Chaos
An Interdisciplinary Journal of Nonlinear Science

Detection of cardiac arrhythmia patterns in ECG through $H \times C$ plane

Cite as: Chaos **32**, 123118 (2022); https://doi.org/10.1063/5.0118717
Submitted: 05 August 2022 • Accepted: 10 November 2022 • Published Online: 05 December 2022

P. Martínez Coq, A. Rey, O. A. Rosso, et al.

AIP Publishing

Detection of cardiac arrhythmia patterns in ECG through $H \times C$ plane

Cite as: Chaos **32**, 123118 (2022); doi: 10.1063/5.0118717
Submitted: 5 August 2022 · Accepted: 10 November 2022 ·
Published Online: 5 December 2022

View Online Export Citation CrossMark

P. Martínez Coq,[1,a)] A. Rey,[1] O. A. Rosso,[2,3] R. Armentano,[4] and W. Legnani[1]

AFFILIATIONS

[1] Signal and Image Processing Center (CPSI), Facultad Regional Buenos Aires. Universidad Tecnológica Nacional, Ciudad Autónoma de Buenos Aires C1179AAQ, Argentina
[2] Physics Institute, Universidade Federal de Alagoas (UFAL), Maceió CEP 57072-900, Brazil
[3] Instituto de Física (IFLP), Universidad Nacional de la Plata, CONICET, CCT La Plata, La Plata B1900, Argentina
[4] Bioengineering Research and Development Group (GIBIO), Facultad Regional Buenos Aires. Universidad Tecnológica Nacional, Ciudad Autónoma de Buenos Aires C1179AAQ, Argentina

Note: This paper is part of the Focus Issue on Ordinal Methods: Concepts, Applications, New Developments and Challenges.

[a)] **Author to whom correspondence should be addressed:** pmartinez@frba.utn.edu.ar

ABSTRACT

The aim of this study is to formulate a new methodology based upon informational tools to detect patients with cardiac arrhythmias. As it is known, sudden death is the consequence of a final arrhythmia, and here lies the relevance of the efforts aimed at the early detection of arrhythmias. The information content in the time series from an electrocardiogram (ECG) signal is conveyed in the form of a probability distribution function, to compute the permutation entropy proposed by Bandt and Pompe. This selection was made seeking its remarkable conceptual simplicity, computational speed, and robustness to noise. In this work, two well-known databases were used, one containing normal sinus rhythms and another one containing arrhythmias, both from the MIT medical databank. For different values of embedding time delay τ, normalized permutation entropy and statistical complexity measure are computed to finally represent them on the horizontal and vertical axes, respectively, which define the causal plane $H \times C$. To improve the results obtained in previous works, a feature set composed by these two magnitudes is built to train the following supervised machine learning algorithms: random forest (RF), support vector machine (SVM), and k nearest neighbors (kNN). To evaluate the performance of each classification technique, a 10-fold cross-validation scheme repeated 10 times was implemented. Finally, to select the best model, three quality parameters were computed, namely, accuracy, the area under the receiver operative characteristic (ROC) curve (AUC), and the F1-score. The results obtained show that the best classification model to detect the ECG coming from arrhythmic patients is RF. The values of the quality parameters were at the same levels reported in the available literature using a larger data set, thus supporting this proposal that uses a very small-sized feature space to train the model later used to classify. Summarizing, the attained results show the possibility to discriminate both groups of patients, with normal sinus rhythm or arrhythmic ECG, showing a promising efficiency in the definition of new markers for the detection of cardiovascular pathologies.

To detect arrhythmic electrocardiogram (ECG) signals, the corresponding points in the Shannon permutation entropy and statistical complexity plane were computed and considered as a feature space to train three machine learning classification algorithms. The best results were achieved with the random forest methodology after a 10-times 10-fold cross-validation scheme was applied to compute the corresponding quality parameters.

I. INTRODUCTION

The number of deaths worldwide reported by the World Health Organization (WHO) was 55.4 million in 2019. Fifty five percent of them were related to top ten causes. Ischemic heart disease and stroke were the world's biggest killers, causing about 16% and 11% of total deaths worldwide, respectively (for more information, cf. Ref. 1). It is important to point out that an arrhythmia does not necessarily mean an irregularity of the heart rate, since the heart can frequently have regular arrhythmias with absolute stability, such as a tachycardia or a flutter, where the heart rate is within a normal range.[2] A diagnosis of arrhythmia in itself does not mean an evident pathology. In fact, in healthy subjects, the sporadic presence of certain arrhythmias can occur. For the American Heart Association,[3] some arrhythmias are so brief (a temporary pause or a premature beat) that the overall heart rate or rhythm is not affected at the clinical level. However, if arrhythmias last longer, these may cause the heart rate to be too slow, too fast, or become erratic—so the heart pumps less effectively. Cardiac arrhythmias are classified according to their place of origin, underlying mechanism, electrocardiographic pattern, or the clinical point of view.

The greatest relevance of arrhythmias is related to an association with sudden cardiac death.[4] It is also important to remember that frequent arrhythmias (especially atrial fibrillation) may lead to embolism, including cerebral embolism, often with severe consequences. Sometimes, fast arrhythmias may trigger or worsen a heart failure;[5] and the incidence of the majority of arrhythmias increases progressively with age and they are not frequent in children.[6]

It is important to point out that the ECG signals are obtained in a non-invasive way, and are widely available for use in medical research. In particular, single lead ECG signal acquisition is useful in wearable applications (see for example, Refs. 7 and 8 and within references).

An arrhythmia may or may not be suspected before the ECG is performed. Sometimes, a patient describes a change in heart rhythm suggestive of an arrhythmia, but nothing appears on the surface of the ECG tracing, even if an arrhythmia is detected during the physical examination.

The possibility of linking the diagnosis of arrhythmia with measures of information theory that provide some orientation about the dynamic process of this manifestation concerning a reference state considered as healthy is a fact that partly motivated the present study. Entropy and statistical complexity were the variables selected to test the hypothesis, constituting the $H \times C$ plane. Both are widely used in contemporary signal analysis. Assuming, as indicated in Ref. 9, that a progressively diseased cardiovascular system decomplexes, we attempt to detect which are the locations in the $H \times C$ plane for both classes of interest: normal rhythm ECGs and arrhythmias. Computer learning algorithms are then implemented to obtain more robust results. Since mobile monitoring devices (holter and recorders) require in all cases the validation of specialists for the final reports, the present research could contribute to performing the first step to distinguish between records of healthy people, and those which require more detailed analysis in the diagnostic stage once the patients hand over their acquisition devices in the health care centers. This might be done once the study has been extended to a substantial number of patients to be validated accordingly and has been sufficiently analyzed in the specialized medical community.

