length, a longer recording time increases the potential of detecting the intermittent arrhythmic events and providing an earlier diagnosis of AF [7]. Furthermore, conventional ECG devices involve burdensome wires and adhesive electrodes, which interfere with the user's daily activities and provoke skin irritation.

* Correspondence to: Inst. Tecnologías Audiovisuales, Campus Universitario s/n, C.P. 16071, Cuenca, Spain.
E-mail address: alvaro.huerta@uclm.es (Á. Huerta).

https://doi.org/10.1016/j.bspc.2023.105920
Received 17 August 2023; Received in revised form 11 December 2023; Accepted 29 December 2023
Available online 9 January 2024

Some recent wearable systems overcome these problems by continuously acquiring an ECG signal for long weeks or months in a comfortable, unobtrusive way for the patient [8]. They present the potential to revolutionize the diagnosis and management of AF and other CVDs, as they are able to detect these pathologies without altering the patient's daily life [8]. However, the signal recorded by these devices in a non-resting state is often contaminated by strong artifacts and noises, which can seriously mask the ECG morphology and confound the physicians' diagnosis [9,10]. Hence, bearing the large amounts of captured ECG data in mind, automated and accurate signal quality assessment for identification and rejection of the low-quality excerpts before every diagnosis is essential to reach a massive use of these wearable ECG acquisition systems in daily clinical practice [9–11].

In the last years, a wide variety of algorithms based on traditional machine learning and modern deep learning approaches have been proposed for quality evaluation of the standard 12-lead ECG [9,10]. However, many are not applicable to the ECG recording acquired by wearable systems, as it only contains a reduced number of leads (often among 1 and 3 leads). Nonetheless, in the context of single-lead ECG quality assessment, most of the published methods are still based on the same supervised learning concepts, and two large groups of algorithms can be identified. On the one hand, some are based on merging hand-crafted features with common machine learning classifiers, such as support vector machine (SVM), decision tree, and random forest, among others [9,10]. In this case, statistical, morphological, and time–frequency domain features are mainly derived from the raw or preprocessed ECG signal or from its delineated fiducial points [9,10]. Although they have reported promising results when short, resting ECG signals are assessed, their performance is often degraded on long-term ECG signals obtained by wearable systems in dynamic, ever-changing environments [9,10].

On the other hand, deep learning-based schemes have been more recently proposed for ECG quality assessment. In this case, manual selection of ECG features is not required, thus reducing subjectivity and redundant information among the variables. These algorithms are mainly based on convolutional neural networks (CNNs), which usually achieve more abstract, low-level ECG representations, and consequently better classification between high- and low-quality excerpts, than the previous machine learning methods [11,12]. Given the excellent results obtained in the field of image processing and computer vision [13], two-dimensional (2-D) CNN schemes have been mostly used. Moreover, the analysis of ECG-based images has proven additional advantages regarding the study of original one-dimensional (1-D) signals, such as identification of a richer set of features, easier application of data augmentation, and longer insensitivity to some characteristics of ECG acquisition process, such as sampling rate, amplitude, and intrinsic noise [14]. To turn the ECG signal into a 2-D image, diverse kinds of time–frequency transformations have been mainly considered to date, including short-time Fourier transform [15], continuous Wavelet transform [12], modified frequency slice Wavelet transform [16], and Stockwell transform [17], among others.

Although these time–frequency analyses have proven to be effective to represent non-stationary and time-varying physiological signals and noises [18], they are only based on linear concepts. Hence, other approaches more optimized to visualize and feature nonlinear, complex, and chaotic dynamics in a time series, such as diverse phase space (PS) representations, could reveal useful novel information to identify low-quality ECG excerpts. Indeed, some CNN schemes fed with PS-based images have already provided promising results in diverse ECG-based scenarios, e.g., in the assessment of patient's eligibility for cardioverter implantation [19], or in biometric authentication [20]. However, the use of diverse PS portraits has not still been thoroughly explored in the context of ECG quality assessment. Hence, the main goal of the present work is to analyze for the first time whether ECG transformation into two common PS graphs, such as the Poincaré plot (PP) and first order difference graph (FODG), might be helpful in quality evaluation of long-term, single-lead ECG signals, when both machine learning and deep learning classifiers are used.

All the classification models will be generated and assessed in the challenging context of paroxysmal AF. Although several previous algorithms, both based on machine learning and deep learning approaches, have shown to successfully work on long-term ECG recordings acquired from healthy subjects, their performance has proven to be much more limited on signals obtained from patients with such an intermittent arrhythmia [15,21]. The presence of paroxysmal AF episodes often provokes significant changes in the ECG signal, resembling its waveform and time–frequency features to the most common transient noise found during the acquisition procedure [22]. In fact, this kind of noise is the most relevant source of false alarms of AF in current continuous ECG monitoring, both by insertable Holters in dynamic environments [23] and by bedside monitors in intensive care units [24]. These false alarms account for more than 70% of the total detected events in both cases [24], and therefore automated quality assessment of long-term, single-lead ECG signals is a challenge of great significance for accurate diagnosis of paroxysmal AF using wearable devices [25].

2. Databases

Two separate databases were analyzed to consider a wide variety of ECG morphologies, artifacts and noises. For each dataset, the ECG signals were obtained within different contexts, as well as making use of diverse wearable acquisition systems. On the one hand, a proprietary database (PDB) was firstly enrolled. This consisted of 2 hour-length, single-lead ECG intervals extracted from much longer recordings obtained from 25 AF patients (12 women and 13 men, aged between 52 and 68 years), who presented intermittent arrhythmic episodes over time. After catheter ablation, they were continuously monitored for several weeks through a textile wearable Holter system (Nuubo™), placed on the thorax and capturing an ECG signal with a frequency of 250 Hz and a resolution of 12 bits over a dynamic range of ±5 mV. All the patients gave consent to be monitored, and the study was approved by the Ethical Review Board of Hospital Universitario San Juan de Alicante (Protocol Number UGP–14–219). The detection of paroxysmal AF episodes was tackled by an automatic algorithm [26] and visually supervised by two expert physicians. Similarly, noisy and clean ECG segments were manually labeled by two expert reviewers on the basis of their ability to detect R-peaks. Those ECG excerpts where the reviewers were able to unequivocally identify R-peaks were labeled as high-quality, whereas the remaining were marked as low-quality. Arrhythmias other than AF and premature ventricular and atrial contractions in the subset of high-quality ECG intervals were also labeled by the reviewers as other rhythms (OR). Eventually, from these labeled ECG segments, 10,000 noisy and 10,000 high-quality excerpts of 5 s in length were randomly selected to build the final database. Note that the group of high-quality ECG intervals was composed of 7650, 1750, and 600 ECG excerpts from normal sinus rhythm (NSR) segments, AF episodes, and OR intervals, respectively, such as Table 1 summarizes.

On the other hand, the publicly available training set of the PhysioNet/CinC Challenge 2017 Database (PC2017DB) [27,28] was also analyzed. This dataset contains 8528 ECG recordings with a duration ranging from 9 to 60 s. They were acquired by a portable AliveCor™ device, which operates linked to a smartphone with a sampling frequency of 300 Hz and 16 bits of resolution over a dynamic range of ±5 mV. Several experts annotated the ECG recordings, discerning among four different classes, i.e., NSR, AF, OR, and noisy signals. Once the recordings were segmented into 5 second-length excerpts, the resulting 47,439 high-quality and 1168 low-quality ECG intervals formed the second dataset included in the study. As can be seen in Table 1, a total of more than 68,600 ECG excerpts of 5 s in length were finally analyzed.

Table 1
Total number of ECG segments included in the two databases enrolled in the study.

Class		Database		Total
		PDB	PC2017DB	
High-quality	NSR	7650	28,413	57,439
	AF	1750	4329	
	OR	600	14,697	
Low-quality		10,000	1168	11,168
Total		20,000	48,607	**68,607**

3. Methods

3.1. Data preprocessing

An acquisition rate of 250 Hz is commonly considered sufficient to detect R-peaks, and accurately analyze and interpret ECG signals [29]. Bearing in mind that it was the lowest sampling rate between the two databases, all ECG excerpts from the PC2017DB were resampled to this frequency. A forward/backward, 8th-order Chebyshev low-pass filtering was used to avoid aliasing. Additional filtering stages were also considered to reduce typical artifacts and noises acquired along with the ECG signal. Hence, the baseline wandering was removed using a moving median filtering of order equal to half of the sampling frequency [30]. An algorithm based on the stationary Wavelet transform was also employed to remove the powerline interference and other high-frequency noises, but mostly preserving the original ECG morphology [31]. Lastly, all ECG excerpts were normalized to avoid any bias related to the amplitude and highlight the different morphology from high- and low-quality ECG segments. A generalized min–max scaling was applied to restrict the range of values in the ECG between −1 and 1. Precisely, denoting the preprocessed N sample-length ECG interval as $x(n) = \{x(1), x(2), \dots, x(N)\}$, its normalized version was obtained as

$$e(n) = -1 + \frac{2 \cdot \Big(x(n) - \min\{x(n)\}\Big)}{\max\{x(n)\} - \min\{x(n)\}}. \tag{1}$$

3.2. Phase space reconstruction of the ECG signal

Phase space representations obtain a geometric view of the underlying dynamics of a system to facilitate the study of its behavior. A valid PS is any vector space where the dynamical system's behavior can be unequivocally defined at every point [32]. The most used way to reconstruct full dynamics of a system from the generated time series is the well-known Takens' delay embedding theorem, which is based on plotting the original time series versus its time-delayed copies [32]. Accordingly, for a normalized ECG segment $e(n)$, the time-lagged PS vectors are given by

$$\mathbf{E}(i) = \{e(i), e(i+\tau), \dots, e(i+(m-1)\cdot\tau)\}, \tag{2}$$

i ranging from 1 to $N-(m-1)\cdot\tau$, m being the embedding dimension of the PS, and τ the time delay between points in the series. The optimal selection of both parameters m and τ is a key step to obtain fully representative PS portraits of the system's behavior [32]. Although several approaches have been proposed to automatically optimize their values [32], the special case of $m = 2$ and $\tau = 1$ has been widely considered to feature the ECG morphology [19] and heart rate variability [33] in a broad variety of contexts. This simplified 2-D PS is commonly referred to as the PP [33]. As an example, Fig. 1 displays the PP for typical 5 second-length ECG excerpts presenting NSR, AF, OR, and noise. Note that, whereas the first three cases, all belonging to the high-quality group, do not show large differences in the PP, the distribution of points in the last case (low-quality ECG excerpt) was markedly different.

