Comparison of the performance of features based on symbolic dynamics and recurrence microstates for time series classification

H. C. Costa [1,2] J. V. M. Silveira [1,2] B. R. R. Boaretto [3,4] C. Masoller [3] S. R. Lopes [1,2] and T. L. Prado [1,2,*]

[1]*Department of Physics, Federal University of Paraná, Curitiba 81531-980, Brazil*

[2]*Interdisciplinary Center for Science, Technology and Innovation CICTI, Federal University of Paraná, Curitiba, Brazil*

[3]*Physics Department, Universitat Politècnica de Catalunya, Barcelona, Spain*

[4]*Science and Technology Institute—ICT, Federal University of São Paulo, São José dos Campos, Brazil*

(Received 2 September 2025; accepted 17 March 2026; published 14 May 2026)

Symbolization is a well-established tool to transform a time series into a sequence of symbols. Here, we compare two symbolization approaches: recurrence microstate (RM), based on recurrence patterns in phase space, and symbolic dynamics (SD), derived from ordinal binary sequences. Specifically, we compare the performance of two sets of features, based on RM and SD, for parameter estimation and classification. For parameter estimation, we generate synthetic data simulated from chaotic dynamical systems with a parameter that varies in a range that is divided into ten equal bins, and for classification, we use freely available electrocardiogram (ECG) recordings from patients with four different cardiac pathologies. We train a Random Forest (RF) classifier to infer to which parameter bin a synthetic time series belongs and to which pathology an ECG recording corresponds. We find that using RM features to train the RF algorithm consistently provides higher accuracy; however, SD features encapsulate complementary information that can improve the performance. Combining the two sets of features generally enhances the performance.

DOI: 10.1103/1m3f-jshf

I. INTRODUCTION

Chaotic trajectories are deterministic yet highly sensitive to initial conditions, exhibiting irregular, nonperiodic long-term behavior while remaining confined within a bounded region of an M-dimensional phase space [1]. In real-world scenarios, however, we typically have access to only a single observable from such systems, namely, a finite time series that represents a low-dimensional projection of the full M-dimensional dynamics. As a result, distinguishing between short time series generated by chaotic systems can be particularly challenging, especially when they arise from nearly identical parameters.

To analyze and classify dynamical systems from their time series data, a variety of methodologies have been proposed. Classical approaches typically focus on extracting statistical and dynamical properties such as mean, standard deviation, amplitude, frequency content, or probability distribution functions [2,3]. Many other methods have been proposed to reduce the dimensionality of the data by using symbolic representations, transforming complex signals into simplified patterns that retain essential dynamical features [4,5]. The Symbolic Dynamics (SD) constitutes a powerful framework for analyzing complex systems where properties of the real data are, typically, achieved by partitioning the phase space into a finite number of regions and assigning a distinct symbol to each region. The resulting symbolic sequences provide a simplified yet rigorous representation of the underlying dynamics, enabling the study of invariant properties, entropy measures, and long-term behavior with reduced complexity.

Among symbolic time series techniques, ordinal binary pattern analysis configures a simplified variant of ordinal patterns [4] and has gained popularity due to its simplicity and ability to capture local directional trends [6]. These binary patterns encode the relative ordering of values over time, offering a compact representation of the system's dynamics. In parallel, recurrence-based methods provide a different perspective for the same question by examining how often the system revisits similar states in its phase space [5,7]. Rooted in Poincaré's recurrence theorem [8], recurrences quantify the structure of repeated states and transitions. Here, we focus on a recent development in this area known as Recurrence Microstates (RMs) [9], which groups together temporally localized recurrence patterns into discrete states.

Furthermore, it is important to note that machine learning techniques for time series analysis are gaining enormous popularity, particularly when applied to representative symbolic representations derived from the underlying dynamics, rather than relying directly on raw data [10–12].

Recent studies have shown the effectiveness of symbolic and recurrence-based approaches for classifying dynamical systems from time series data [4,11–16]. These methods are computationally efficient and robust to noise, making them well suited for real-world applications. Recurrence plots (RPs), initially developed as a qualitative visualization tool

*Contact author: thiago@fisica.ufpr.br

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

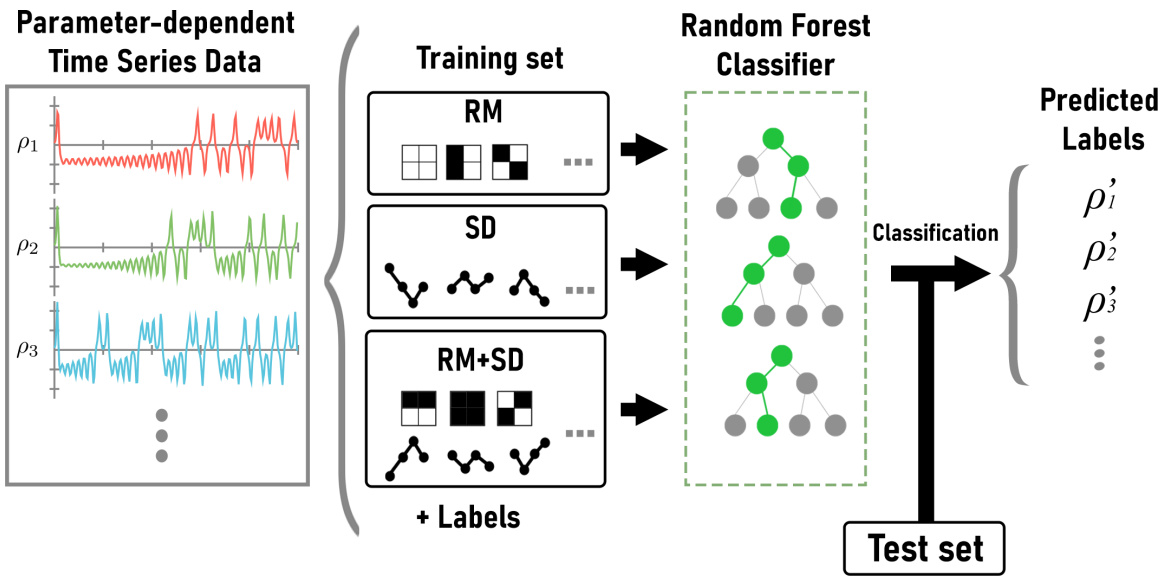


FIG. 1. Schematic illustration of the classification procedure applied to parameter-dependent time series generated from dynamical systems. On the left, time series of a representative dynamical system are shown (Lorenz system in this example) for different values of a control parameter [Lorenz ρ parameter (see Eq. (12))], which defines the labels to be predicted. These labeled time series are then processed through three strategies for feature extraction: (1) RM, where the features are the probabilities of the 16 possible microstates shown in Fig. 2; (2) SD, where the features are the probabilities of the 16 possible ordinal binary patterns shown in Fig. 3; and (3) the combination of RM and SD strategies, where the two sets of probabilities are concatenated in a joint feature vector. The resulting features, along with the corresponding labels, compose the training set used to train a Random Forest classifier. Finally, a test set composed of unseen data is provided to the trained model for classification. The output consists of the predicted labels ρ_1', ρ_2', ρ_3', . . ., which aim to match the original generative parameter values of the corresponding dynamical systems.

[7], have also evolved into quantitative techniques for assessing dynamical complexity through metrics like recurrence rate, determinism, and entropy [5]. Symbolic encodings of recurrence, including RMs [9] and entropy maximization strategies [12,17], provide compact and interpretable features for classification tasks. More broadly, recurrence quantification analysis (RQA) has been successfully applied across disciplines: in climate science to detect regime shifts [18,19], in health to analyze cardiac and respiratory signals [20,21], and in economics to identify financial regimes and instabilities [13,22].

It is believed that these two approaches, SD [6] and RM [9,17] from recurrence analyses provide complementary insights into the underlying dynamics. In this work, we evaluate their effectiveness in classifying time series generated by different nonlinear systems under varying parameter settings. For the classification task, we employ a Random Forest (RF) algorithm, a widely used and interpretable machine learning method.

In this work, we address three main questions: (1) Do the recently developed Recurrence Microstates provide a complementary or supplementary approach to traditional Symbolic Dynamics? (2) Are certain symbol probabilities more informative than others for parameter classification tasks? (3) Is there any invariance with respect to the type of system, and does RM consistently outperform SD? These questions are central to understanding how SD and RM methods interact within machine learning classification (MLC). Moreover, our analysis contributes to the interpretability of MLC by offering physical insights into the underlying mechanisms of this newly defined feature space.