In the following sections, the probability distribution function (PDF) of ordinal pattern permutations proposed by Bandt and Pompe is computed using embedding time delay values ranging from 1 to 35. Then, the normalized Shannon entropy and the statistical complexity measure based on the Jensen–Shannon divergence are calculated to define the points in the $H \times C$ plane. Finally, three conceptual different and well-known supervised machine learning techniques are tuned in order to discriminate arrhythmic patients from those with a normal sinus rhythm.

II. MATERIALS AND METHODS

The concept of entropy was first introduced in the theory of communications by Shannon[10] as a tool to measure the degree of organization in a system with physical properties. Nowadays, it has become one of the most emblematic notions to quantify information of a dynamical system. This definition was extended to dynamic systems by Kolmogorov[11] and amended by Sinai[12] due to its application of symbolic encoding of phase space. In Ref. 5, the authors provided a number of entropy interpretations emerged from a wide range of science and technology topics, such as, disorder, state space volume, and lack of information.

Let $X(t) = \{x_t : t = 1, \ldots, M\}$ be a time series of M observations of the variable X. Its associated PDF is given by $P = \{p_i : i = 1, \ldots, N\}$, where N is the number of possible states of the system under study and $\sum_{i=1}^{N} p_i = 1$. Then, Shannon entropy is defined as

$$S[P] = -\sum_{i=1}^{N} p_i \ln(p_i). \tag{1}$$

Expression (1) can be seen as a measure of the uncertainty related to the physical process described by P. When $S[P] = 0$, it means that the underlying structure is fully deterministic, so that the knowledge of the process is maximal at this instance. On the opposite corner, for maximal uncertainty, such as an uncorrelated stochastic process with uniform distribution, the knowledge of the dynamic system is minimal, which implies that all the states have the same probability of occurrence. This PDF is denoted by $P_e = \{p_i = 1/N : i = 1, \ldots, N\}$ (cf. Refs. 13–15). It is worth noting that no distribution is required to be known since the calculus is based on the state probabilities.

The well-known normalized Shannon entropy is defined by

$$H[P] = \frac{S[P]}{S[P_e]} = \frac{S[P]}{\ln N} \tag{2}$$

and satisfies $0 \leq H[P] \leq 1$.

In order to introduce a definition of complexity in physics, it is crucial to consider it as an indicator of plausible undetected patterns that depict the system as dynamic.[16] In Ref. 17, it is suggested that a kind of distance to a reference PDF must be included in the complexity computation. Thus, the disequilibrium can be defined by

$$Q[P] = Q_0 D[P, P_e], \tag{3}$$

where Q_0 is a normalization constant and D is a stochastic distance. In this work, D is considered as the Jensen–Shannon divergence given by

$$D_{JS}[P, P_e] = S\left[\frac{P + P_e}{2}\right] - \frac{1}{2}S[P] - \frac{1}{2}\ln(N), \tag{4}$$

for which

$$Q_0 = -2\left[\frac{N+1}{N}\ln(N+1) - 2\ln(2N) + \ln(N)\right]^{-1}. \tag{5}$$

It can be observed that the disequilibrium vanishes when the distribution of the time series resembles a cloud of uniformly sparse points, which is maximized when the time series is periodic.

Thus, the complexity can be measured as a combination between the information inherent to the system and its disequilibrium. Explicitly, using Eqs. (3) and (2), the complexity can be computed as follows:

$$C[P] = Q[P]H[P]. \tag{6}$$

The qualitative information extraction without parametric model assumptions and temporal ordering structure quantifies the degree of complexity in terms of expression (6). Hence, this statistical complexity allows us to identify different levels of periodicity and chaos.[18]

Since both information measures mentioned above are calculated in terms of a PDF, the approach proposed by Bandt and Pompe[19] is used. This methodology consists in obtaining the probabilities related to the ordinal dynamism of the elements in a time series. Consider the time series $X(t)$. If $m > 1$ represents the embedding dimension and τ the embedding time delay given by the length of the interval between two consecutive observations in the resampling, then $M - m + 1$ overlapping partitions of length m are constructed as follows:

$$s \rightarrow \{x_{s-\tau(m-1)}, x_{s-\tau(m-2)}, \dots, x_{s-\tau}, x_s\}, \tag{7}$$

with $s = m, m+1, \dots, N$. For each s, the permutations of the set $\{0, 1, \dots, m-1\}$ are denoted by $\pi_j = \{r_0, r_1, \dots, r_{m-1}\}$ and given by the ordering $x_{s-\tau r_{m-1}} \leq x_{s-\tau r_{m-2}} \leq \dots \leq x_{s-\tau r_0}$. For $j = 1, \dots, m!$, the permutation π_j has the following probability of occurrence:

$$p_j(\pi_j) = \frac{\#\{s \text{ is of type } \pi_j\}}{M - m + 1}. \tag{8}$$

The condition $M \gg m!$ is necessary to ensure statistical reliability and a proper dissimilitude for deterministic and stochastic systems.[20]

If PDF given in Eq. (8) is replaced in Eq. (2), an extension of Shannon entropy is obtained, which is called permutation entropy (PE). One of the main advantages of this technique lies in that there is no need of a statistical model for the signal, so that there are no assumptions about the nature of the underlying process. Besides, most of these models are prone to outliers.

A. Data set

The collection of ECG time series studied in this work was obtained from the PhysioNet platform managed by members of the Computational Physiology Laboratory of the Massachusetts Institute of Technology, and it is available at https://physionet.org/. This set contains 18 normal sinus rhythms ECGs from 5 men and 13 women who were found to have had no significant arrhythmias[21] and 48 long-term recordings of ECG belonging to patients with cardiac arrhythmias from 26 men and 22 women, 37 of whom were taking medication.[22] Since the original records of arrhythmic ECGs, which last half an hour, consist of about 650 K samples, and the normal ECG records correspond to 24-h holter's information, only the first 650 K samples were considered for the normal sinus rhythm records, which means a duration of 85 min per record approximately. These characteristics are summarized in Table I. An example of a normal sinus rhythm and a cardiac arrhythmic ECG recording are exhibited in Fig. 1.