To avoid the use of a specific value of τ and make the resulting distribution of points (or attractor) independent on the parameters associated with PS reconstruction, a modified version of the PP has also been recently proposed [34]. This plot is obtained by displaying the first order difference (i.e., $e(n+1)-e(n)$) regarding the original time series (i.e., $e(n)$). In this case, the resulting geometrical figure is different from the PP, such as Fig. 2 presents for the same ECG excerpts displayed in Fig. 1. The largest semi-circle in this FODG is usually corresponding to the QRS complex [34], thus exhibiting a clearly defined pattern in high-quality ECG segments (i.e., panels (a), (b) and (c)), regardless of whether they come from NSR, AF, and OR episodes. Hence, the usefulness of both kinds of graphs, PP and FODG, to discern between high- and low-quality ECG intervals using machine learning and deep learning approaches was analyzed.

3.3. Classification algorithms based on machine learning concepts

To classify high- and low-quality ECG excerpts through machine learning models, different descriptive features and statistics from the PP and FODG were manually derived. Thus, elliptical geometry exhibited by the distribution of points in the PP was firstly quantified through two well-known standard deviations. In short, these measures, referred to as $\mathcal{S}_1$ and $\mathcal{S}_2$, were computed by fitting an ellipse on the identity line [35]. The index $\mathcal{S}_1$ is the standard deviation of the dispersion of points perpendicular to the identity line and is linked to the short-term variability in the time series. Similarly, $\mathcal{S}_2$ is the standard deviation of the dispersion of points along the identity line and is related to the long-term variability in the time series. To define computation of $\mathcal{S}_1$ and $\mathcal{S}_2$ in mathematical terms, two new time series have to be defined from the normalized ECG, $e(n)$, such that

$$e_1(n) = e(n) - e(n+1), \tag{3}$$

and

$$e_2(n) = e(n) + e(n+1), \tag{4}$$

for $n = 1, 2, \dots, N-1$. Then, both standard deviations can be computed as

$$\mathcal{S}_1 = \sqrt{\frac{1}{2(N-1)} \sum_{i=1}^{N-1} \big(e_1(i) - \mu_1\big)^2} \tag{5}$$

and

$$\mathcal{S}_2 = \sqrt{\frac{1}{2(N-1)} \sum_{i=1}^{N-1} \big(e_2(i) - \mu_2\big)^2}, \tag{6}$$

being μ_1 and μ_2 the mean of $e_1(n)$ and $e_2(n)$, respectively, i.e.,

$$\mu_1 = \frac{1}{N-1} \sum_{i=1}^{N-1} e_1(i) \tag{7}$$

and

$$\mu_2 = \frac{1}{N-1} \sum_{i=1}^{N-1} e_2(i). \tag{8}$$

In addition to these two variables, the well-established ratio of them, i.e. $\mathcal{S}_{12} = \mathcal{S}_1/\mathcal{S}_2$, was also obtained to reflect the relationship between the short- and long-term variabilities in the time series [33].

As an alternative to this elliptical fitting, a novel approach to characterize the distribution of points in the PP has been recently proposed by dividing the graph into a grid of $C \times C$ cells with a fixed size [36]. Because the ECG signal was normalized between −1 and 1 and both axes in the PP ranging between these extremes, the variable C serves as a coarse-graining parameter of the plot. Several measures were then derived from the gridded plot with the idea of quantifying the differences between well-defined, regular distributions of points associated with the high-quality ECG excerpts and irregular, chaotic attractors presented by the low-quality and noise-corrupted ECG intervals. Precisely, the number of void grids ($\mathcal{V}$), i.e., the number of cells containing no points, the interquartile range ($\mathcal{I}$) of non-void

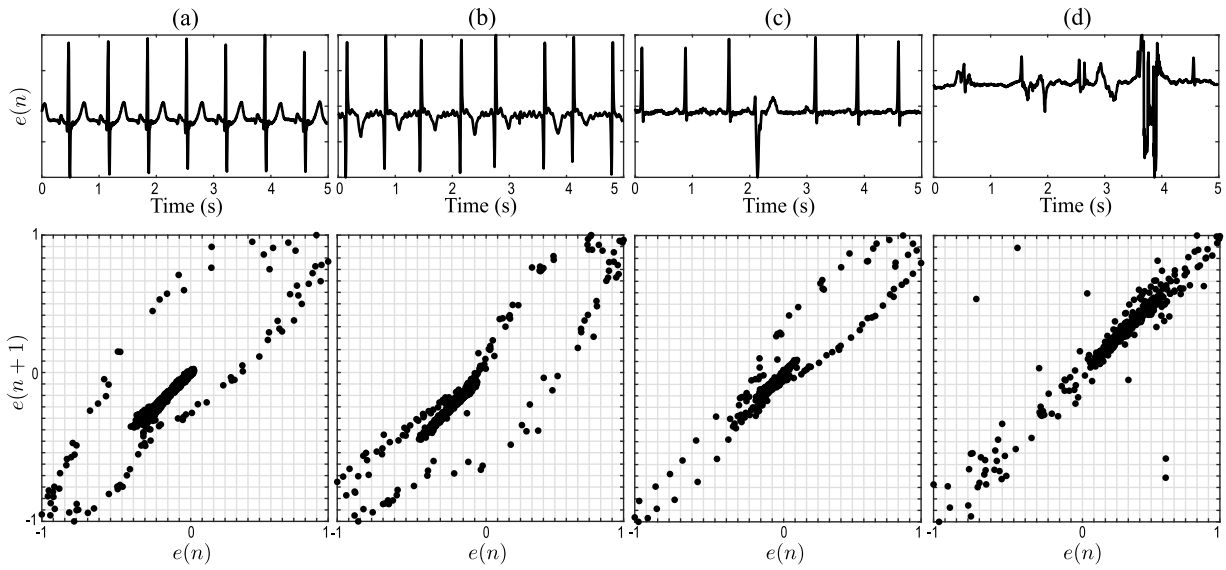


Fig. 1. Poincaré plots (bottom panel) obtained for typical 5 second-length ECG segments (top panel) from (a) NSR, (b) AF, (c) OR, and (d) noisy episodes.

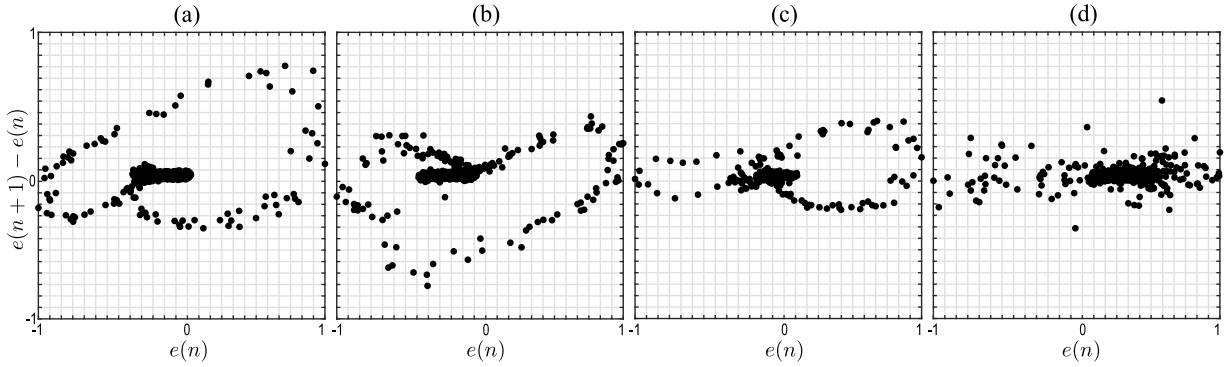


Fig. 2. First order difference graphs obtained for the same 5 second-length ECG segments presented in Fig. 1 from (a) NSR, (b) AF, (c) OR, and (d) noisy episodes.

grids, i.e., of those containing one or more points, the median absolute deviation ($\mathcal{M}$) of non-void grids, the number of points accumulated by the grid cells located on the main diagonal (D) and at a normalized distance less than ± 0.2, the largest number of points accumulated by a grid cell ($\mathcal{L}$) and the relative position ($\mathcal{P}$) of this cell within the $C \times C$ matrix, and finally the gridded distribution entropy ($\mathcal{E}$). To compute this last parameter, the percentage of points in each cell regarding the total one (i.e., p_i for $i = 1, 2, \ldots, C \times C$) was firstly computed. Then, after discarding those void grids, $\mathcal{E}$ was estimated by computing common Shannon Entropy, such that

$$\mathcal{E} = -\sum_{i=1}^{C \times C - \mathcal{V}} p_i \cdot \ln(p_i). \tag{9}$$

In general terms, irregular, chaotic attractors are expected to result in less void grids (i.e., lower values of $\mathcal{V}$), greater dispersion in distribution of points (i.e., larger values of $\mathcal{I}$ and $\mathcal{M}$), less points accumulated around the main diagonal (i.e., lower values of D), more variable content and position of the cell including the largest number of points (i.e., larger differences in values of $\mathcal{L}$ and $\mathcal{P}$), and more irregularity in the displayed information (i.e., larger values of $\mathcal{E}$) than regular, repetitive ones.