To address these questions, we construct a synthetic dataset comprising eight chaotic dynamical systems, ranging from one- and two-dimensional nonlinear maps to well-known continuous systems. For each system, we vary a single control parameter and segment the resulting time series into multiple classes based on parameter regimes. Feature vectors are then extracted from the probability distributions of RMs and SD, and these features are used as inputs to a Random Forest classifier for parameter classification. Finally, to test the robustness and applicability of our approach, we extend the analysis to a real-world dataset of 12-channel electrocardiogram (ECG) signals from patients exhibiting four distinct cardiac rhythms. This step highlights the potential of combining RM and SD features for the classification of complex biological signals. A schematic illustration of the classification procedure applied to all parameter-dependent time series is depicted in Fig. 1.

This paper is organized as follows: Section II presents the methodological framework, including details about the dataset, feature extraction based on RM and SD, and the machine learning algorithms employed. In Sec. III, we report and analyze the classification results. Finally, Sec. IV provides a discussion of the findings and concluding remarks.

II. METHODS

A. Recurrence microstates

Recurrence plots are graphical constructions obtained from the recurrence matrix [7]. The recurrence matrix is computed from a data series consisting of a sequence of K data states in

an M-dimensional space as

$$\mathbf{R}_{ij} = \Theta(\varepsilon - |\mathbf{x}_i - \mathbf{x}_j|), \tag{1}$$

where ε is a tolerance radius and Θ is the Heaviside function. If the distance between the dynamic states $\mathbf{x}_i - \mathbf{x}_j$ is less than the threshold ε, the states are considered recurrent ($\mathbf{R}_{i,j} = 1$); otherwise, the event is classified as nonrecurrent ($\mathbf{R}_{i,j} = 0$).

Important quantification can be derived from the statistical properties of recurrent and nonrecurrent structures within a given recurrence matrix. This type of symbolic analysis of dynamics is well known and extensively explored in the specialized literature [5]. Studies have investigated the quantification of diagonal structures [23], vertical structures [24], and nonrecurrent elements, among others [25]. These measures have already been applied in diverse fields such as physics, econophysics, climatology, and health [5].

A natural generalization of the concept of recurrence of states in phase space is obtained by considering not only single states but also all possible recurrences of *sequences* of consecutive states. As an illustrative example, let us take two short sequences of length $N = 2$ embedded in an M-dimensional data

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \underbrace{\mathbf{x}_i, \mathbf{x}_{i+1}}, \ldots, \underbrace{\mathbf{x}_j, \mathbf{x}_{j+1}}, \ldots, \mathbf{x}_{K-1}, \mathbf{x}_K. \tag{2}$$

This leads to the concept of an RM [9,17], defined as the set of all possible recurrence relations between two short sequences of size $N \ll K$, randomly selected from the dataset. For the case $N = 2$, the recurrences given by Eq. (1), when applied to the pairs $\mathbf{x}_i, \mathbf{x}_{i+1}$ and $\mathbf{x}_j, \mathbf{x}_{j+1}$, yield a total of 16 distinct RMs. These can be represented as four-digit binary numbers ranging from $\mathrm{RM}_1 = \{0, 0, 0, 0\}$ to $\mathrm{RM}_{16} = \{1, 1, 1, 1\}$. In general, for an arbitrary sequence length N, the number of possible RM is $\mathcal{N} = 2^{N^2}$.

In this way, RMs are submatrices extracted from the recurrence plot that encapsulate both local and nonlocal information about the relationships between different parts of the system [9].

Shannon entropy is a key measure that can be derived from the probability distribution of the RM. The probability distribution can be calculated as follows:

$$P(\mathrm{RM}) = n(\mathrm{RM})/s, \tag{3}$$

where s is the total number of samples and $n(\mathrm{RM})$ is the number of occurrences of a specific RM. The recurrence entropy can then be calculated as

$$S(\varepsilon) = -\sum_{i=1}^{\mathcal{N}} P(\mathrm{RM}_i) \ln[P(\mathrm{RM}_i)]. \tag{4}$$

The concept of maximum entropy can be used to automatically determine the tolerance radius ε [17] and the time sampling τ in continuous dynamical systems [26]. Additionally, the occurrence probabilities, Eq. (3), of RMs derived from a time series of length K can serve as an equally relevant set of quantifiers.

These probabilities have been shown to be important features for the classification of dynamical systems [12]. The equal probability distribution of RMs in stochastic systems, particularly for groups with $N = 2$, was explored as a means to distinguish stochastic from chaotic systems [11]. The submatrices can be graphically illustrated as black-and-white boxes, as shown in Fig. 2 for $N = 2$.

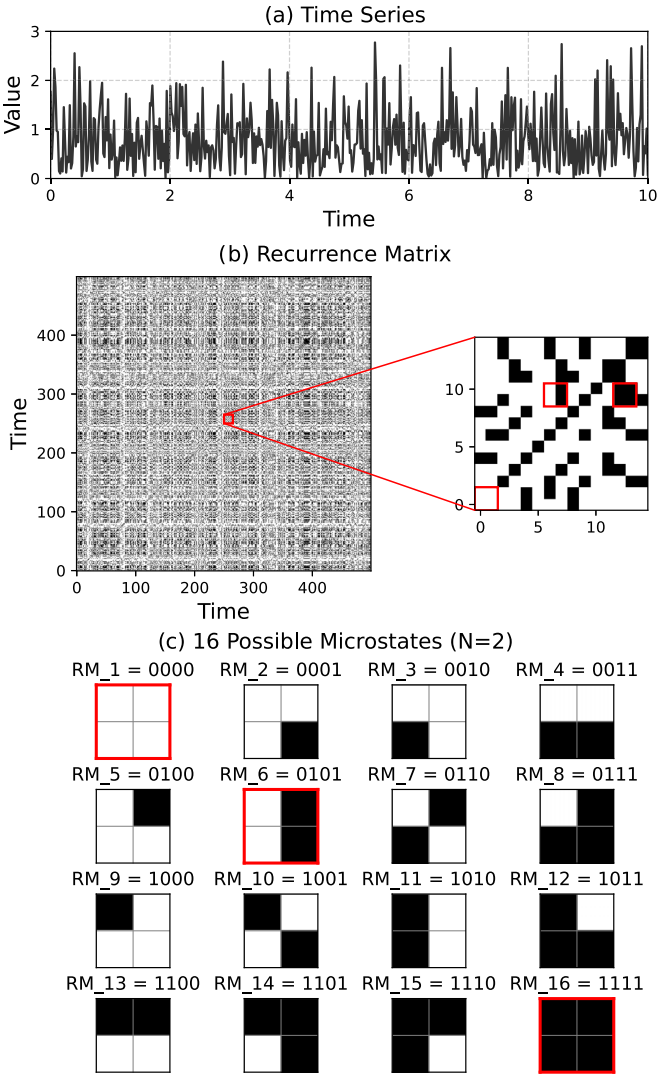


FIG. 2. Procedure to define the RMs from a time series generated by data [panel (a)]. Recurrent plot computed based on the data [panel (b)] from which submatrices known as RMs can be randomly extracted. In a Recurrence Matrix, RMs can be defined [shown as red squares in the inset of panel (b)]. Panel (c) depicts the 16 possible RMs with 2×2 structure.