TABLE I. General characteristics of the ECG database.

	Normal sinus rhythm	Cardiac arrhythmias
Recordings	18	48
Males	5	26
Females	13	22
Sampling acquisition frequency	128 Hz	360 Hz
Record's length (time in min/samples)	85/650 K	30/650 K

In many research studies, the authors apply a noise filter. However, the interest in the present study is to analyze the information and structure of the ECG signal in the raw form. This decision is inspired by the work.[23]

B. Definition of the feature space

As mentioned above, for the computation of PE two parameters are required. Due to the length of the signals of interest, an embedding dimension $m = 6$ is chosen. This value option is motivated by a reasonable statistical estimation of the ordinal patterns related to the number of data in the signal, as well as the recommendation in the specific literature (e.g., Refs. 24 and 25). In practice, for the selection of τ, it is usual to adopt a value suggested by experimental tests developed by scientists and researchers from the available literature.

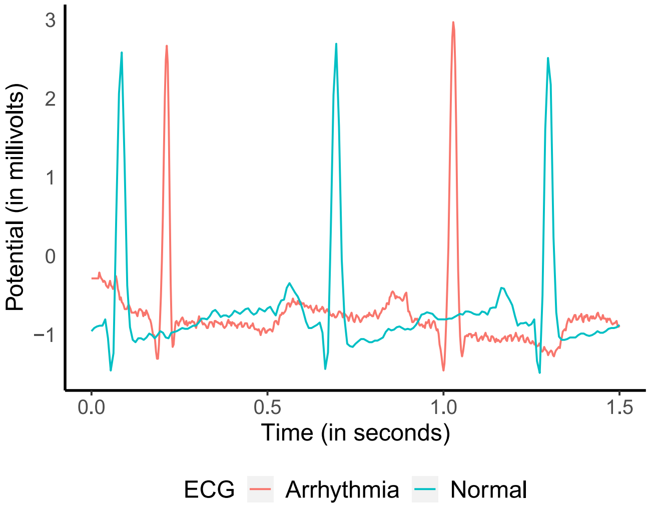

FIG. 1. Examples of normal and arrhythmic ECG signals, respectively, named "16273" from the MIT-BIH Normal sinus rhythm database and "100" from the MIT-BIH Arrhythmia database.

However, this habit has drawbacks as the strong dependence of PE on the sampling frequency[26] or the need of an expert advice.

In Ref. 27, the authors based the computations with $\tau = 15$. Due to the absence of a unique criterion for τ selection and with the aim to improve the results obtained in this work, automatic methods to select a proper embedding time delay can be found in references such as Refs. 24, 25, and 28. The implementation of these algorithms yielded similar values for τ in terms of the type of ECG signal.

The feature space from which the classifier is going to learn is defined as follows. For each ECG signal in the database introduced in Sec. II A and for a fixed τ value varying in the range of $1, 2, \ldots, 35$, the PE and the complexity are computed. Finally, every signal is labeled as "normal" or "arrhythmia" referring to the original database. This means that the group of patients can be divided into two classes.

From now on, the PE will be simply denoted by H and the complexity using Jenssen–Shannon divergence by C. Therefore, a point in the $H \times C$ plane is associated to every ECG signal. It is known that for a given value of H, the possible values for C are bounded between two curves denoted by $C_{\min}$ and $C_{\max}$.[29] This particularity is illustrated in Fig. 3 for all the signals under analysis and for some values of τ.

C. Classification models

The approach to improve ECG classification using machine learning techniques has been used for several years now. The main research effort made was focused on the feature space constructed from time variables, a kind of signal transform or a combination of both. In that sense, a valuable source of information is compiled in Ref. 30. Considering the wide set of algorithms applied, there are cases in which the researchers use methods based on neural networks, Markov chain models, support vector machine, and multilayer sensor classifiers, as can be seen in Refs. 31–34.

To detect arrhythmias in the ECG database, three well-known supervised classification techniques are applied: random forest (RF), support vector machine (SVM,) and k nearest neighbors (kNN). This selection is inspired in the dissimilar intrinsic properties of the three approaches, based on decision trees, hyperplanes, and distances. Thus, these differences contribute to a strong differentiation and clear comparison. It is worth mentioning that in spite of the small sample size, the results obtained by machine learning techniques can be reliable as argued in Ref. 35.

RF[36] is a method that uses sets of decision trees on either splits with randomly generated vectors or random subsets of training data, and computes the score as a function of these different components. In the classification context, the prediction made by random forest results from the most frequent class in the set of predictions obtained per every decision tree. Bagging with decision trees can be considered a special case of random forest depending on how the sample is selected (bootstrapping). Being an ensemble model, variance is reduced compared to training a single tree. Pruning is not necessary to avoid potential over-fitting as in the use of a single tree, since the samples used to train the individual trees in the forest are bootstrapped.

The SVM,[37] initially thought as a binary classification algorithm in an n-dimensional space, attempts to find a hyperplane, i.e.,

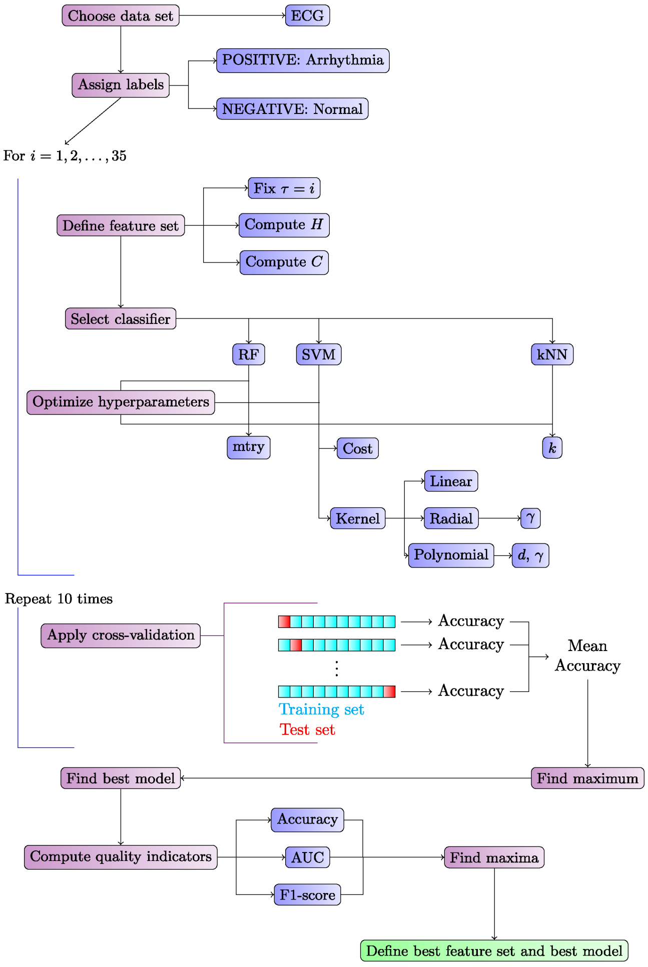

FIG. 2. Workflow.