The same grid mapping approach was also used to qualitatively characterize the distribution of points in the FODG. However, in this case the parameters S_1, S_2, and S_{12} were not computed, and additionally the index D was modified for a better quantification of the resulting attractors. Thus, the number of points accumulated by the cells placed on the abscissa axis and at a normalized distance of $\pm$ 0.2 was computed. As a summary, the indices $\mathcal{V}$, $\mathcal{I}$, $\mathcal{M}$, D, $\mathcal{L}$, $\mathcal{P}$, and $\mathcal{E}$ were obtained with the same idea of discerning between regular and chaotic attractors on the FODG.

The variables derived from each one of the two PS representations were then used to generate two different classification models between high- and low-quality ECG excerpts. Those finally included in each model were automatically chosen by making use of a common wrapper-type feature selection algorithm, such as a forward sequential selection technique [37]. Thus, according to their statistical relevance, the features were sequentially added to an empty candidate set until the addition of further ones did not decrease the classification error, assessed inside repeated cross-validation loops to reduce overfitting. The optimal subset of indices obtained for each PS portrait was then combined through a GentleBoost algorithm, which creates a strong ensemble classifier by weighted voting from a set of weak learners. In the present work a decision tree model was used as weak learner, and the Gini's index was employed to split data. It was also established a maximum number of decision splits of 8, a maximum number of learning cycles of 20, and a learning rate of 0.015.

3.4. Classification algorithms based on deep learning concepts

In contrast to the machine learning classifiers, the deep learning models do not combine hand-crafted, knowledge-based discrete variables. Indeed, they use neural networks to holistically assess global

Table 2
Sequence of layers and their characteristics included in the proposed CNN.

Layer	Type of layer	# parameters	# Filters or neurons	Kernel size/Stride	Output size
0	Input layer	–	–	–	$C \times C \times 3$
1	2-D Convolution layer	448	16	3/1	$C \times C \times 16$
2	Bath normalization	32	–	–	$C \times C \times 16$
3	Average pooling layer	–	–	k/k	$22 \times 22 \times 16$
4	Fully connected layer	154,900	20	–	$1 \times 1 \times 20$
5	Dropout function (0.25)	–	–	–	$1 \times 1 \times 20$
6	Fully connected layer	42	2	–	$1 \times 1 \times 2$
7	Softmax function	–	–	–	$1 \times 1 \times 2$
8	Classification layer	–	–	–	$1 \times 1 \times 2$

information in inputted images and then extract more abstract, low-level features [38]. The architecture of a common 2-D CNN consists of a set of different layers operating in a sequential and/or parallel way [38]. Thus, the network initially presents, at least, a convolutional layer that extracts local features from the input image by its convolution with different filters. This layer is generally followed by a pooling one, which combines similar features to make the model simpler and more robust to noise and input deformations. The features resulting from this layer represent the original image from different angles, achieving more abstract representations when the number of convolutional and pooling layers interconnected in series or parallel increases. The network ends with, at least, a fully-connected layer, which converts the 2-D feature map obtained by a pooling layer into a 1-D vector to estimate the probability distribution of belonging to each output class. Apart from the layers, other mathematical functions, such as rectified linear units (ReLU), data normalizations, and dropout regularizations, are often included in intermediate points of the network to enhance its generalization capability and reduce overfitting [38].

To discern between low- and high-quality ECG segments using this kind of network, the gridded versions of the PP and FODG were firstly transformed into 2-D images. For that purpose, a Jet colormap with 128 colors was applied to the range of values obtained by computing the logarithm of the number of points in each grid cell. As an example, the images resulting from the ECG segments displayed in Figs. 1 and 2 for $C = 25$ are respectively presented in the top and bottom panels of Fig. 3. Next, with the idea of developing a lightweight network able to operate even in resource-constrained environments (i.e., with computational power, memory capacity, and battery limitations) [39], a minimal structure of layers was designed, such as Table 2 displays. Precisely, after the input image passing through a convolutional layer, a batch normalization, and an average pooling layer, the obtained 2-D feature map was converted to a 1-D vector through a first fully-connected layer. To minimize overfitting, a dropout function with a rate of 25% was introduced before the final fully-connected layer, which was connected to a softmax classifier. The obtained softmax mapping score was lastly compared with the corresponding input label to calculate cross-entropy loss and then classify the input image. Note that ReLU activation functions were inserted after the batch normalization and the first fully-connected layer. Moreover, the main goal of the pooling layer was to keep the number of network parameters constant as the $C \times C$ matrix of cells increased. Thus, the pooling region size and stride took different values (k) to maintain a feature map of size $22 \times 22 \times 16$ regardless of the value of C.

3.5. Previous classification algorithms

To serve as a reference, two well-known methods based on combining hand-crafted, ECG-based features through conventional machine learning classifiers were implemented. The first algorithm was proposed by Behar et al. [21] and blended seven ECG-based parameters through an SVM classifier. The variables were the fraction of beats simultaneously detected by two previously published R-peak detectors, the ratio of the number of beats detected by these two detectors, the relative power in the QRS complex, the third moment (i.e., skewness) of the signal, the fourth moment (i.e., kurtosis) of the signal, the relative power in the baseline, and finally the ratio of the sum of the eigenvalues associated with the five principal components over the sum of all eigenvalues obtained by a principal component analysis applied to the time-aligned ECG beats detected by one of the previous R-peak detectors. More recently, Albaba et al. [29] have analyzed a generic machine learning pipeline and a broad variety of ECG-based features to discern between high- and low-quality ECG excerpts. The algorithm reporting the best performance was based on an SVM classifier and seven variables, i.e., mean, maximum, kurtosis and skewness of the spectral distribution of the ECG signal, median absolute deviation of the wavelet scales 3 and 5 of the ECG signal, and finally the location of the first zero-crossing in the autocorrelation function of the ECG signal. Note that, an SVM classifier with a Gaussian kernel (scale of 1 and maximum penalty on margin-violating observations of 25) was employed in both cases.

Moreover, other two previous ECG quality indices based on more advanced deep learning concepts were also implemented for comparison purposes. On the one hand, Liu et al. [11] introduced a method composed of a three-layer wavelet scattering network and a bidirectional long short-term memory (Bi-LSTM) architecture. In brief, scattering coefficients of order 0, 1, and 2 were generated from the normalized ECG segment, $e(n)$, using a wavelet scattering network, constructed with a Morlet wavelet function. The resulting scattering feature matrix was then inputted into a Bi-LSTM network to discern between high- and low-quality ECG excerpts. On the other hand, because 2-D CNN schemes fed with ECG-based images have provided promising classification results in a variety of scenarios [40], the algorithm proposed by Zhao et al. [16] was also considered. In this case, the raw ECG signal was transformed into a 2-D image using a modified frequency slice Wavelet transform and then inputted to a CNN architecture, consisting of 9 learnable layers. Thus, it included an input layer designed to receive a grayscale image of 200 × 50 pixels, three convolution layers, three maximum pooling layers, one flatten layer, and one fully-connected layer.

3.6. Training, validation, and testing of the classification algorithms

The most accurate and unbiased overview of the performance of a classification model can only be obtained by conducting its training and testing on separate databases, because other validation approaches where data from the same patients are shared in both stages often provide inflated results [41,42]. Such an external validation with independent datasets has been vigorously advocated for every clinical application by the Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative [43]. Hence, in the present work the PDB and PC2017DB were detachedly used for training and testing of all generated classification algorithms.