B. Symbolic dynamics based on variation patterns

In this work, we introduce an SD approach based on the sequence of monotonic variations, specifically, the detection of ups and downs, within a sliding window of the time series. This method is conceptually related to ordinal analysis [4] but differs by focusing solely on the direction of successive changes, rather than the full rank permutation of the values. Given a window of N consecutive data points, $\{x_1, x_2, \ldots, x_L\}$, we compute a binary symbolic sequence of length $L - 1$ by comparing each pair of adjacent values. For

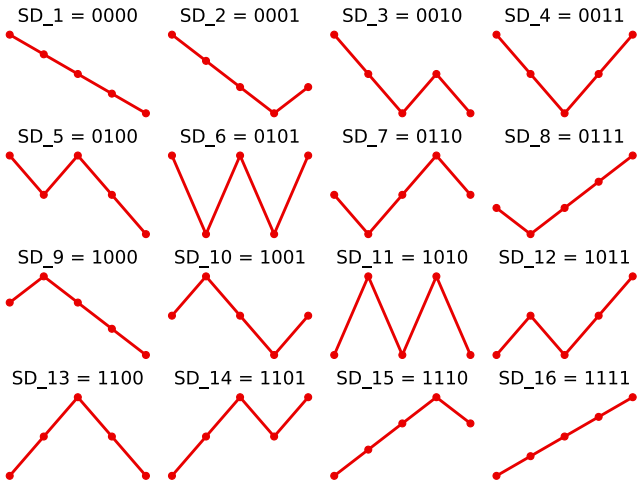


FIG. 3. Graphical representation of the 16 ordinal binary patterns that can be defined from five data points in a time series, using Eq. (5): "ups" ($x_i < x_{i+1}$) are represented by "1s" and downs ($x_i \geqslant x_{i+1}$) by "0s".

each position i, we define the symbolic component s_i as

$$s_i = \begin{cases} 1, & \text{if } x_{i+1} > x_i \quad \text{(up)} \\ 0, & \text{if } x_{i+1} \leqslant x_i \quad \text{(down)} \end{cases} \quad \text{for } i = 1, \ldots, L-1. \tag{5}$$

This binary representation [6] maps each data window to a binary word of length $L-1$, which encodes the pattern of variations. For instance, the sequence $\{4, 5, 3, 1\}$ produces the symbolic word $(1, 0, 0)$, as $5 > 4$ (up), $3 < 5$ (down), and $1 < 3$ (down). In this framework, the number of possible symbolic patterns is 2^{L-1}. Figure 3 illustrates all 16 symbolic patterns obtained from segments of $L = 5$ points. The patterns are labeled similarly to the RMs introduced in the previous section, following a binary rule in which *ups* are represented by "1" and *downs* by "0". In contrast to the recurrence-based approach, which characterizes the system by identifying recurrences in phase space and thus captures the geometrical organization of trajectories through neighborhoods of reference points, this variation-based SD method focuses on the local dynamical variations of the time series. By encoding only the direction of successive changes, the symbolic patterns emphasize how the system evolves from one state to the next, rather than how often it revisits similar regions of state space. This variation-based SD method offers a simplified representation of the time series structure. It retains information about the local dynamics while avoiding the complexity and redundancy often present in full permutation-based ordinal analyses. Notably, the number of symbolic patterns generated for a given window size is directly comparable to the number of RMs obtained in the recurrence-based approach, allowing for consistent comparisons between the two frameworks. One advantage of this methodology, compared to the RM approach, is the absence of a recurrence threshold. Additionally, while computing the Shannon entropy of RM often requires optimizing the threshold to maximize S, the ordinal patterns in this method are computed only once.

C. Datasets

In this paper, we analyze well-known [1] time series generated by six discrete and two continuous chaotic systems, and an empirical open dataset of ECG data from the PhysioNet database [27]. All the synthetic time series evolutions of the dynamical systems were performed using the *DynamicalSystems.jl* package [28] in the Julia language and *ScikitLearn* [29] in Python. The systems are described below.

1. Synthetic chaotic systems

(1) The generalized Bernoulli chaotic map, also known as the βx map, is described by

$$x_{n+1} = \beta x_n \ (\text{mod}1), \tag{6}$$

where β is a positive parameter that controls the dynamical characteristics of the map. The term "mod1" means that only the fractional part of the product βx is retained. This operation ensures that x_n is limited in $[0, 1)$.

(2) The Gauss map is described by

$$x_{n+1} = e^{-\alpha x_n^2} + \gamma, \tag{7}$$

where α controls the sharpness of the decay and γ is a constant offset.

(3) The Hénon map is a two-dimensional discrete-time dynamical system introduced by Michel Hénon in 1976 as a simplified model of the Lorenz system's Poincaré section [30], and is defined by

$$\begin{aligned} x_{n+1} &= 1 - a x_n^2 + y_n, \\ y_{n+1} &= b x_n, \end{aligned} \tag{8}$$

where a and b are real parameters.

(4) The Ikeda map is a two-dimensional discrete-time dynamical system that models the behavior of light in an optical cavity with a nonlinear medium [31]. The system is given by

$$\begin{aligned} x_{n+1} &= 1 + u(x_n \cos\theta_n - y_n \sin\theta_n), \\ y_{n+1} &= u(x_n \sin\theta_n + y_n \cos\theta_n), \\ \theta_n &= k - \frac{p}{1 + x_n^2 + y_n^2}, \end{aligned} \tag{9}$$

where u is a dissipation parameter, and k and p control the strength and nonlinearity of the phase shift θ_n.

(5) The well-known logistic map [1] is a simple yet powerful tool for illustrating chaotic transitions:

$$x_{n+1} = r x_n (1 - x_n), \tag{10}$$

where r is a control parameter.

(6) The tent map [1] is another classical example of a one-dimensional map that exhibits chaotic behavior depending on the control parameter $r \in [0, 2]$:

$$x_{n+1} = \begin{cases} r x_n, & \text{if } x_n < \frac{1}{2}, \\ r(1 - x_n), & \text{if } x_n \geqslant \frac{1}{2}. \end{cases} \tag{11}$$

(7) The Lorenz system [32] is a classical continuous-time dynamical system originally developed as a simplified model for atmospheric convection. It is defined by the set of three

coupled nonlinear differential equations:

$$\frac{dx}{dt} = \sigma(y - x),$$
$$\frac{dy}{dt} = x(\rho - z) - y,$$
$$\frac{dz}{dt} = xy - \beta z, \tag{12}$$

where σ, ρ, and β are positive parameters.

(8) The Rössler system [33] is a continuous-time dynamical system introduced as a simplified model to explore chaotic behavior. It is defined by the following set of differential equations:

$$\frac{dx}{dt} = -y - z,$$
$$\frac{dy}{dt} = x + ay,$$
$$\frac{dz}{dt} = b + z(x - c), \tag{13}$$

where a, b, and c are real parameters.

For continuous systems, the time sampling affects the sequencing, but it was carefully selected to maximize entropy (i.e., minimize mutual information) [17,34], giving $dt = 0.25$ for Lorenz (7) and $dt = 1.1$ for Rössler (8).

2. *Empirical dataset*

(9) ECG dataset from Physionet [27,35,36], which contains over 40 000 time series, each with 5000 data points recorded across 12-lead channels. Although the dataset includes dozens of different ECG conditions, most are rare, so we focused on the four most frequent: atrial fibrillation (AF), sinus bradycardia (SB), sinus rhythm (SR), and sinus tachycardia (ST). These four conditions account for 19 524 subjects, if we selected subjects with only one isolated condition: 3224 AF, 8909 SB, 5908 SR, and 1483 ST. Among these, sinus rhythm is the typical healthy heart rhythm, while the others are considered abnormalities.

D. Random forest classifier

We employ the Random Forest algorithm for classification tasks using the `RandomForestClassifier` implementation from the `scikit-learn` library [29]. The features used for training were the symbolic probability distributions extracted from dynamical patterns, either based on RMs, Sds, or both combined (resulting in 16 or 32 features, respectively). The target variable is the parameter value of the dynamical system, discretized into 10, 20, or 40 classes; however, we present only the results with 10 classes for visualization through confusion matrices. The datasets are partitioned into three independent subsets: 60% for training, 20% for validation, and 20% for final testing. To ensure statistical robustness and prevent overfitting, we implement a fivefold cross-validation protocol within the training set for model selection, while the validation set is used to monitor performance and tune hyperparameters, ensuring that the test set remains strictly isolated for the final evaluation of the model's generalization across different dynamical regimes. The classifier is configured with 100 decision trees (`n_estimators = 100`) and a fixed seed (`random_state = 42`) to ensure reproducibility.

Using the Random Forest classifier from the `scikit-learn` library, the input data containing the extracted features from the symbolic analysis are loaded. Each sample includes 16 features from RMs and/or 16 features from SD, which are concatenated into a single feature vector of length 32. The classifier is then trained to predict the class corresponding to the value of the dynamical system parameter, based on the probability distributions associated with these patterns. After training, the relative importance of each feature is extracted using the `clf.feature_importances_` attribute, which quantifies how much each feature contributes, on average, to reducing impurity in the decision nodes of the forest trees. This process is repeated across five different files, and the results shown represent the average feature importance over these five runs.