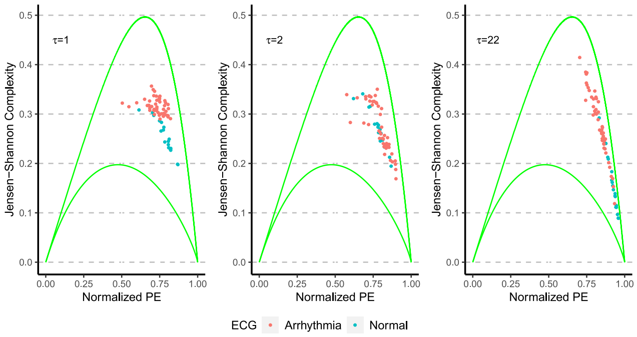

FIG. 3. Scatterplot in the plane $H \times C$ with the boundary curves $C_{\min}$ and $C_{\max}$ for different values of τ.

an $n-1$-dimensional subspace, which can be used to separate the regions occupied by elements in each class. The hyperplane is a curve that splits the feature space in such a way that the margin—distance between the data points in the two different groups—is maximized. The name of this technique is due to data points closest to the hyperplane are called support vectors. SVM performs very well with higher-dimensional data sets and is one of the most memory-efficient classifiers. Despite this algorithm is originally in the linear context, the kernel trick allows us to extend its application to other frameworks. Although this methodology is very stable, it is not recommended if the sample size is large and it may not be suitable in the presence of overlapped or noisy data.

The kNN[38] classifies an object by a majority vote of the object's neighbors, in the space of input parameters, i.e., the new object is assigned to the most common class among its k nearest neighbors. This non-parametric method is simple to implement, robust to noisy training data, and effective if training large data sets. On the contrary, it is necessary to determine the value of k and the computation cost is high as it needs to compute the distance from each instance to all the training samples. This kind of classifier is memory-based and requires no model to be fit. Despite its simplicity, the kNN algorithm has been successful in a large number of classification problems, even if each class has many possible prototypes or the decision boundary is very irregular.

For a deeper treatment and more details of these classifiers, see Refs. 39–41. In this work, packages `randomForest`,[42] `e1071`[43] and `kknn`,[44] from R language, were used to implement these classifiers.

A 10-fold cross-validation repeated ten times was applied to each model for tuning the following hyperparameters:

- the number of variables randomly sampled as candidates at each split in RF, denoted by "mtry," where the number of trees is fixed as 500;
- the type of kernel—linear, polynomial or radial—and cost c of constraints violation in SVM, as well as the degree d in polynomial kernel and the constant factor γ in both, polynomial and radial kernels; and
- the maximum number of neighbors k in kNN.