The PDB was specifically collected to be balanced and then minimize misclassification of the samples belonging to the minority class in a prospective performance. In this respect, models trained with unbalanced datasets often exhibit a classification bias towards the majority class, due to its increased prior probability [44]. Moreover, to assess learning of the models during training, this dataset was divided

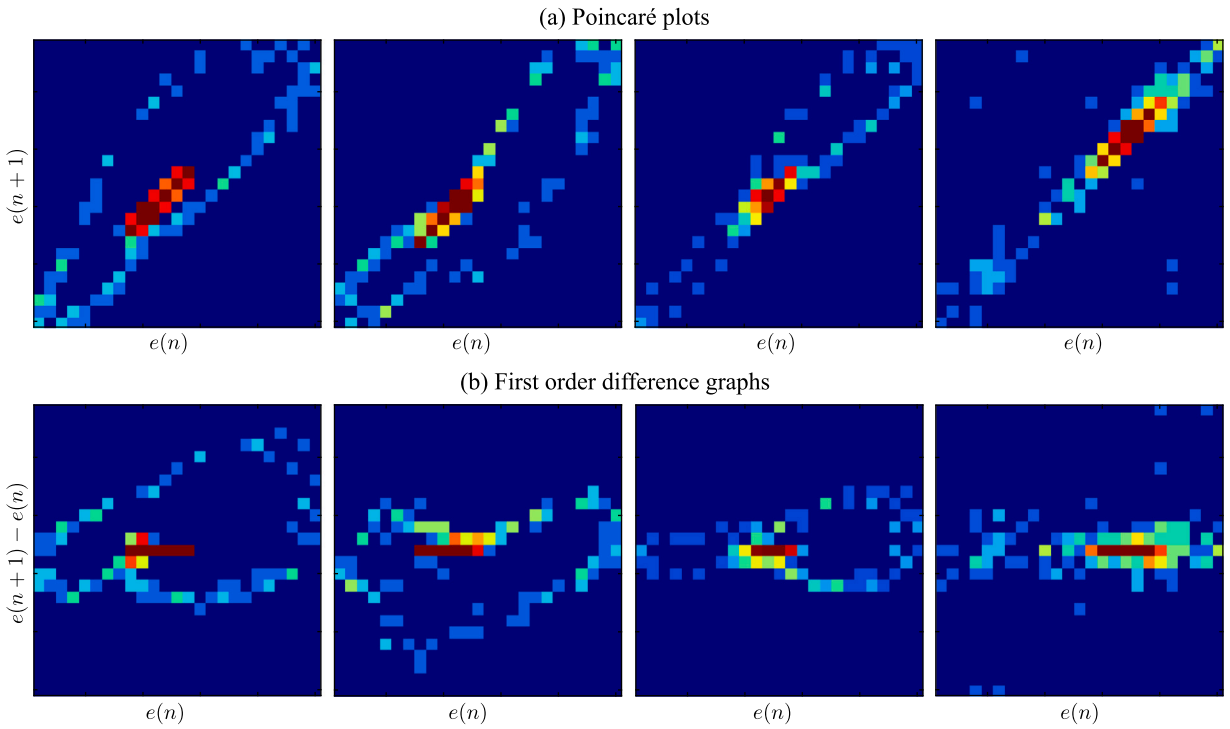


Fig. 3. Color images resulting for the parameter $C = 25$ from the PS representations displayed in (a) Fig. 1 and (b) Fig. 2, respectively.

into two stratified groups, so that 80% of samples were used for training and the remaining 20% for validation. Only in the case of deep learning models and with the idea of improving their learning and generalization ability, diversity of information in the training subset was increased using a well-established data augmentation approach. Precisely, the dataset was duplicated by randomly applying to every ECG-based image a rotation between −15° and 15° or a translation in x-direction or y-direction between −3 and 3 pixels. Moreover, an early stopping approach was used to avoid overtraining of these models [45]. Thus, the training process was stopped when validation accuracy (i.e., the percentage of correctly classified ECG excerpts on the validation subset) was not increased for 10 consecutive epochs. It should also be noted that a stochastic gradient descent algorithm and a constant learning rate of 0.01 on a batch size of 512 samples were used to train the proposed deep learning models based on PP and FODG images, whereas the parameters recommended by the authors in [11,16] were used for the Liu et al.'s and Zhao et al.'s methods, respectively.

Despite being highly unbalanced, the PC2017DB was used for external testing. The fact that this database was freely available makes comparison of the obtained results with others previously published easier and fairer. Furthermore, severe unbalance is not a major problem in the testing phase. Although the most common performance metric, i.e., Accuracy, usually shows overoptimistic inflated results on unbalanced datasets, alternative indices have been proposed to provide more truthful scores, including Balanced Accuracy (BAcc), F_1, and Matthews correlation coefficient (MCC) [46]. Assuming the positive class as the group of high-quality ECG excerpts and the negative class as the group of low-quality ECG intervals, and TP, TN, FN, and FP being the abbreviations of true positive, true negative, false negative, and false positive, respectively, these performance metrics, along with sensitivity (Se) and specificity (Sp), were computed as

$$Se = \frac{TP}{TP + FN}, \tag{10}$$

$$Sp = \frac{TN}{TN + FP}, \tag{11}$$

$$BAcc = \frac{Se + Sp}{2}, \tag{12}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \text{ and} \tag{13}$$

$$MCC = \left(1 + \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}\right)/2. \tag{14}$$

Moreover, to analyze how AF excerpts were discerned from noisy ECG segments, the rates of correctly classified NSR ($\mathcal{R}_{NSR}$), AF ($\mathcal{R}_{AF}$), and OR ($\mathcal{R}_{OR}$) intervals were also estimated within the high-quality group.

Finally, average time required by all the analyzed machine and deep learning models to classify a 5 second-length ECG sample in training, validation, and testing stages was also measured as an estimation of computational cost. Note that all the experiments were conducted with MATLAB 2023a, running on an HP workstation with 32 GB RAM memory, an Intel Xeon @ 3.60 GHz processor, and a graphic processing unit (GPU) GeForce GTX 1060 with 6 GB dedicated VRAM memory.

4. Results

To build the machine learning models, automated feature selection was conducted on the training dataset (i.e., 80% of the samples in the PDB) using a 10-fold cross-validation approach. This procedure was repeated 5 times with random splits to reduce the bias resulting from a single split of the PDB into training and validation subsets. In most iterations three features were selected for the two PS representations. The variables $\mathcal{D}$ and $\mathcal{E}$ were firstly chosen in both cases, whereas the third index was $\mathcal{S}_{12}$ for the PP and $\mathcal{P}$ for the FODG. The classification outcomes yielded by the resulting models of merging these variables on the validation (20% of the samples in the PDB) and testing (PC2017DB) datasets are presented in Table 3. As can be seen, for both datasets highly balanced values of Se and Sp were noticed in the case of using PP-based features. Moreover, all the performance metrics computed for

Table 3
Classification results obtained by the machine learning models on the validation and testing datasets.

PS portrait	Database	C	# Parameters	Se	Sp	BAcc	F_1	MCC	$\mathcal{R}_{NSR}$	$\mathcal{R}_{AF}$	$\mathcal{R}_{OR}$
PP	Validation	25	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.951	0.936	0.943	0.944	0.944	0.951	0.997	0.805
		45	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.949	0.936	0.942	0.943	0.943	0.950	0.994	0.808
		65	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.954	0.933	0.943	0.945	0.944	0.955	0.994	0.803
		85	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.949	0.934	0.941	0.942	0.942	0.950	0.994	0.808
	Testing	25	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.847	0.834	0.841	0.915	0.639	0.873	0.843	0.800
		45	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.843	0.826	0.835	0.913	0.636	0.871	0.843	0.791
		65	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.851	0.825	0.838	0.917	0.639	0.879	0.842	0.798
		85	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{S}_{12}$)	0.855	0.813	0.834	0.920	0.639	0.882	0.851	0.805
FODG	Validation	25	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.933	0.883	0.908	0.911	0.909	0.933	0.979	0.800
		45	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.929	0.880	0.904	0.907	0.905	0.928	0.982	0.783
		65	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.926	0.877	0.902	0.905	0.902	0.922	0.991	0.800
		85	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.925	0.868	0.897	0.900	0.897	0.922	0.994	0.775
	Testing	25	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.802	0.872	0.837	0.889	0.626	0.829	0.793	0.753
		45	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.800	0.865	0.833	0.888	0.624	0.828	0.786	0.752
		65	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.777	0.862	0.820	0.873	0.615	0.802	0.787	0.725
		85	3 ($\mathcal{D}$, $\mathcal{E}$, $\mathcal{P}$)	0.762	0.849	0.805	0.863	0.608	0.787	0.756	0.715

these PP-based models obtained quite stable values regardless of the grid size C. Thus, for the analyzed values of C = 25, 45, 65, and 85, the metrics Se, Sp, BAcc, F_1, and MCC always remained about 0.950, 0.935, 0.940, 0.945, and 0.945 on the validation dataset and about 0.850, 0.830, 0.840, 0.920, and 0.640 on the testing database, respectively. Within the group of high-quality ECG excerpts, the rates of properly identified NSR, AF, and OR intervals were also steady around 0.950, 0.995, and 0.805 on the validation dataset and around 0.880, 0.840, and 0.800 on the testing database, respectively. A similar behavior was also noticed for the models combining FODG-based features. However, a slightly lower global performance was outlined by the metrics BAcc, F_1, and MCC both in validation and testing phases. Moreover, in this case a difference between the values of Se and Sp of about 5%–10% was also noticed, as well as a mild trend towards poorer classification outcomes when the grid size decreased (i.e., C increased). Additionally, the rates $\mathcal{R}_{NSR}$ and $\mathcal{R}_{AF}$ were about 10% lower than for the models based on PP features in both validation and testing stages, but the index $\mathcal{R}_{OR}$ was only about 5%–7% lower in the case of testing.