III. RESULTS

In this section, we evaluate the effectiveness of nonlinear representations, specifically, RM and SD defined in Sec. II, in classifying time series generated by various chaotic dynamical systems. We aim to test whether these representations, combined with machine learning algorithms, can successfully distinguish between subtly different dynamical regimes. To that end, all systems described in Sec. II C are analyzed independently. For the multidimensional systems, we analyze the x variable.

Each system is divided into C parameter classes, with each class corresponding to time series generated from distinct but closely related parameter values designed to produce similarly chaotic, yet distinguishable, dynamics. Within each class, 100 time series are simulated from different initial conditions to ensure statistical robustness. Classification performance is assessed both qualitatively, using the confusion matrix, and quantitatively, via accuracy scores.

For each system, we define a parameter range $(p_i,\ p_f]$ in which one of the system's parameters is varied to generate different classes. Given C total classes, the interval is divided into C uniformly spaced values, excluding the lower bound p_i, which is used solely to define the start of the range. The greater the number of classes, the closer the parameter values are between adjacent classes. As an illustrative case, for $C = 10$ classes, the βx map [Eq. (6)] is explored by uniformly sampling the parameter β in the interval $(p_i = 1.99,\ p_f = 6.99]$, resulting in ten distinct values: [2.49, 2.99, 3.49, …, 6.99]. These parameter values generate slightly different chaotic regimes. Table I summarizes the full range of parameter variations used for all dynamical systems analyzed.

A. Recurrence microstate probabilities as features for parameter classification

First, we test the 16 RMs as features for the six discrete maps, considering $C = 10$ distinct classes. Figure 4 shows the confusion matrices for (a) the βX map, (b) the Gauss map, (c) the Hénon map, (d) the Ikeda map, (e) the Logistic map, and (f) the tent map. The y axis represents the true parameter values that generated the data series, while the x axis shows

TABLE I. Parameter variation ranges for each dynamical system $(p_i,\ p_f]$. The parameter is varied linearly in this range in C steps, excluding p_i.

System	Parameter	Lower bound (p_i)	Upper bound (p_f)
βX	β	1.99	6.99
Gauss Map	γ	−0.7	−0.3
	α (fixed)	6.2	–
Hénon Map	a	1.1	1.2
	b (fixed)	0.3	–
Ikeda Map	u	0.6	0.8999
Logistic Map	r	3.6	4
Tent Map	μ	1.8	1.9999
Lorenz System	ρ	27.99	37.99
	σ (fixed)	10	–
	b (fixed)	8/3	–
Rössler System	a	0.20	0.30
	b (fixed)	0.2	–
	c (fixed)	5.7	–

the parameters predicted by the random forest algorithm. The numbers inside the cells associated with the color code from white 0 to dark red 100 correspond to the percentage of classification, which is fully accurate when the diagonal lines are filled with 100.

We observe that, for each system, the predictions are mostly accurate, with only a few instances falling outside the main diagonal of the confusion matrix. Notably, the Hénon map exhibits unexpected behavior for one of the parameter classes, where the predicted value of $a = 1.17$ is not consistently classified and instead appears scattered across all possible classes. This discrepancy likely arises from subtle dynamical transitions occurring within regions typically associated with chaos. In particular, we detect a transition near $a \approx 1.1723$ due to intermittency, where the chaotic attractor gives way to a periodic orbit. When the chaotic attractor is still present ($a < 1.1723$), one can observe "nearly-periodic" trajectories, intermittently interrupted by bursts of chaotic behavior [1]. These pseudoperiodic windows interfere with our classification approach by generating time series with ambiguous signatures. As a result, such cases pose a genuine challenge for the random forest algorithm, which struggles to distinguish them from nearby dynamical regimes.

B. Symbolic dynamics probabilities as features for parameter classification

Next, we consider a similar scenario shown in Fig. 5, but using only the 16 symbols from SD as features for the random forest algorithm applied to the six discrete maps.

In general, the classification performance using SD features shows reduced accuracy compared to the results obtained with RMs, as illustrated in Fig. 4, with several predictions falling outside the main diagonal of the confusion matrix. This outcome is expected, since previous studies have demonstrated that RM features are strongly correlated to dynamical properties of the system, for example the Lyapunov exponent [9], providing a more reliable representation of the

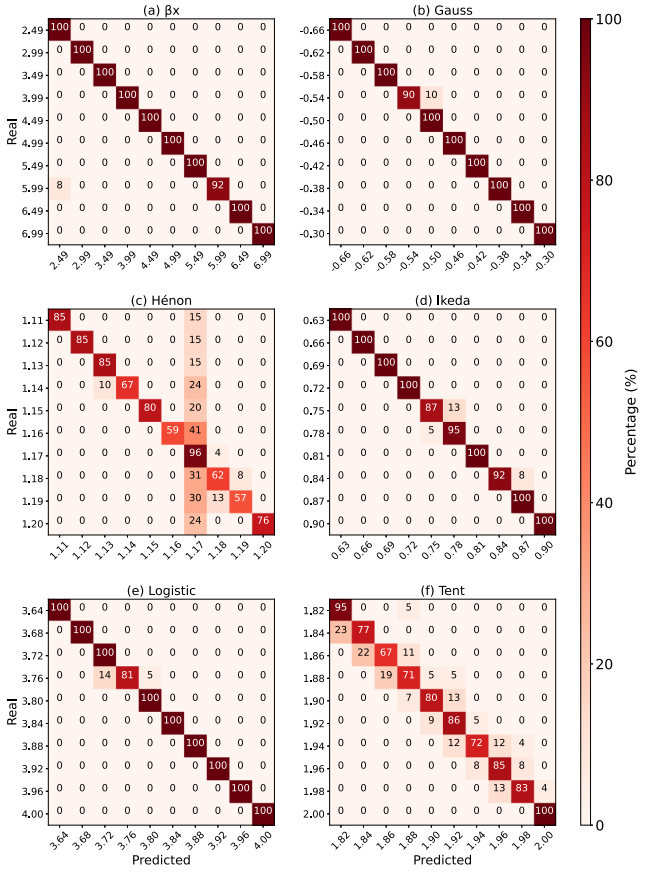


FIG. 4. Performance of the random forest algorithm when the input features are the occurrence probabilities of the 16 RMs shown in Fig. 2(c). Panels (a)–(f) show the confusion matrices obtained for the six maps analyzed (βX, Gauss, Hénon, Ikeda, Logistic, and tent, respectively). For each map, 100 time series were classified into ten classes according to the parameter value used to generate the time series. The time series were split evenly into 60% for training, 20% for validation, and 20% for final testing. Each time series contained 1000 data points that followed the first 1000 transient points that were disregarded. Overall, a high classification accuracy can be observed across all maps, with the exception of one class for the Hénon map (as discussed in the text).

system's underlying dynamics. While visual inspection already suggests a clear advantage of RM for data reduction in machine learning classification tasks, it is particularly noteworthy that the transition of the Hénon map near $a = 1.17$, which challenged the RM-based classification, does not pose a problem when SD features are used. This observation suggests that combining both RM and SD features may lead to improved classification performance, offering a more robust joint feature representation.

C. Joint features for parameter classification

Here, we propose combining both methods, the 16 features of RMs and 16 from SD, to form a new space of 32 features for the random forest algorithm. Figure 6 demonstrates visually that the results suffer an improvement compared to those shown in Figs. 4 and, especially, 5.