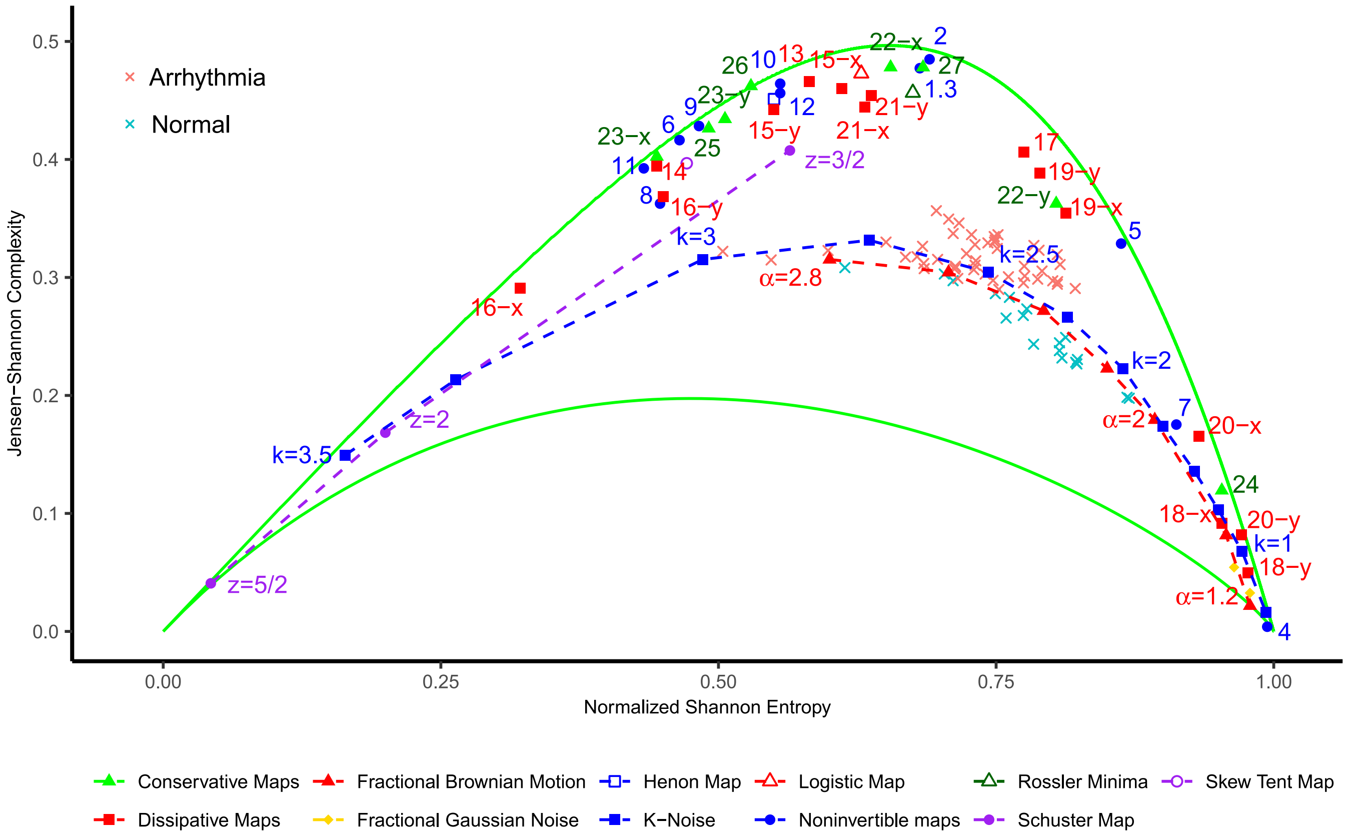

FIG. 4. $H \times C$ plane: maximum and minimum complexity curves (continuous green lines), normal sinus rhythm ECGs and ECGs from patients with arrhythmia, and dynamical systems to locate a diversity of behaviors in this plane following Ref. 9.

D. Classifier quality indicators

A positive label is assigned to arrhythmic ECG records. The possible predictions can be: true positive (TP), true negative (TN), false positive (FP), or false negative (FN). The classification quality indicators to be used are accuracy, the area under the receiver operating characteristics (ROC) curve (AUC), and the F1-score. Accuracy is defined as the proportion of well predicted data; i.e., $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})$. The area under the ROC curve (AUC) can be regarded as the probability that the classifier ranks a random positive observation more highly than a random negative observation (cf. Ref. 45). AUC ranges from 0 (all predictions are wrong) to 1 (all predictions are correct). The F1-score is the harmonic mean between the precision given by $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, and the recall defined by $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. Thus, $F_1 = 2(\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall})$.

As a summary of Sec. II, the workflow developed in this research is presented in Fig. 2. Besides, for the reproducibility of the results of the present work, the implemented codes are freely available at https://github.com/arey1911/ECG-classification-by-HxC-plane.

III. RESULTS AND DISCUSSION

Beginning with the formulation of the feature set calculated from the ECG time series of patients with cardiac normal sinus rhythms and arrhythmic ones from the Physionet database, a cloud of points is built as the feature space providing information for the classification process. As it is shown in Fig. 3, ECGs with a normal sinus rhythm show, in their waveform, a higher PE, and a lower statistical complexity measure for $\tau = 1$. In addition, there is a confirmation that there is a displacement of the cloud of points from the patients with cardiac arrhythmia toward the space of a larger statistical complexity measure and a smaller PE. That feature was also found in Ref. 27. All the points are placed in the plane $H \times C$ within the area limited by the two boundary curves $C_{\min}$ and $C_{\max}$. Those coming from the arrhythmic ECG group remain in a zone occupied above the k-noises boundary, approximately $k = 2.5$, as shown in Ref. 9. In Fig. 4, the clouds of points corresponding to both classes are located with a low level of overlap. This fact encourages the use of a computer learning algorithm to increase the efficiency of a classification model distinguishing both groups of ECG signals. Most ECG recordings from patients with diagnosed arrhythmia have coordinates in the $H \times C$ plane that place them above the dashed blue curve on which the k-noises fall, whereas ECGs from normal sinus rhythm recordings are located below the aforementioned line. These relative locations in the $H \times C$ plane would indicate that the arrhythmia dynamic behavior of the cardiac system would cause a decrease in the entropy value with a consequent increase in the corresponding statistical complexity value. This fact suggests that the disease renders the cardiac regulatory system less capable of regulation, an effect that coincides with the behavior of cardiovascular disease states that have been reported in the literature. As was pointed out in Ref. 46, the loss of complexity in physiological signals is related to a disease process or aging. In this particular case, the lower signal entropy could contribute to explain the underlying mechanism that drives a healthy cardiovascular system to develop an arrhythmia. In turn, the increment of the statistical complexity in the case of arrhythmic records was verified in Ref. 47, locating atrial fibrillation ECGs records in an $H \times C$ region over the fractional Brownian motion locations, thus providing partial confirmation of the results of the present research. Continuing with this reasoning, the normal sinus rhythm ECG group is located close to the region characterized by the random processes, whose main characteristic is the high level or variation. In contrast, the arrhythmia disease shows a lower variability as indicated by the lower entropy level, the points approaching the regions that begin to be occupied by the chaotic systems. The chaotic behavior is supposed to try to set a degree of order. In other words, if the system shows a chaotic behavior, then there is some order. Meanwhile, noise does not introduce order but randomly distributes frequency spectra.

TABLE II. Best models for each value of τ after a 10-fold cross-validation repeated 10 times. The best, average, and worst values are, respectively, highlighted in green, blue, and red.

τ	Accuracy	Best model
1	**0.944**	RF, mtry = 2
2	**0.737**	SVM radial, $c = 0.5$, $\gamma = 6.992$
3	0.744	SVM radial, $c = 1$, $\gamma = 4.508$
4	0.815	SVM linear, $c = 1$
5	0.798	SVM linear, $c = 1$
6	0.905	kNN, $k = 5$
7	0.822	RF, mtry = 2
8	0.869	SVM linear, $c = 1$
9	0.912	SVM linear, $c = 1$
10	0.925	SVM linear, $c = 1$
11	0.933	SVM linear, $c = 1$
12	0.927	SVM radial, $c = 1$, $\gamma = 4.181$
13	0.924	SVM polynomial, $c = 1$, $d = 3$, $\gamma = 0.1$
14	0.895	SVM polynomial, $c = 1$, $d = 2$, $\gamma = 0.1$
15	0.892	SVM polynomial, $c = 1$, $d = 3$, $\gamma = 0.1$
16	0.884	SVM polynomial, $c = 1$, $d = 3$, $\gamma = 0.1$
17	0.934	RF, mtry = 2
18	0.942	SVM linear, $c = 0.5$
19	0.916	SVM linear, $c = 1$
20	0.908	SVM polynomial, $c = 0.5$, $d = 1$, $\gamma = 0.1$
21	0.879	SVM linear, $c = 1$
22	**0.881**	SVM radial, $c = 1$, $\gamma = 0.106$
23	0.867	SVM polynomial, $c = 0.25$, $d = 3$, $\gamma = 0.1$
24	0.858	SVM radial, $c = 0.5$, $\gamma = 16.961$
25	0.877	RF, mtry = 2
26	0.871	SVM radial, $c = 1$, $\gamma = 0.115$
27	0.880	SVM polynomial, $c = 1$, $d = 1$, $\gamma = 0.1$
28	0.886	SVM linear, $c = 0.25$
29	0.887	SVM linear, $c = 1$
30	0.889	SVM linear, $c = 0.25$
31	0.880	SVM radial, $c = 1$, $\gamma = 0.09$
32	0.911	SVM radial, $c = 1$, $\gamma = 17.95$
33	0.882	RF, mtry = 2
34	0.880	SVM linear, $c = 1$
35	0.885	SVM linear, $c = 0.25$