Regarding the classification results obtained by the deep learning models, the averaged performance metrics for the 5 iterations conducted on the validation and testing databases are displayed in Table 4. As for the machine learning algorithms, in this case the models inputted with the PP-based images exhibited a classification performance stable for every value of C, as well as well-balanced in terms of Se and Sp. A similar behavior was also noticed for the deep learning models fed with FODG-based images, additionally reporting comparable values in most performance metrics. Indeed, for both cases, values of BAcc, F_1 and MCC about 0.970 were noticed on the validation database and about 0.830, 0.910, and 0.630 on the testing dataset, respectively. Within the group of high-quality ECG excerpts, comparable results were also observed for all the models based on both kinds of PS images, but a trend towards a better classification of OR episodes was seen for those fed with FODG-based images. To this respect, the rate $\mathcal{R}_{OR}$ reported by the FODG-based models on the validation dataset was about 4%–5% higher than that obtained by the PP-based algorithms.

The classification outcomes obtained by the four previous ECG quality assessment algorithms implemented for comparison purposes are presented in Table 5. On the validation dataset, Behar et al.'s [21], Liu et al.'s [11], and Zhao et al.'s [16] methods yielded a very high performance, with values of BAcc, F_1, and MCC larger than 0.960, whereas the Albaba et al.'s method [29] reported more modest values about 0.900. However, a notable reduction in the discriminant power on the testing dataset was observed for the four techniques, reporting values of BAcc, F_1, and MCC lower than 0.82, 0.888, and 0.62, respectively. Moreover, larger unbalances between values of Se and Sp than in validation were also noticed, especially in the case of Behar et al.'s method [21]. As well, it is worth noting that in most cases these outcomes were between 2% and 10%, between 3% and 9%, and between 2% and 5% poorer in terms of BAcc, F_1, and MCC, respectively, than those obtained by the proposed machine and deep learning algorithms based on PP and FODG features and images. Similarly, within the group of high-quality ECG excerpts previous methods also reported lower rates of $\mathcal{R}_{NSR}$, $\mathcal{R}_{AF}$, $\mathcal{R}_{OR}$ between 4% and 16%, between 2% and 16%, and between 1 and 15%, respectively.

Finally, Table 6 presents average time spent by all the analyzed algorithms on a 5 second-length ECG excerpt in training, validation and testing. As can be seen, the deep learning models fed with both PP and FODG images required a similar training time, which increased as a function of C. This time was much higher (approximately between 2 and 31 times) than that required for training the machine learning algorithms based on both PP and FODG features. Contrarily, the time spent in classifying each sample on validation and testing was similar for most of these machine and deep learning algorithms, because it was between 1 and 2 ms in all the cases. Moreover, this time was notably lower (approximately between 3 and 145 times) than that required by the previous techniques implemented for comparison purposes, both in validation and testing stages. In the same line, all these previous methods except the Behar et al.'s one [21] also needed more training time.

5. Discussion

5.1. Main findings

In the last years, diverse kinds of PS representations have been widely used to characterize the response of physiological systems [32,47], as well as to detect events and diseases [48,49] and extract clinically relevant information from biomedical signals [19,20]. However, to the best of our knowledge, this is one of the earliest works proving the usefulness of two common PS portraits, such as the PP and FODG, for quality assessment of single-lead ECG recordings. Only one study combining a few hand-crafted, FODG-based features with an SVM has been previously proposed to discern among ECG excerpts artificially disturbed with different levels of noise [50]. Besides not considering real-world noisy ECG signals contaminated during their acquisition, a limited database without patients suffering from AF and other supraventricular arrhythmias was only analyzed in that pilot study. Hence, the present work has conducted a broader, more systematic and robust experimentation about PS reconstruction of the ECG for its quality evaluation, introducing novel PS-based features and images for feeding conventional machine learning and advanced deep learning classifiers.

In a straightforward comparison of the same experimental protocol and datasets, the combination of manually and automatically derived features from the proposed gridded versions of the PP and FODG reported notably better binary classification outcomes between high-

Table 4
Classification results obtained by the deep learning models on the validation and testing datasets.

PS portrait	Database	C	# Parameters	Se	Sp	BAcc	F_1	MCC	$\mathcal{R}_{NSR}$	$\mathcal{R}_{AF}$	$\mathcal{R}_{OR}$
PP	Validation	25	155,422	0.976	0.975	0.976	0.976	0.976	0.988	0.994	0.775
		45	155,422	0.979	0.972	0.976	0.976	0.976	0.991	0.994	0.788
		65	155,422	0.981	0.954	0.967	0.969	0.968	0.991	0.994	0.817
		85	155,422	0.982	0.948	0.965	0.966	0.965	0.993	0.994	0.804
	Testing	25	155,422	0.813	0.854	0.833	0.895	0.627	0.793	0.792	0.803
		45	155,422	0.851	0.826	0.838	0.917	0.640	0.832	0.827	0.841
		65	155,422	0.853	0.835	0.844	0.919	0.643	0.834	0.832	0.842
		85	155,422	0.817	0.838	0.827	0.897	0.626	0.800	0.788	0.806
FODG	Validation	25	155,422	0.982	0.966	0.974	0.974	0.974	0.990	0.994	0.833
		45	155,422	0.985	0.973	0.979	0.979	0.979	0.994	0.994	0.838
		65	155,422	0.978	0.975	0.977	0.977	0.977	0.989	0.994	0.788
		85	155,422	0.987	0.941	0.964	0.965	0.965	0.995	0.994	0.867
	Testing	25	155,422	0.850	0.819	0.834	0.917	0.638	0.831	0.835	0.839
		45	155,422	0.837	0.840	0.838	0.909	0.636	0.818	0.817	0.826
		65	155,422	0.816	0.854	0.835	0.896	0.630	0.798	0.790	0.805
		85	155,422	0.828	0.843	0.835	0.904	0.632	0.810	0.801	0.817

Table 5
Classification results obtained by the previous ECG quality assessment algorithms implemented for comparison purposes on the validation and testing datasets.

Database	Algorithm	# Parameters	Se	Sp	BAcc	F_1	MCC	$\mathcal{R}_{NSR}$	$\mathcal{R}_{AF}$	$\mathcal{R}_{OR}$
Validation	Behar et. al. [21]	7	0.971	0.959	0.965	0.965	0.965	0.978	0.988	0.825
	Albaba et. al. [29]	7	0.878	0.925	0.901	0.900	0.902	0.887	0.882	0.758
	Liu et. al. [11]	27,532	0.971	0.960	0.966	0.966	0.966	0.978	0.999	0.800
	Zhao et. al. [16]	35,394	0.984	0.998	0.991	0.991	0.991	0.990	0.984	0.900
Testing	Behar et. al. [21]	7	0.712	0.890	0.801	0.830	0.600	0.723	0.710	0.690
	Albaba et. al. [29]	7	0.739	0.804	0.771	0.847	0.593	0.752	0.685	0.729
	Liu et. al. [11]	27,532	0.791	0.719	0.755	0.880	0.594	0.806	0.830	0.750
	Zhao et. al. [16]	35,394	0.772	0.860	0.816	0.868	0.617	0.774	0.759	0.762

Table 6
Average time required by all the analyzed algorithms on a 5 second-length ECG excerpt in training, validation, and testing. The values are expressed in seconds.

Algorithm	PS portrait	C	Computational time		
			Training	Validation	Testing
Machine learning	PP	25	0.002	0.001	0.001
		45	0.002	0.001	0.001
		65	0.002	0.001	0.001
		85	0.002	0.001	0.001
	FODG	25	0.002	0.001	0.001
		45	0.002	0.001	0.001
		65	0.002	0.001	0.001
		85	0.002	0.001	0.001
Deep learning	PP	25	0.007	0.001	0.001
		45	0.030	0.001	0.001
		65	0.050	0.002	0.002
		85	0.053	0.002	0.002
	FODG	25	0.006	0.001	0.001
		45	0.023	0.001	0.001
		65	0.065	0.002	0.002
		85	0.073	0.002	0.002
Behar et. al. [21]	–	–	0.008	0.007	0.007
Albaba et. al. [29]	–	–	0.086	0.084	0.085
Liu et. al. [11]	–	–	0.084	0.028	0.028
Zhao et. al. [16]	–	–	0.167	0.145	0.145

and low-quality ECG excerpts than four well-known previous algorithms, i.e., those proposed by Behar et al. [21], Albaba et al. [29], Liu et al. [11], and Zhao et al. [16]. Precisely, the machine and deep learning models built from PP- and FODG-based features and images provided values for the global performance metrics BAcc, F_1, and MCC about 2%–10% greater than those previous methods (see Tables 3–5), then reaching the best results on the external testing dataset of about 84%, 90%, and 64%, respectively. Likewise, they also reported about 2%–16% greater rates in well classification of NSR, AF, and OR episodes within the group of high-quality ECG segments. These findings suggest that the proposed PS-based techniques achieved the best trade-off between sufficient generalization to disregard intrinsic inter-patient variability in morphology and amplitude presented by high-quality ECG excerpts and high sensitivity to discern most of the transient dynamics associated with noises and artifacts from those linked to high-quality NSR, AF, and OR segments. To this respect, Figs. 1 and 2 show how these three rhythms resulted in similar attractors both in the PP and FODG, whereas a completely different distribution of points was seen for the low-quality excerpt.