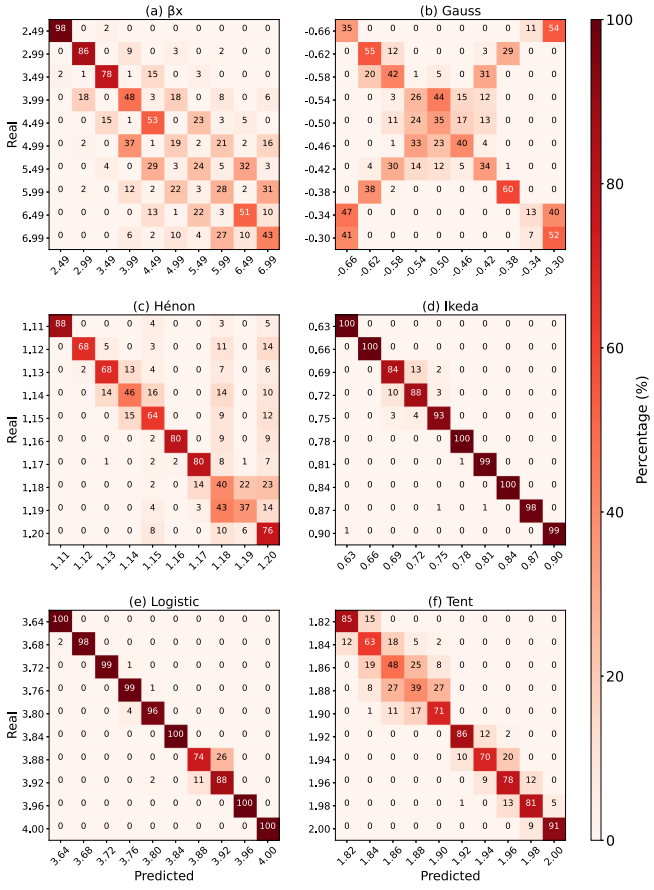


FIG. 5. Performance of the random forest algorithm when the input features are the occurrence probabilities of the 16 ordinal binary patterns shown in Fig. 3. Panels (a)–(f) show the confusion matrices obtained for the six maps analyzed; the datasets and training, validation, and testing procedures are as in Fig. 4. We observe good classification accuracy only for the Ikeda and Logistic maps, whereas the Gauss map exhibits a high level of misclassification.

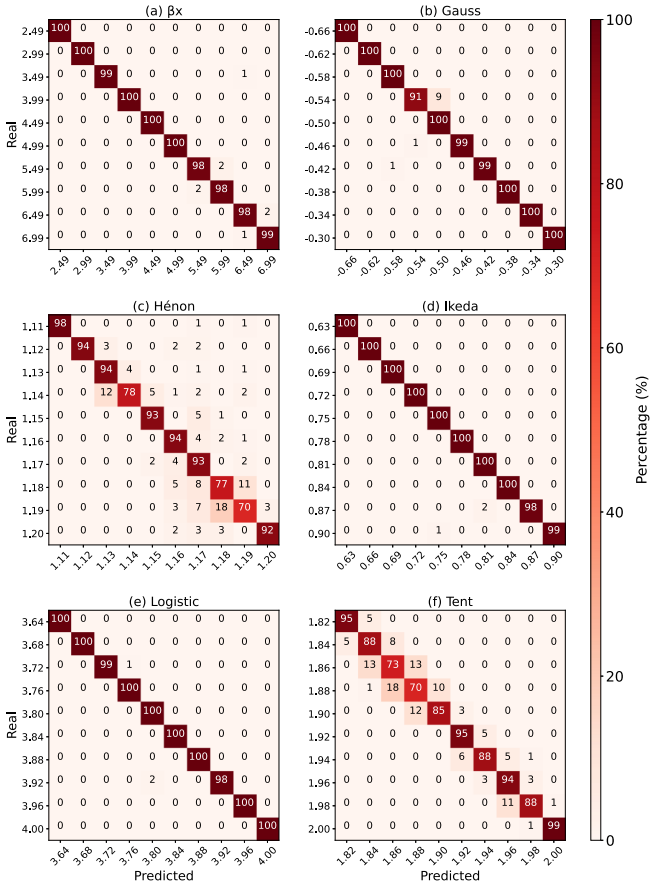


FIG. 6. Confusion matrices obtained for the six maps analyzed when the input features are the probabilities of 16 RMs [shown in Fig. 2(c)] and of 16 ordinal binary patterns (shown in Fig. 3). The datasets and training, validation, and testing procedures are as in Figs. 4 and 5. We observe a significant increase in accuracy with respect to Figs. 4 and 5.

We observe that, for most dynamical maps, the results using only RM are similar to those obtained with the combined features. However, when RM alone underperformed, the inclusion of SD features appears to solve the issue. This complementarity reflects the sensitivity of RM to certain system characteristics in which SD is more robust, and vice versa. This approach offers the best of both worlds: a compact feature space, high classification accuracy, and a fast technique grounded in fundamental physical concepts.

The analysis can be extended to continuous dynamical systems. Consider Fig. 7, where column (a) corresponds to the Lorenz system and column (b) to the Rössler system. Each row represents a different feature set: the first row uses RM, the second row uses SD, and the third row combines RM and SD. We observe a pattern similar to that seen in discrete dynamical systems. In general, the combination of both methodologies (RM and SD) yields superior classification performance across both continuous cases. However, the Lorenz system shows notably poor performance. Interestingly, the misclassifications often occur near the main diagonal of the confusion matrix, suggesting that the model tends to confuse classes with very similar dynamics. This observation indicates that performance may be improved by increasing the number of features extracted or by employing a more complex classification algorithm (e.g., a deeper or more optimized Random Forest model).

To stress our methodology, we consider the dataset composed of an experimental ECG database from Physionet [27,35,36]. Although the dataset includes dozens of different ECG conditions, most of them are rare, so, in this work, we focused on the four most frequent: atrial fibrillation (AF), sinus bradycardia (SB), sinus rhythm (SR), and sinus tachycardia (ST). These four conditions account for approximately 20 000 events (see Sec. II C for details). The dataset was partitioned into 70% for training and 30% for testing. Each time series was further segmented into five windows of 1000 data points each. For every segment and across all 12 channels, RM and SD features were computed in parallel, yielding a total of $12 \times 32 = 384$ features per subject, with half derived from SD and half from RM. Figure 8 shows the confusion matrices for three situations: (a) using only RM features, (b) only SD features, and (c) both feature sets combined, as done previously for the discrete and continuous dynamical systems.

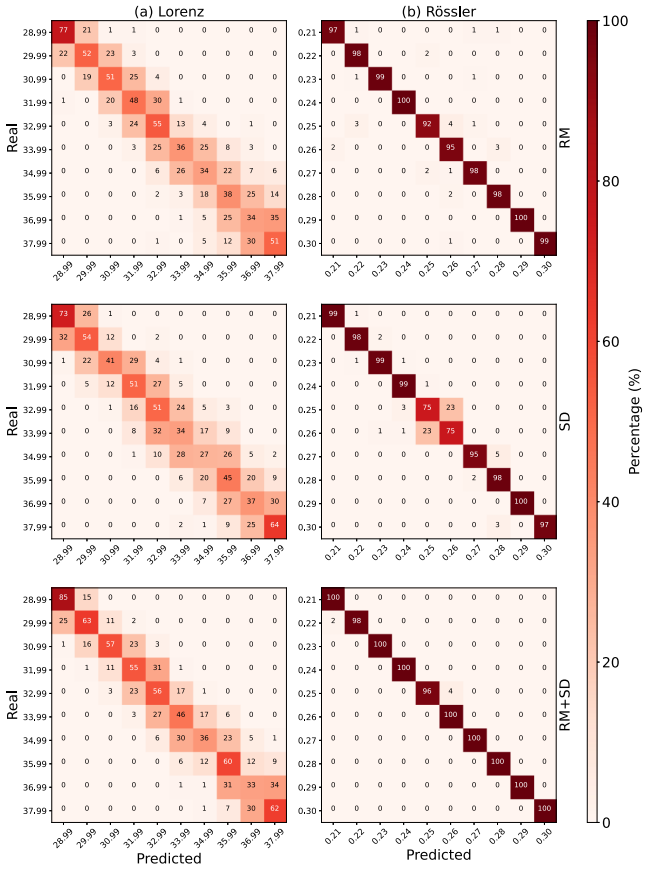


FIG. 7. Confusion matrices obtained for the two continuous dynamical systems analyzed (Lorenz in the left column, Rössler in the right column) when the input features are the probabilities of 16 RMs (first row), of 16 ordinal binary patterns (second row), and of both (third row). The length and number of time series used, and the training, validation, and testing procedures are as in Figs. 4–6.