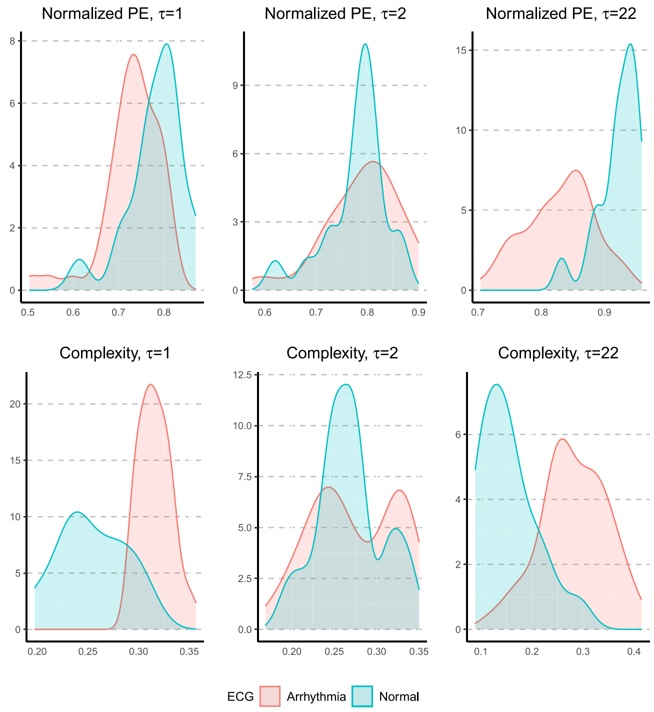

FIG. 5. Histograms of each feature according to the type of ECG for different values of τ.

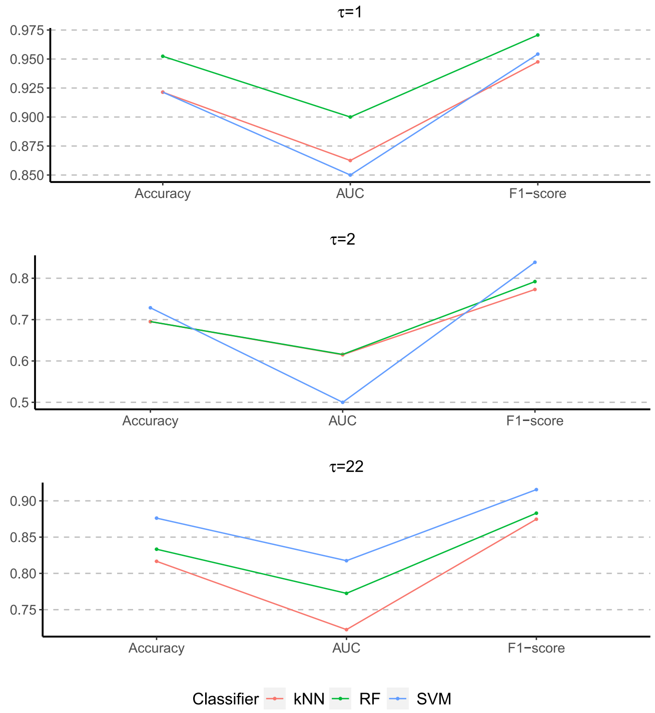

FIG. 6. Values of the quality indicators of the models using different classifiers after implementing a 10-time 10-fold cross-validation for different values of τ.

Notice that the behavior described above does not hold for $\tau = 2$. In the case $\tau = 22$, the points in the plane belonging to each group of ECG signals are not so clearly separated as in the case of $\tau = 1$. The point locations in the $H \times C$ plane might justify the results shown in Table II.

Although the peaks in the histograms of the PE entropy are separated as it is shown in Fig. 5 for $\tau = 1$ and $\tau = 22$, the strong overlapping between the two distributions is notorious, especially in the first case. On the other hand, the histograms for the statistical complexity measure computed with these two values of τ, which are presented in the same figure, show a clear separation between their peaks, which is more evident for $\tau = 1$. However, the normal sinus rhythm ECG group also has a second modal peak whose position is aligned with the one reached by the peak of the other group. This feature may explain a certain degree of deficiency in the classification quality of the normal ECG group. As a consequence of these graphics, the statistical complexity measure appears to be the most meaningful data set variable to train the classifier. Again, these properties are not verified for $\tau = 2$, where the entropy and complexity histograms suggest a misunderstanding between the two classes.

The best models after the hyperparameter tuning using a 10-fold cross-validation repeated 10 times are exhibited in Table II, as well as the accuracy obtained in each case. It can be seen that the maximum, minimum and mean values for accuracy are 0.944 for $\tau = 1$; 0.737 for $\tau = 2$; and 0.881 for $\tau = 22$, respectively. In this sense, these three values of τ are considered as references to exemplify several characteristics throughout the present study.

With the aim to illustrate the behavior of the three classification techniques proposed, the values of accuracy, AUC and F1-score, after applying 10-fold cross-validation for $\tau = 1, 2, 22$, are shown in Fig. 6. Besides, the best results for every value of τ are summarized in Fig. 7. It can be noticed that the best performance was achieved by the RF model using $\tau = 1$, for which the average accuracy is 0.952, the average AUC is 0.900, and the average F1-score is 0.971. With the aim to compare existing results in the available literature, but without the depth of a review, it is worth noticing that in Ref. 23 using SVM, kNN, RF, and other two different options, with the same database, the overall accuracy values range from 0.725 to 0.944 and the F1-scores run from 0.391 to 0.662. It is important to keep in mind that those results were obtained using a larger feature space than the one of our proposal.

As indicated in Ref. 48, PE could be susceptible to the presence of noise. Hence, the authors suggest being careful in the application of these tools to time series for which the signal-to-noise ratio is low. Since the complexity decreases for noisy signals, our method is extremely sensitive to noise perturbation. In this sense, it is noteworthy that the conclusions obtained in this study are biased by the presence of noise superposed to ECG traces.