It should also be noted that the proposed PS-based methods required less or comparable time for training than three out of the four previous algorithms implemented for comparison purposes (see Table 6). Moreover, once trained, they provided a classification of ECG excerpts in validation and testing stages notably quicker than all the previous techniques. Both outcomes could be explained by the fact that reconstruction of the PP and FODG do not require mathematical computation, whereas some hand-crafted features employed by previous algorithms, such as approximate entropy, as well as most time–frequency transformations are computationally cost [51,52]. Overall, ECG transformation to PP- and FODG-based portraits appears to reveal novel information of great utility for ECG quality assessment, enabling the development of both machine and deep learning models able to classify high- and low-quality ECG excerpts in a simple, efficient, and computational low-cost way.

5.2. Indirect comparison with other previous works

Beyond these last findings derived by direct comparison with the four previous ECG quality assessment indices implemented to serve as a reference, it is worth noting that the classification outcomes yielded by the proposed machine and deep learning models based on PP and FODG features and images were also comparable or slightly lower than those presented by the majority of previous works, which have reported accuracy values of 90% or above [9,10]. However, every indirect comparison of results obtained on different datasets should be considered with caution. Thus, many previous works only dealt

with a small amount of ECG signals, acquired from a homogeneous group of healthy subjects with a single recording system [9,10]. Indeed, most analyzed a freely available database that was specifically designed for quality assessment of 12-lead ECG signals and contains about 1500 excerpts of 10 s in length [28]. In contrast, more than 68,000 single-lead, 5-second length ECG excerpts with a wide variety of morphologies, noises, and artifacts were examined in the present work. They were acquired with two different wearable devices from diverse body positions. While the PDB was formed by ECG recordings obtained from a non-standard lead with a textile Holter placed on the patient's thorax, the PC2017DB collected signals captured from an equivalent lead I between the two patient's hands with a portable ECG monitor [27]. Moreover, both datasets included ECG recordings from healthy subjects, but also from patients with AF and other supraventricular arrhythmias. It is well-known that these cardiac disorders often provoke abnormal atrial waves that can be easily confounded with high frequency noise [21,22], thus involving a harder challenge than quality evaluation of ECG recordings exclusively obtained from healthy subjects. Actually, many previous algorithms presented an excellent accuracy in assessing quality of ECG signals from healthy subjects, but their performance was drastically reduced by 10%–40% in patients with AF and OR. This behavior was observed both for machine learning algorithms based on combining diverse kinds of hand-crafted features [53], including the Behar et al.'s method [21], as well as for deep learning methods based on 1-D and 2-D CNN schemes with sequential and parallel architectures [15,54].

Of note is also that resubstitution validation and ECG segment-wise cross-validation approaches were mainly used to estimate the proposed algorithm's performance in most previous works [9,10]. Although a more unbiased and general overview of the performance of a model is obtained with cross-validation than with resubstitution validation (i.e., when the method is trained and tested on the same dataset), the inclusion of ECG excerpts from the same patients in training and testing subsets during cross-validation loops often inflates classification results because the models are able to memorize subject-specific features [41,42]. For instance, whereas Albaba et al.'s method achieved values of BAcc about 90% on three separate databases using ECG segment-wise cross-validation, the same performance metric fell to 50% when a database was used for training and the remaining two for testing [29]. Similarly, the Liu et al.'s technique also reported values of accuracy near 99% and less than 80% when ECG segment-wise cross-validation and external validation were respectively used [11]. In contrast, as recommended by TRIPOD guidelines [43], all the results obtained in the present work were obtained by external validation using two separate databases for training and testing, and they could therefore be considered more robust than those presented by most previous works. This finding together with the high values of BAcc, F_1, and MCC obtained on the testing subset suggest that the proposed machine and deep learning models based on PP and FODG features and images achieved heavy levels of generalization when discerning between high- and low-quality ECG excerpts.

Another interesting advantage of these proposed PS-based methods regarding many previous approaches, both based on machine learning classifiers [9,10] and CNN-based schemes [55], is their ability to deal directly with the preprocessed ECG recording. Indeed, no delineation of fiducial points and detection of R peaks were required, reducing the impact of the frequent errors associated with these procedures, especially in presence of artifacts and QRS complexes with abnormal morphologies [56]. Moreover, as previously mentioned, PP and FODG reconstructions do not require any kind of mathematical transformation, thus making the proposed algorithms more easily interpretable and computationally efficient than most of the previous ECG quality indices found in the literature.

5.3. Comparison between PS representations

Comparing the obtained results by the two analyzed PS representations, the machine learning models based on PP features reported a more stable performance, as a function of the grid cell's size C, than those based on FODG variables both for the validation and testing datasets (see Table 3). This finding might be motived by the fact that the features S_1 and D are by definition independent and the index $\mathcal{E}$ slightly dependent on the used grid mapping approach [50]. Contrarily, the very definition of the variable $\mathcal{P}$ involves an intrinsic variability as a function of C, which could explain the gently greater differences noticed in the results obtained by the machine learning models based on the FODG features. Thus, although the relative position of the largest accumulation of points within the $C \times C$ matrix should always remain in the same region, the specific location of the cell containing the greatest number of points will be strongly dependent on its size (i.e., on the value of C).

Contrarily, all the deep learning models based on both PP and FODG images yielded a very similar performance on the validation and testing datasets for every value of C (see Table 4). Moreover, they presented classification metrics on the testing dataset totally comparable to those reported by the machine learning algorithms. To this respect, differences in the metrics BAcc, F_1, and MCC provided by all the machine and deep learning algorithms were lower than 2%. In the case of classification of the different rhythms within the high-quality group, a slightly different trend was observed for both kinds of models on the testing database. Whereas the deep learning algorithms presented well-balanced classification rates of about 80%–83% for NSR, AF, and OR episodes, the machine learning methods tended to classify NSR and AF intervals mildly better than OR excerpts. Anyway, it should be noted that the main goal of the present work was to explore the usefulness of two common PS representations for ECG quality assessment and not to compare the performance of diverse machine and deep learning techniques. In fact, to work even in scenarios with limited resources of memory, battery, and processing, a lightweight CNN architecture was only considered. The use of deeper CNN schemes and transfer learning from diverse pre-training stages could improve classification between high- and low-quality ECG intervals [12,57]. However, these aspects will be tackled in the future.

Another interesting point to highlight is that the deep learning models initially trained on the PDB were only able to provide values of Sp about 15%–20% larger than Se, when externally tested on the PC2017DB. As in many clinical scenarios [58], this situation was undesirable because it increased the probability of misinterpreting brief AF episodes as noisy excerpts. Hence, to compensate the rates of false positives and false negatives and achieve the well-balanced results displayed in Table 4, training of all the deep learning models was improved by duplicating the PDB via data augmentation. In this way, the training dataset remained balanced for both groups of high- and low-quality ECG excerpts, but more diverse PP- and FODG-based attractors were offered for the models' learning. In fact, it is reasonable to think that the improvement in detection of high-quality ECG excerpts (i.e., Se) noticed for all the models was due to the increase of such diversity of attractors rather than to the increase in the number of images themselves. In general terms, clean ECG excerpts show great similarity and scarce morphological variability, thus leading to highly similar distributions of points both in the PP and FODG. On the contrary, noisy ECG segments inherently involve a significantly larger pool of different and chaotic morphologies, which might explain why high values of Sp were even achieved without the need for data augmentation.

Beyond the PP and FODG, other PS representations based on the Takens' delay embedding theorem can also be found in the literature. However, the former alternatives were considered as a first approximation to study the usefulness of transforming ECG into simple and free of tunable parameters PS representations for its quality evaluation. To this respect, values of $m = 2$ and $\tau = 1$ are well-established for the PP

reconstruction [35], while the FODG presents the additional advantage of removing every dependency with the parameter τ [34]. On the contrary, another well-known PS reconstruction, such as the recurrence plot, involves the need to define more parameters in addition to m and τ. Indeed, this graph represents the times when a time series roughly recurs the same area in the PS, and therefore it is mandatory to define the concept of recurrence (i.e., when two points in the PS are sufficiently close to be considered as visiting the same area) [59]. Although a variety of methods have been proposed to optimize the parameters required by this PS representation, they have provided highly variable values as a function of the ECG-based application [60,61]. Nonetheless, recent classification of recurrent plots with a 2-D CNN scheme has provided promising results in ECG-based detection of AF and other cardiac arrhythmias [62,63], and that approach could also be expected to successfully work in ECG quality assessment. Thus, further experiments to this respect will be conducted in the future.