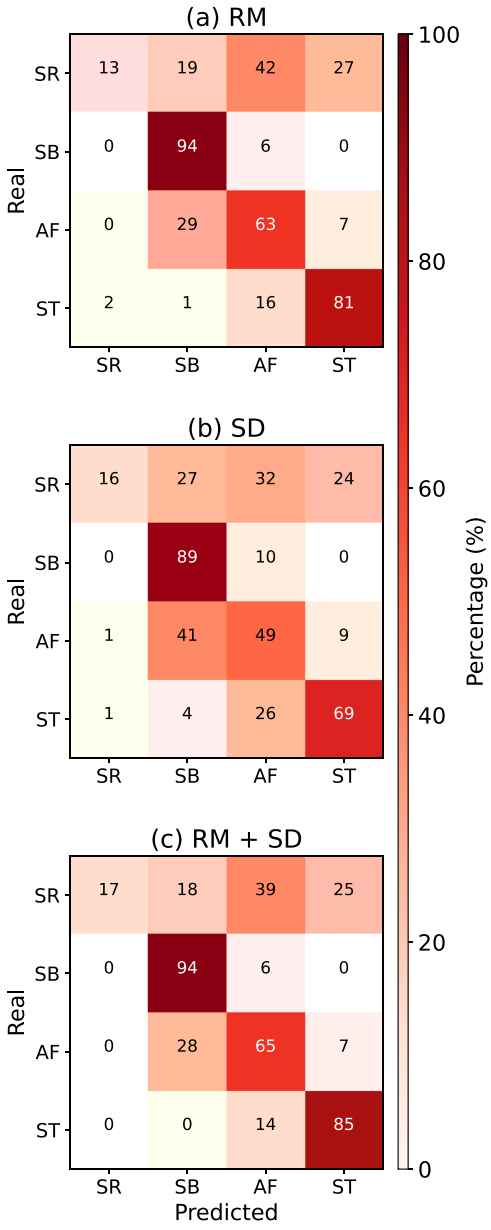


FIG. 8. Confusion matrices obtained when ECG recordings were classified in four cardiac conditions (atrial fibrillation, sinus bradycardia, sinus rhythm, and sinus tachycardia), using as features the probabilities of 16 (a) RMs, (b) symbolic ordinal patterns, and (c) both. See Sec. II C 2 for the description of the dataset. Sixty percent of the data were used for training, 20% for validation, and 20% for testing. We observe a high classification accuracy that improves from symbolic analysis [panel (b)] to RM analysis [panel (a)], and further with the combination of both [panel (c)].

Similar to the previous cases analyzed here, the combination of RM and SD features yields the highest classification accuracy. Overall, the SB class is most accurately predicted, followed by ST and SR, while AF consistently represents the most challenging scenario for correct classification. This result can be partially attributed to the imbalance present in the dataset, where SB subjects represent approximately 45% of the dataset. As a consequence, the classifier tends to perform better on this class. A similar pattern has been reported in Ref. [37], where the authors observed the same accuracy ranking (SB, ST, SR, followed by AF) using recurrence-based entropic features. These findings suggest that both the data distribution and the nature of the extracted features play a critical role in determining classification performance.

The ECG dataset used in this study was selected due to its large number of labeled recordings and the presence of several common cardiac rhythms, including sinus rhythm, atrial fibrillation, bradycardia, and tachycardia. Similar datasets have been explored in previous works using a variety of classification methodologies, particularly deep learning models [38,39], which have reported high classification accuracies, often exceeding 90%, especially when using convolutional architectures trained on large datasets. Our objective here is not to compete with such methods but to assess whether symbolic features, derived from time series dynamics alone, can provide a compact and interpretable alternative. In this context, the overall accuracy achieved (around 80%) is promising. We also note that the difficulty in correctly identifying atrial fibrillation is not unique to our approach and has been reported in prior studies using traditional feature-based methods [37,40,41]. Nonetheless, to the best of our knowledge, the

TABLE II. Summary of classification accuracies using RM, SD, and both (RM and SD) as features for a Random Forest classifier. Results are reported for six discrete maps, two continuous systems, and an ECG dataset. The bold face values are the highests accuracies for each data set.

Data Class	RM	SD	RM + SD
βX map	**0.9950 ± 0.0000**	0.5220 ± 0.0211	0.9920 ± 0.0060
Gauss Map	**0.9900 ± 0.0000**	0.3770 ± 0.0583	0.9890 ± 0.0049
Hénon Map	0.7550 ± 0.0000	0.6560 ± 0.0193	**0.8840 ± 0.0136**
Ikeda Map	0.9800 ± 0.0000	0.9570 ± 0.0098	**0.9970 ± 0.0040**
Logistic Map	0.9800 ± 0.0000	0.9530 ± 0.0133	**0.9980 ± 0.0024**
Tent Map	0.8100 ± 0.0000	0.7080 ± 0.0232	**0.8700 ± 0.0195**
Lorenz Oscillator	0.4710 ± 0.0314	0.4650 ± 0.0202	**0.5410 ± 0.0220**
Rössler Oscillator	0.9780 ± 0.0093	0.9390 ± 0.0143	**0.9960 ± 0.0037**
ECG Data	0.7610 ± 0.0072	0.6830 ± 0.0062	**0.7788 ± 0.0035**

classification strategy presented here, combining recurrence-based and variation-based symbolic features, has not yet been applied to this particular ECG dataset.

The advantages of combining RM and SD as features in the Random Forest classifier are summarized in Table II and bar-plotted in Fig. 9. We can clearly observe that in most cases (except in the Gauss map), the RM-SD combination results in higher or comparable accuracy when compared to using RM alone, including for the more challenging experimental dataset. To gain better insight into how this approach works, we can directly extract from the random forest algorithm the features that carry the most weight in the classification process (Sec. II D for details). In Fig. 10, we can see the proportional weight relevance of each feature from RM and SD for the classification. Moreover, we evaluate the impact of increasing the number of classes for each dynamical system and analyze the specific classification weight assigned to each feature type (SD or RM). It is evident that while the number of classes has a relatively small effect on feature importance, the nature of the dynamical system predominantly determines which features are most relevant. For example, in panel (b), several symbols from SD contribute little or nothing to the classification process. Conversely, for the Ikeda map, SD features appear to be as important as RM features. Overall, RM features show greater stability; in every system studied, all RM features contribute meaningfully to parameter classification. Two main explanations arise: First, in SD, much of the information carried by one symbol may overlap with that of another; this redundancy is less evident and direct in RM. Second, the complementarity provided by SD arises primarily from capturing upward and downward trends, whereas recurrence analysis considers only the magnitude of distances between dynamical states. Although this aligns with the theoretical framework of the methods, it inevitably results in some loss of information, which underpins the observed complementarity between SD and RM.

In Fig. 11, we have a similar bubble for ECG classification. The scenario is much more challenging due to the experimental nature of the data. We can see several asymmetries,

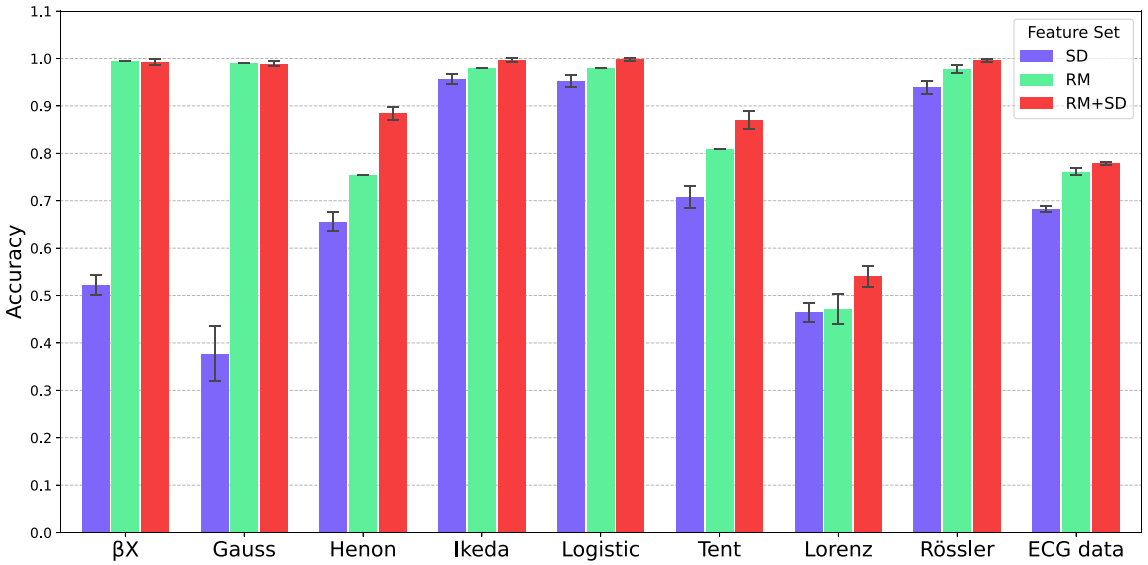


FIG. 9. Classification accuracy across dynamical systems and ECG data. Bars show mean accuracy (with standard deviation) for: SD in green, RM in blue, and RM + SD (both methods) in red. The combined RM + SD approach consistently outperformed individual methods.