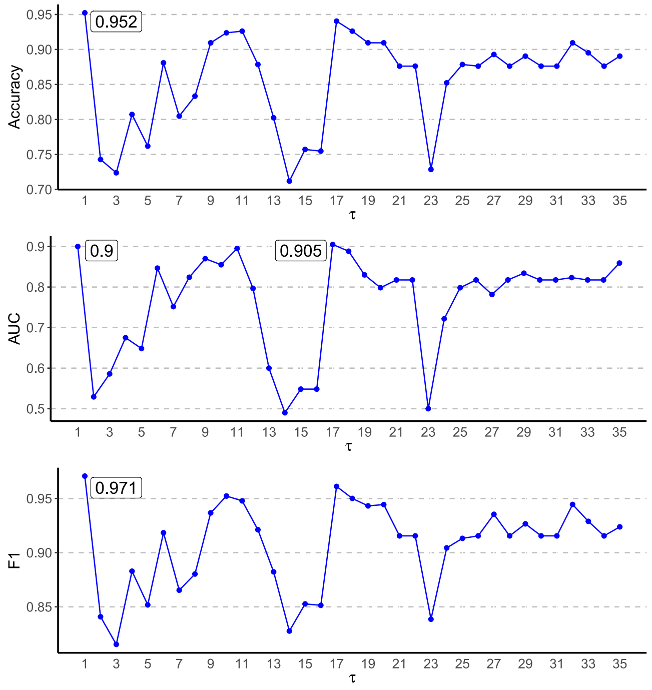

FIG. 7. Values of the quality indicators of the best model for each τ using a 10-times 10-fold cross-validation.

IV. CONCLUSION

This research has two cornerstones. The first one consists of the fact that it is possible to use a reduced set of features formed by informational measures to classify signals of ECGs coming from normal sinus rhythms and arrhythmias. The second one is that the classification model presented in this work has an explaining character of the process which leads from health to disease, under the light of dynamical systems. With reference to the first cornerstone, this work presents the possibility to use only the $H \times C$ plane as a feature space to train classifiers with the aim to construct models. Not only is the reduced dimension of the feature space a clear advantage of this proposal, but also the interpretation of the features selected gives an important insight into the dynamical characteristics of the disease process. In reference to the second cornerstone, the lower entropy values and higher complexity values contribute to understanding the mechanism by which arrhythmias occur. This would indicate a dismissing capacity of the cardiovascular system to control the dynamics of the heart, given by the variation in the values of the Shannon permutation entropy between normal sinus rhythms and arrhythmic records. On the other hand, the increment of the statistical complexity would show the emergence of more structures in the shape of ECG signals from patients with arrhythmia. Since the selection of the embedding time delay and of the embedding dimension are two relevant aspects to consider for the calculations of the permutation distributions leading to the permutation entropy, in the present work this choice was carried out in a way based on the bibliographic background and its subsequent application to the data used in the study. The value of the embedding dimension was chosen based on the selection of the signal length to obtain the improved experimental statistics from the data series, while the embedding time delay was varied on a range based upon the criteria existing in the literature and the optimum value was selected based on the best performance obtained by the classifier.

It should be noted that the points in the $H \times C$ plane that correspond to fractional Brownian motion and k-noises constitute an approximated border which makes it possible to differentiate the cloud of points of both ECG groups of interest.

At the stage of formulating the models, the selected classification algorithms cover a wide spectrum of the most frequently used ones in contemporary research. This is because one of them is based on decision trees, such as Random Forest, another one uses distance calculation, such as kNN, and the remaining one looks for the best hypersurface to be used to separate the analysed classes. The promising results obtained by the computational performance during the calculation of Shannon Permutation Entropy and Statistical Complexity, added to the use of a single channel of ECG recordings, motivate us to consider as feature work, the formulation of a pilot protocol for home follow-up of patients clinically suspected of suffering from arrhythmia. The used databases have been employed in a wide variety of scientific studies, so their validity has been demonstrated. However, given their small size, it is essential to increase the number of cases from both databases to improve the statistical aspects of the proposal and to test it on larger sets, if possible, coming from different sources.

ACKNOWLEDGMENTS

This work was developed within the framework of Project Nos. PID UTN 4729 and PID UTN 8120. R. Armentano gives thanks to project No. UNDP URU-84-002.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

P. Martínez Coq: Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Validation (equal) (equal); Writing – original draft (equal); Writing – review & editing (equal). **A. Rey:** Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **O. A. Rosso:** Methodology (equal); Supervision (equal); Validation (equal); Writing – review & editing (equal). **R. Armentano:** Supervision (equal); Writing – review & editing (equal). **W. Legnani:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1] See https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death for "WHO: The top 10 causes of death" (last accessed August 15, 2021), 2019.

[2] A. B. De Luna and A. Baranchuk, *Clinical Arrhythmology* (John Wiley & Sons, 2017).

[3] American Heart Association, see https://www.heart.org/en/health-topics/arrhythmia/about-arrhythmia for "About Arrhythmia" (accessed August 15, 2021), 2016.

[4] S. Goldstein, J. G. Soldevila, and A. B. de Luna, *Sudden Cardiac Death* (Futura Publishing Company, 1994).

[5] F. Olivares, L. Souza, W. Legnani, and O. A. Rosso, "Informational time causal planes: A tool for chaotic map dynamic visualization," in *Nonlinear Systems-Theoretical Aspects and Recent Applications* (IntechOpen, 2019).

[6] B. Gray, M. J. Ackerman, C. Semsarian, and E. R. Behr, "Evaluation after sudden death in the young: A global approach," Circ.: Arrhythmia Electrophysiol. **12**, e007453 (2019).

[7] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (IEEE, 2017), pp. 1–4.

[8] Y. Xia, H. Zhang, L. Xu, Z. Gao, H. Zhang, H. Liu, and S. Li, "An automatic cardiac arrhythmia classification system with wearable electrocardiogram," IEEE Access **6**, 16529–16538 (2018).

[9] O. A. Rosso, F. Olivares, L. Zunino, L. De Micco, A. L. Aquino, A. Plastino, and H. A. Larrondo, "Characterization of chaotic maps using the permutation Bandt–Pompe probability distribution," Eur. Phys. J. B **86**, 1–13 (2013).