5.4. Limitations

A binary classification between high- and low-quality ECG excerpts was only considered in the present study to minimize the impact of the implicit subjectivity in grading ECG quality [21] on the performance of the analyzed models. Given that there are no standard and strict limits to discern among several levels of ECG quality and previous works have used diverse criteria to establish three or more levels [64], the most objective criterion possible for a binary categorization of the ECG excerpts from the PDB was chosen on the basis of whether R peaks could be clearly detected. In this way, accurate subsequent analysis of the heart ventricular response and its variability may only be ensured, since other ECG waveforms and intervals (e.g., P-wave, T-wave, TQ-interval, etc.) might still be corrupted by noise. Nonetheless, heart rate variability analysis is often sufficient to detect most cardiac arrhythmias, including AF, on continuous ECG monitoring [65,66]. Moreover, this analysis of R peaks is today predominant on long-term ECG recordings, because P- and T-waves are often drowned out by noise in a large part of the signal acquired from most wearable devices [67].

Another limitation of the developed experimental setup is that the segmentation approach of the ECG recordings from the PC2017DB might have led to mislabel some 5 second-length excerpts. Although the duration of the ECG signals ranged between 9 and 60 s, a single quality label was assigned to each one. Thus, in ECG signals labeled as noisy by the presence of a highly localized artifact in time, it could only affect a few ECG excerpts into the recording and the remaining ones would present high-quality but be erroneously classified as low-quality. These ECG excerpts would have had a negative impact on the performance of all the tested models, leading to undervalued figures. However, for the sake of comparison with other works, a relabeling of these ECG excerpts with the same criteria used in the PDB was not conducted. Finally, although all the proposed PS-based methods revealed high generalization ability on an external dataset, they were trained on ECG excerpts acquired from only a non-standard lead. Thus, to extend the pool of ECG patterns both in training and testing stages, further experiments will be conducted on signals obtained from paroxysmal AF patients with a variety of conventional and wearable ECG acquisition systems, where standard and non-standard leads will be recorded.

6. Conclusions

The transformation of the single-lead ECG into two common phase space portraits, such as the Poincaré plot and first order difference graph, has resulted to be highly useful for quality assessment of the signal using classification algorithms based both on machine and deep learning concepts. Novel features and images derived from the grid mapping of these phase space representations have led to train models able to notably improve the performance of well-known and widely used previous algorithms in the challenging context of paroxysmal AF. These features and images present the advantage of being obtained from the preprocessed ECG signal without requiring mathematical transformations and detection of its fiducial points, thus avoiding the inherent errors introduced by such approaches, minimizing the cost of their computation, and facilitating their implementation in wearable recording systems. Hence, their use, along with machine and deep learning classifiers, for automated quality assessment of single-lead ECG signals might provide a definitive boost to the deployment of wearable devices in continuous cardiac monitoring of patients suffering from paroxysmal AF and other supra-ventricular arrhythmias.

CRediT authorship contribution statement

Álvaro Huerta: Investigation, Methodology, Software, Writing – original draft. **Arturo Martinez-Rodrigo:** Investigation, Methodology. **Vicente Bertomeu-González:** Resources, Supervision. **Óscar Ayo-Martin:** Resources, Validation. **José J. Rieta:** Supervision. **Raúl Alcaraz:** Project administration, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research has received financial support from Daiichi Sankyo SLU and from public grants PID2021-00X128525-IV0, PID2021-12380 4OB-I00, and TED2021-130935B-I00 of the Spanish Government 10.13039/501100011033 jointly with the European Regional Development Fund (EU), SBPLY/21/ 180501/000186 from Junta de Comunidades de Castilla-La Mancha, Spain, and AICO/2021/286 from Generalitat Valenciana.

References

[1] Cardiovascular diseases, 2023, Available online: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), accessed on 23rd March 2023.
[2] A. Timmis, N. Townsend, C.P. Gale, A. Torbica, M. Lettino, S.E. Petersen, E.A. Mossialos, A.P. Maggioni, D. Kazakiewicz, H.T. May, D. De Smedt, M. Flather, L. Zuhlke, J.F. Beltrame, R. Huculeci, L. Tavazzi, G. Hindricks, J. Bax, B. Casadei, S. Achenbach, L. Wright, P. Vardas, European Society of Cardiology, European society of cardiology: Cardiovascular disease statistics 2019, Eur. Heart J. 41 (1) (2020) 12–85.
[3] A. Odutayo, C.X. Wong, A.J. Hsiao, S. Hopewell, D.G. Altman, C.A. Emdin, Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis, BMJ 354 (2016) i4482.
[4] G. Lippi, F. Sanchis-Gomar, G. Cervellin, Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge, Int. J. Stroke 16 (2) (2021) 217–221.
[5] B.J.J.M. Brundel, X. Ai, M.T. Hills, M.F. Kuipers, G.Y.H. Lip, N.M.S. de Groot, Atrial fibrillation, Nat. Rev. Dis. Primers 8 (1) (2022) 21.
[6] S. Blum, S. Aeschbacher, P. Meyre, L. Zwimpfer, T. Reichlin, J.H. Beer, et al., Incidence and predictors of atrial fibrillation progression, J. Am. Heart Assoc. 8 (20) (2019) e012554.
[7] Z. Kalarus, G.H. Mairesse, A. Sokal, G. Boriani, B. Średniawa, R. Casado-Arroyo, et al., Searching for atrial fibrillation: looking harder, looking longer, and in increasingly sophisticated ways. an EHRA position paper, Europace 25 (1) (2023) 185–198.
[8] E.Y. Ding, G.M. Marcus, D.D. McManus, Emerging technologies for identifying atrial fibrillation, Circ. Res. 127 (1) (2020) 128–142.
[9] U. Satija, B. Ramkumar, M.S. Manikandan, A review of signal processing techniques for electrocardiogram signal quality assessment, IEEE Rev. Biomed. Eng. 11 (2018) 36–52.