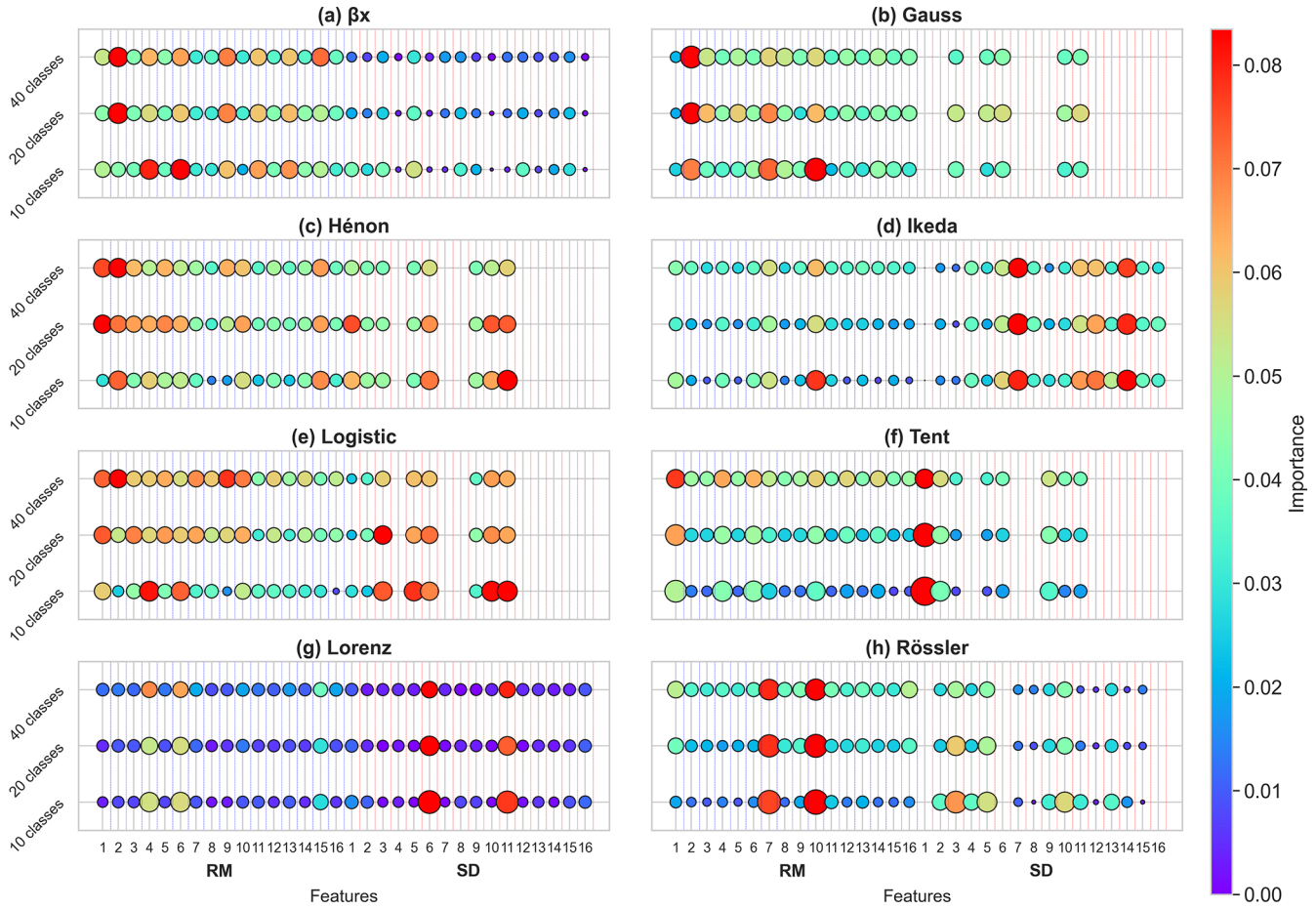


FIG. 10. Mean feature importance of the descriptors extracted from time series generated by eight different dynamical systems, considering classifications into 10, 20, and 40 classes. For each system, five repetitions of time series with 1000 data points (after discarding 1000 transient steps) were generated using different values of dynamic parameters. The features correspond to the probabilities of RMs and SD patterns with 16 symbols (SD), combined into a single input vector. The importance of each feature was estimated using a Random Forest classifier with 100 trees, trained on the combined RM and SD dataset. Importance values were averaged across repetitions. The size and color of each circle indicate the relative importance of the feature, with visual separation between the RM group (on the left, delimited by dashed blue vertical lines) and the SD group (on the right, delimited by dashed red vertical lines).

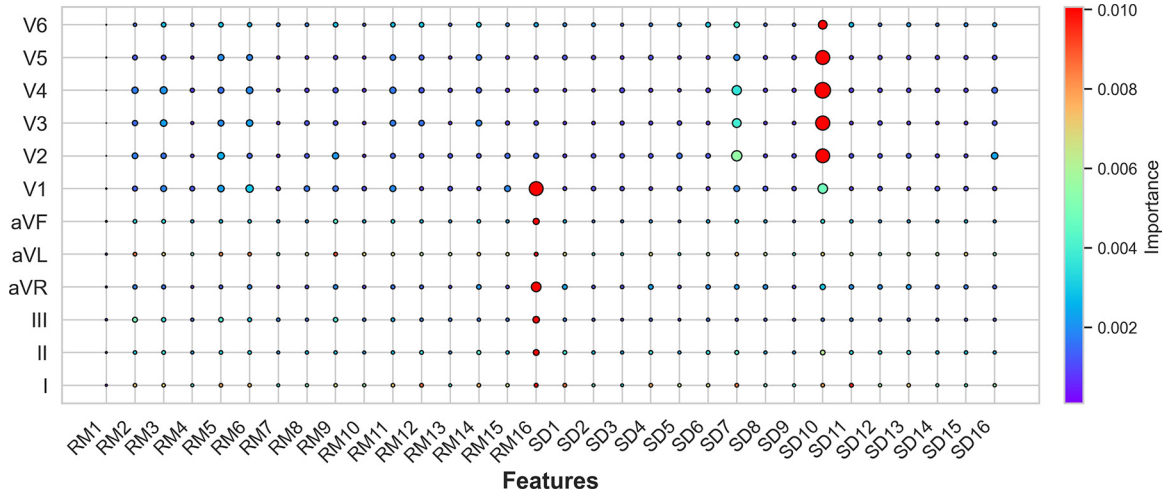


FIG. 11. Feature importance across 12 ECG channels based on RM and SD descriptors. Each row corresponds to a specific ECG channel, and each column represents one of the 32 features (16 RM and 16 SD) used in the classification task. The size and color of the circles indicate the relative importance of each feature for that channel, as estimated by a Random Forest classifier with 100 trees. Importance values were normalized and averaged over five repetitions.

from the RM and SD contribution, to the particularities of each ECG channel. For instance, every six V channels have a greater contribution than the other channels. Furthermore, there are some channel symbols that almost do not contribute to the classification. Of particular attention are RM16 and SD10. The first one is the complete RM, while the latter is a symbol connected to local fluctuations. Moreover, SD7 is also relevant in the V channels. This symbol is just an axis reflection of SD10. The joint evaluation reveals a promising insight: Dynamical fluctuations are key ingredients for the classification.