[10] C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423 (1948).

[11] A. N. Kolmogorov, "New metric invariant of transitive dynamical systems and endomorphisms of lebesgue spaces," Dokl. Russ. Acad. Sci. **119**, 861–864 (1958).

[12] I. Sinai, "On the concept of entropy for a dynamic system," Dokl. Akad. Nauk SSSR **124**, 768–771 (1959).

[13] J.-B. Brissaud, "The meanings of entropy," Entropy **7**, 68–96 (2005).

[14] L. Zunino, A. F. Bariviera, M. B. Guercio, L. B. Martinez, and O. A. Rosso, "On the efficiency of sovereign bond markets," Physica A **391**, 4342–4349 (2012).

[15] A. F. Bariviera, L. Zunino, M. B. Guercio, L. B. Martinez, and O. A. Rosso, "Efficiency and credit ratings: A permutation-information-theory analysis," J. Stat. Mech.: Theory Exp. **2013**, P08007 (2013).

[16] O. A. Rosso, L. De Micco, H. A. Larrondo, M. T. Martín, and A. Plastino, "Generalized statistical complexity measure," Int. J. Bifurcation Chaos **20**, 775–785 (2010).

[17] R. Lopez-Ruiz, H. L. Mancini, and X. Calbet, "A statistical measure of complexity," Phys. Lett. A **209**, 321–326 (1995).

[18] L. Zunino, M. C. Soriano, and O. A. Rosso, "Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach," Phys. Rev. E **86**, 046210 (2012).

[19] C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series," Phys. Rev. Lett. **88**, 174102 (2002).

[20] A. Kowalski, M. Martín, A. Plastino, and O. A. Rosso, "Bandt–Pompe approach to the classical-quantum transition," Physica D **233**, 21–31 (2007).

[21] 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, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," Circulation **101**, e215–e220 (2000).

[22] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," IEEE Eng. Med. Biol. Mag. **20**, 45–50 (2001).

[23] S. K. Pandey, R. R. Janghel, and V. Vani, "Patient specific machine learning models for ECG signal classification," Procedia Comput. Sci. **167**, 2181–2190 (2020).

[24] M. Riedl, A. Müller, and N. Wessel, "Practical considerations of permutation entropy," Eur. Phys. J. Spec. Top. **222**, 249–262 (2013).

[25] A. Myers and F. A. Khasawneh, "On the automatic parameter selection for permutation entropy," Chaos **30**, 033130 (2020).

[26] A. Popov, O. Avilov, and O. Kanaykin, "Permutation entropy of EEG signals for different sampling rate and time lag combinations," in *2013 Signal Processing Symposium (SPS)* (IEEE, 2013), pp. 1–4.

[27] P. Martinez Coq, W. Legnani, and R. Armentano, "Detection of arrhythmic cardiac signals from ECG recordings using the entropy-complexity plane," in *Multidisciplinary Digital Publishing Institute Proceedings* (MDPI, 2019), Vol. 46, p. 8.

[28] A. D. Myers and F. A. Khasawneh, "Delay parameter selection in permutation entropy using topological data analysis," arXiv:1905.04329 (2019).

[29] M. Martin, A. Plastino, and O. A. Rosso, "Generalized statistical complexity measures: Geometrical and analytical properties," Physica A **369**, 439–462 (2006).

[30] J. R. Annam, S. Kalyanapu, S. Ch, J. Somala, and S. B. Raju, "Classification of ECG heartbeat arrhythmia: A review," Procedia Comput. Sci. **171**, 679–688 (2020).

[31] P. Shimpi, S. Shah, M. Shroff, and A. Godbole, "A machine learning approach for the classification of cardiac arrhythmia," in *2017 International Conference on Computing Methodologies and Communication (ICCMC)* (IEEE, 2017), pp. 603–607.

[32] H. I. Bulbul, N. Usta, and M. Yildiz, "Classification of ECG arrhythmia with machine learning techniques," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (IEEE, 2017), pp. 546–549.

[33] G. Nalbantov, S. Ivanov, and J. Van Prehn, "Multi-class classification of pathologies found on short ECG signals," in *2020 Computing in Cardiology* (IEEE, 2020), pp. 1–4.

[34] M. A. Reyna, E. A. P. Alday, A. Gu, C. Liu, S. Seyedi, A. B. Rad, A. Elola, Q. Li, A. Sharma, and G. D. Clifford, "Classification of 12-lead ECGs: The PhysioNet/computing in cardiology challenge 2020," in *2020 Computing in Cardiology* (IEEE, 2020), pp. 1–4.

[35] P. Kokol, M. Kokol, and S. Zagoranski, "Machine learning on small size samples: A synthetic knowledge synthesis," Sci. Prog. **105**, 003685042110297 (2022).

[36] L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).

[37] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (ACM Digital Library, 1992), pp. 144–152.

[38] E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination: Consistency properties," Int. Stat. Rev. **57**, 238–247 (1989).

[39] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The Elements of Statistical Learning*, Springer Series in Statistics Vol. 1 (Springer, New York, 2001).

[40] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning* (Springer, 2013), Vol. 112, p. 18.

[41] C. C. Aggarwal, "Data classification," in *Data Mining* (Springer, 2015), pp. 285–344.

[42] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News **2**, 18–22 (2002), http://CRAN.R-project.org/doc/Rnews/.

[43] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071)," R package version 1.7-7 (TU Wien, 2021).

[44] K. Schliep and K. Hechenbichler, "kknn: Weighted k-nearest neighbors," R package version 1.3.1 (2016).

[45] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett. **27**, 861–874 (2006).

[46] A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, "What is physiologic complexity and how does it change with aging and disease?," Neurobiol. Aging **23**, 23–26 (2002).

[47] K. N. Aronis, R. D. Berger, H. Calkins, J. Chrispin, J. E. Marine, D. D. Spragg, S. Tao, H. Tandri, and H. Ashikaga, "Is human atrial fibrillation stochastic or deterministic? Insights from missing ordinal patterns and causal entropy-complexity plane analysis," Chaos **28**, 063130 (2018).

[48] A. Porta, V. Bari, A. Marchi, B. De Maria, P. Castiglioni, M. Di Rienzo, S. Guzzetti, A. Cividjian, and L. Quintin, "Limits of permutation-based entropies in assessing complexity of short heart period variability," Physiol. Meas. **36**, 755–765 (2015).