[10] K. van der Bijl, M. Elgendi, C. Menon, Automatic ECG quality assessment techniques: A systematic review, Diagnostics (Basel) 12 (11) (2022).
[11] F. Liu, S. Xia, S. Wei, L. Chen, Y. Ren, X. Ren, Z. Xu, S. Ai, C. Liu, Wearable electrocardiogram quality assessment using Wavelet scattering and LSTM, Front. Physiol. 13 (2022) 905447.
[12] Á. Huerta, A. Martínez-Rodrigo, V. Bertomeu-González, A. Quesada, J.J. Rieta, R. Alcaraz, A deep learning approach for featureless robust quality assessment of intermittent atrial fibrillation recordings from portable and wearable devices, Entropy (Basel) 22 (7) (2020).
[13] J. Chai, H. Zeng, A. Li, E.W. Ngai, Deep learning in computer vision: A critical review of emerging techniques and application scenarios, Mach. Learn. Appl. 6 (2021) 100134.
[14] T.J. Jun, H.M. Nguyen, D. Kang, et al., ECG arrhythmia classification using a 2-D convolutional neural network, 2018, arXiv preprint arXiv:1804.06812.
[15] Q. Zhang, L. Fu, L. Gu, A cascaded convolutional neural network for assessing signal quality of dynamic ECG, Comput. Math. Methods Med. 2019 (2019) 7095137.
[16] Z. Zhao, C. Liu, Y. Li, Y. Li, J. Wang, B.-S. Lin, J. Li, Noise rejection for wearable ECGs using modified frequency slice wavelet transform and convolutional neural networks, IEEE Access 7 (2019) 34060–34067.
[17] G. Liu, X. Han, L. Tian, W. Zhou, H. Liu, ECG quality assessment based on hand-crafted statistics and deep-learned S-transform spectrogram features, Comput. Methods Programs Biomed. 208 (2021) 106269.
[18] B.K. Pradhan, B.C. Neelappu, J. Sivaraman, D. Kim, K. Pal, et al., A review on the applications of time-frequency methods in ECG analysis, J. Healthc. Eng. 2023 (2023).
[19] A.J. Dunn, M.H. ElRefai, P.R. Roberts, S. Coniglio, B.M. Wiles, A.B. Zemkoho, Deep learning methods for screening patients' S-ICD implantation eligibility, Artif. Intell. Med. 119 (2021) 102139.
[20] H.-L. Chan, H.-W. Chang, W.-Y. Hsu, P.-J. Huang, S.-C. Fang, Convolutional neural network for individual identification using phase space reconstruction of electrocardiogram, Sensors 23 (6) (2023) 3164.
[21] J. Behar, J. Oster, Q. Li, G.D. Clifford, ECG signal quality during arrhythmia and its application to false alarm reduction, IEEE Trans. Biomed. Eng. 60 (6) (2013) 1660–1666.
[22] S.K. Bashar, E. Ding, A.J. Walkey, D.D. McManus, K.H. Chon, Noise detection in electrocardiogram signals for intensive care unit patients, IEEE Access 7 (2019) 88357–88368.
[23] M.R. Afzal, J. Mease, T. Koppert, T. Okabe, J. Tyler, M. Houmsse, R.S. Augostini, R. Weiss, J.D. Hummel, S.J. Kalbfleisch, E.G. Daoud, Incidence of false-positive transmissions during remote rhythm monitoring with implantable loop recorders, Heart Rhythm 17 (1) (2020) 75–80.
[24] B.J. Drew, P. Harris, J.K. Zègre-Hemsey, T. Mammone, D. Schindler, R. Salas-Boni, Y. Bai, A. Tinoco, Q. Ding, X. Hu, Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients, PLoS One 9 (10) (2014) e110274.
[25] X. Zhang, J. Li, Z. Cai, L. Zhao, C. Liu, Deep learning-based signal quality assessment for wearable ECGs, IEEE Instrum. Meas. Mag. 25 (5) (2022) 41–52.
[26] J. Ródenas, M. García, R. Alcaraz, J.J. Rieta, Combined nonlinear analysis of atrial and ventricular series for automated screening of atrial fibrillation, Complexity 2017 (2017).
[27] G.D. Clifford, C. Liu, B. Moody, H.L. Li-wei, I. Silva, Q. Li, A. Johnson, R.G. Mark, AF classification from a short single lead ECG recording: The PhysioNet/Computing in cardiology challenge 2017, in: 2017 Computing in Cardiology (CinC), 2017, pp. 1–4.
[28] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) E215–20.
[29] A. Albaba, N. Simões-Capela, Y. Wang, R.C. Hendriks, W. De Raedt, C. Van Hoof, Assessing the signal quality of electrocardiograms from varied acquisition sources: A generic machine learning pipeline for model generation, Comput. Biol. Med. 130 (2021) 104164.
[30] G. Lenis, N. Pilia, A. Loewe, W.H.W. Schulze, O. Dössel, Comparison of baseline wander removal techniques considering the preservation of ST changes in the ischemic ECG: A simulation study, Comput. Math. Methods Med. 2017 (2017) 9295029.
[31] M. García, M. Martínez-Iniesta, J. Ródenas, J.J. Rieta, R. Alcaraz, A novel wavelet-based filtering strategy to remove powerline interference from electrocardiograms with atrial fibrillation, Physiol. Meas. 39 (11) (2018) 115006.
[32] J. de Pedro-Carracedo, D. Fuentes-Jimenez, A.M. Ugena, A.P. Gonzalez-Marcos, Phase space reconstruction from a biological time series: A photoplethysmographic signal case study, Appl. Sci. 10 (4) (2020) 1430.
[33] S. Roy, D.P. Goswami, A. Sengupta, Geometry of the Poincaré plot can segregate the two arms of autonomic nervous system - a hypothesis, Med. Hypotheses 138 (2020) 109574.
[34] Y. Li, X. Tang, Z. Xu, H. Yan, A novel approach to phase space reconstruction of single lead ECG for QRS complex detection, Biomed. Signal Process. Control 39 (2018) 405–415.
[35] M. Brennan, M. Palaniswami, P. Kamen, Do existing measures of Poincaré plot geometry reflect nonlinear features of heart rate variability? IEEE Trans. Biomed. Eng. 48 (11) (2001) 1342–1347.
[36] C. Yan, P. Li, C. Liu, X. Wang, C. Yin, L. Yao, Novel gridded descriptors of Poincaré plot for analyzing heartbeat interval time-series, Comput. Biol. Med. 109 (2019) 280–289.
[37] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.
[38] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (2021) 1–74.
[39] M. Gu, Y. Zhang, Y. Wen, G. Ai, H. Zhang, P. Wang, G. Wang, A lightweight convolutional neural network hardware implementation for wearable heart rate anomaly detection, Comput. Biol. Med. 155 (2023) 106623.
[40] S. Somani, A.J. Russak, F. Richter, S. Zhao, A. Vaid, F. Chaudhry, J.K. De Freitas, N. Naik, R. Miotto, G.N. Nadkarni, J. Narula, E. Argulian, B.S. Glicksberg, Deep learning and the electrocardiogram: review of the current state-of-the-art, Europace 23 (8) (2021) 1179–1191.
[41] M.A. Little, G. Varoquaux, S. Saeb, L. Lonini, A. Jayaraman, D.C. Mohr, K.P. Kording, Using and understanding cross-validation strategies. perspectives on Saeb et al., GigaScience 6 (5) (2017) 1–6.
[42] S. Saeb, L. Lonini, A. Jayaraman, D.C. Mohr, K.P. Kording, The need to approximate the use-case in clinical machine learning, Gigascience 6 (5) (2017) gix019.
[43] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, TRIPOD Group, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, Circulation 131 (2) (2015) 211–219.
[44] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (1) (2019) 1–54.
[45] C.W. Bartlett, J. Bossenbroek, Y. Ueyama, P. McCallinhart, O.A. Peters, D.A. Santillan, M.K. Santillan, A.J. Trask, W.C. Ray, Invasive or more direct measurements can provide an objective early-stopping ceiling for training deep neural networks on non-invasive or less-direct biomedical data, SN Comput. Sci. 4 (2) (2023) 1–12.
[46] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (1) (2020) 6.
[47] E. Brzozowska, M. Borowska, Selection of phase space reconstruction parameters for EMG signals of the uterus, Stud. Logic Gramm. Rhetor. 47 (1) (2016) 47–59.
[48] N. Ilakiyaselvan, A.N. Khan, A. Shahina, Reconstructed phase space portraits for detecting brain diseases using deep learning, Biomed. Signal Process. Control 71 (2022) 103278.
[49] M. Chen, Y. Fang, X. Zheng, Phase space reconstruction for improving the classification of single trial EEG, Biomed. Signal Process. Control 11 (2014) 10–16.
[50] Y. Li, X. Tang, Grid mapping: a novel method of signal quality evaluation on a single lead electrocardiogram, Australas. Phys. Eng. Sci. Med. 40 (2017) 895–907.
[51] G. Manis, M. Aktaruzzaman, R. Sassi, Low computational cost for sample entropy, Entropy 20 (1) (2018) 61.
[52] L.P. Arts, E.L. van den Broek, The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time–frequency analysis, Nat. Comput. Sci. 2 (1) (2022) 47–58.
[53] U. Satija, B. Ramkumar, M.S. Manikandan, Automated ECG noise detection and classification system for unsupervised healthcare monitoring, IEEE J. Biomed. Health Inform. 22 (2018) 722–732.
[54] D. Yoon, H.S. Lim, K. Jung, T.Y. Kim, S. Lee, Deep learning-based electrocardiogram signal noise detection and screening model, Healthc. Inform. Res. 25 (3) (2019) 201–211.
[55] H. Halvaei, E. Svennberg, L. Sörnmo, M. Stridh, Identification of transient noise to reduce false detections in screening for atrial fibrillation, Front. Physiol. 12 (2021) 672875.
[56] H. Dogan, R.O. Dogan, A comprehensive review of computer-based techniques for R-peaks/QRS complex detection in ECG signal, Arch. Comput. Methods Eng. 1 (2023) 1–19.
[57] K. Weimann, T.O. Conrad, Transfer learning for ECG classification, Sci. Rep. 11 (1) (2021) 1–12.
[58] L.G. Portney, Foundations of Clinical Research: Applications to Evidence-Based Practice, FA Davis, 2020, chapter 33 - Diagnostic Accuracy.
[59] N. Marwan, K.H. Kraemer, Trends in recurrence analysis of dynamical systems, Eur. Phys. J. Spec. Top. 232 (1) (2023) 5–27.
[60] H. Ding, S. Crozier, S. Wilson, Optimization of Euclidean distance threshold in the application of recurrence quantification analysis to heart rate variability studies, Chaos Solitons Fractals 38 (5) (2008) 1457–1467.
[61] S. Martín-González, J.L. Navarro-Mesa, G. Juliá-Serdá, G.M. Ramírez-Ávila, A.G. Ravelo-García, Improving the understanding of sleep apnea characterization using recurrence quantification analysis by defining overall acceptable values for the dimensionality of the system, the delay, and the distance threshold, PLoS One 13 (4) (2018) e0194462.

[62] H. Zhang, C. Liu, Z. Zhang, Y. Xing, X. Liu, R. Dong, Y. He, L. Xia, F. Liu, Recurrence plot-based approach for cardiac arrhythmia classification using inception-ResNet-v2, Front. Physiol. 12 (2021) 648950.
[63] H. Zhang, C. Liu, F. Tang, M. Li, D. Zhang, L. Xia, S. Crozier, H. Gan, N. Zhao, W. Xu, et al., Atrial fibrillation classification based on the 2D representation of minimal subset ECG and a non-deep neural network, Front. Physiol. 14 (2023) 182.
[64] J. Xie, L. Peng, L. Wei, Y. Gong, F. Zuo, J. Wang, C. Yin, Y. Li, A signal quality assessment-based ECG waveform delineation method used for wearable monitoring systems, Med. Biol. Eng. Comput. 59 (10) (2021) 2073–2084.
[65] G. Hirsch, S.H. Jensen, E.S. Poulsen, S. Puthusserypady, Atrial fibrillation detection using heart rate variability and atrial activity: A hybrid approach, Expert Syst. Appl. 169 (2021) 114452.
[66] J. Bacevicius, Z. Abramikas, E. Dvinelis, D. Audzijoniene, M. Petrylaite, et al., High specificity wearable device with photoplethysmography and six-lead electrocardiography for atrial fibrillation detection challenged by frequent premature contractions: DoubleCheck-AF, Front. Cardiovasc. Med. 9 (2022) 869730.
[67] H. Xu, W. Yan, K. Lan, C. Ma, D. Wu, A. Wu, Z. Yang, J. Wang, Y. Zang, M. Yan, Z. Zhang, Assessing electrocardiogram and respiratory signal quality of a wearable device (SensEcho): Semisupervised machine learning-based validation study, JMIR Mhealth Uhealth 9 (8) (2021) e25415.