IV. CONCLUSION

In this work, we have proposed a framework for time series analysis to classify synthetic time series simulated using similar parameters, or empirical time series (ECG recordings) that correspond to four cardiac pathologies. The classification is carried out with a standard Random Forest algorithm, using two types of features: the probabilities of 16 RMs, and the probabilities of 16 ordinal binary patterns, defined using SD. These methodologies were applied to a synthetic dataset composed of six discrete and two continuous dynamical systems, as well as to an empirical dataset of 12-channel ECG recordings from cardiac patients. Both approaches demonstrated substantial dimensionality reduction in the feature space while maintaining high classification accuracy across the different dynamical systems.

Building upon this framework, we conducted a comparative analysis of the RM and SD feature sets. In all cases examined, both simulated and empirical, RM consistently yielded higher classification accuracy. Additionally, RM features are more compact: They require fewer data points, needing only four states to generate the same number of possible patterns that SD would require five for. This advantage becomes even more pronounced at larger scales; for instance, generating 512 symbols requires only six data points with RM, while SD would need ten. However, a notable drawback of the RM approach is its dependency on a recurrence threshold, which must often be fine-tuned to achieve optimal entropy values. This is generally avoided with techniques that optimize the choice of threshold [17], but the trade-off is an increase in the computation time. SD provides complementary information that RM alone may overlook. This is evident, for example, in the Hénon map, where RM occasionally misclassifies parameters near a transition point, an error that SD can avoid. SD captures local directional trends, such as negative gradients, whereas RM emphasizes the magnitude of distances in phase space. Consequently, combining RM and SD features often enhances or stabilizes classification performance, highlighting the benefits of integrating both approaches in the analysis of complex dynamical systems.

Regarding the synthetic dataset, the lowest classification accuracy is observed for the Lorenz system, where only half of the samples are correctly classified. Analyzing the confusion matrix reveals that the misclassified samples tend to fall near the main diagonal, suggesting that the classification errors occur between dynamically similar parameter regimes. This highlights the intrinsic challenge of distinguishing transitions in continuous systems like Lorenz, and suggests that future work should focus on improving classification accuracy in such cases. As a final test, we applied our methodology to an empirical dataset composed of ECG time series from four distinct cardiac rhythms. The overall accuracy was close to 80%, with most rhythms accurately classified. However, the model struggled to correctly identify atrial fibrillation, indicating the need for further refinement in detecting this rhythm.

To further explore the potential of the proposed methodology, future work should assess its robustness under more realistic conditions, such as the presence of observational noise or reduced time series length. Another promising line of investigation is to examine whether the combination of RM and SD features can reliably identify the underlying system that generated a time series, for example, distinguishing between signals produced by the Logistic map, the Hénon map, or a hybrid case involving contamination from multiple sources. Exploring these scenarios may enhance the interpretability and applicability of symbolic features in complex dynamical systems.

ACKNOWLEDGMENTS

This work was supported by the following Brazilian research agency: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grants No. 407072/2022-5, No. 305189/2022-0, No. 408254/2022-0, and No. 300064/2023-3; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Grant No. 88881.895032/2023-01. B.R.R.B. would like to thank FAPESP—Fundação de Amparo à Pesquisa do Estado de São Paulo for this financial support in the context of the Projects No. 2021/09839-0, No. 2023/16273-8, and No. 2024/05700-5—"Dinâmica Não Linear." C.M. was supported by Agencia Estatal de Investigación (PID2024-160573NB-I00), the European Office of Aerospace Research and Development (EOARD, Project No. FA8655-24-1-7022).

H.C.C.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. J.V.M.S.: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Visualization. B.R.R.B.: Conceptualization, Investigation, Writing – original draft, Visualization. C.M.: Conceptualization, Investigation, Writing – original draft, Visualization. S.R.L.: Conceptualization, Methodology, Resources, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. T.L.P.: Conceptualization, Methodology, Resources, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

DATA AVAILABILITY

The data that support the findings of this article are openly available [27,35,36,42], embargo periods may apply.

[1] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, Cambridge, 2002).

[2] G. Datseris and U. Parlitz, *Nonlinear Dynamics: A Concise Introduction Interlaced with Code*, Undergraduate Lecture Notes in Physics (Springer International Publishing, Cham, 2022).

[3] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 2004).

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

[5] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. **438**, 237 (2007).

[6] M. Barreiro, A. C. Marti, and C. Masoller, Inferring long memory processes in the climate network via ordinal pattern analysis, Chaos **21**, 013101 (2011).

[7] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, Recurrence plots of dynamical systems, Europhys. Lett. **4**, 973 (1987).

[8] H. Poincaré, *Sur le probleme des trois corps et les equations de la dynamique*, *Acta Mathematica* (F. & G. Beijer, Stockholm, 1890).

[9] G. Corso, T. d. L. Prado, G. Z. d. S. Lima, J. Kurths, and S. R. Lopes, Quantifying entropy using recurrence matrix microstates, Chaos **28**, 083108 (2018).

[10] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, Cambridge, 2019).

[11] T. L. Prado, B. R. R. Boaretto, G. Corso, G. Z. dos Santos Lima, J. Kurths, and S. R. Lopes, A direct method to detect deterministic and stochastic properties of data, New J. Phys. **24**, 033027 (2022).

[12] G. S. Spezzatto, J. V. V. Flauzino, G. Corso, B. R. R. Boaretto, E. E. N. Macau, T. L. Prado, and S. R. Lopes, Recurrence microstates for machine learning classification, Chaos **34**, 073140 (2024).

[13] P. M. Addo, M. Billio, and D. Guegan, Nonlinear dynamics and recurrence plots for detecting financial crisis, North Am. J. Econ. Finance **26**, 416 (2013).

[14] O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, Distinguishing noise from chaos, Phys. Rev. Lett. **99**, 154102 (2007).

[15] J. M. Amigó, S. Zambrano, and M. A. F. Sanjuán, Permutation complexity of spatiotemporal dynamics, Europhys. Lett. **90**, 10007 (2010).

[16] B. R. R. Boaretto, R. C. Budzinski, K. L. Rossi, T. L. Prado, S. R. Lopes, and C. Masoller, Discriminating chaotic and stochastic time series using permutation entropy and artificial neural networks, Sci. Rep. **11**, 15789 (2021).

[17] T. L. Prado, G. Corso, G. Z. dos Santos Lima, R. C. Budzinski, B. R. R. Boaretto, F. A. S. Ferrari, E. E. N. Macau, and S. R. Lopes, Maximum entropy principle in recurrence plot analysis on stochastic and chaotic systems, Chaos **30**, 043123 (2020).

[18] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, The backbone of the climate network, Europhys. Lett. **87**, 48007 (2009).

[19] N. Marwan, J. F. Donges, Y. Zou, R. V. Donner, and J. Kurths, Complex network approach for recurrence analysis of time series, Phys. Lett. A **373**, 4246 (2009).

[20] J. P. Zbilut, N. Thomasson, and C. L. Webber, Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals, Med. Eng. Phys. **24**, 53 (2002).

[21] C. Letellier, H. Rabarimanantsoa, L. Achour, A. Cuvelier, and J. F. Muir, Recurrence plots for dynamical analysis of noninvasive mechanical ventilation, Philos. Trans. R. Soc. A **366**, 621 (2008).

[22] A. Fabretti and M. Ausloos, Recurrence plot and recurrence quantification analysis techniques for detecting a critical regime, Int. J. Mod. Phys. C **16**, 671 (2005).

[23] L. L. Trulla, A. Giuliani, J. P. Zbilut, and C. L. Webber, Recurrence quantification analysis of the logistic equation with transients, Phys. Lett. A **223**, 255 (1996).

[24] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths, Recurrence-plot-based measures of complexity and their application to heart-rate-variability data, Phys. Rev. E **66**, 026702 (2002).

[25] N. Marwan and K. H. Kraemer, Trends in recurrence analysis of dynamical systems, Eur. Phys. J.: Spec. Top. **232**, 5 (2023).

[26] T. L. Prado, V. S. Machado, G. Corso, G. Z. Dos Santos Lima, and S. R. Lopes, How to compute suitable vicinity parameter and sampling time of recurrence analysis, Nonlinear Dyn. **112**, 1141 (2024).

[27] A. L. Goldberger, L. A. N. 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 (2000).

[28] G. Datseris, Dynamicalsystems.jl: A julia software library for chaos and nonlinear dynamics, J. Open Source Software **3**, 598 (2018).

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

[30] M. Hénon, A two-dimensional mapping with a strange attractor, Commun. Math. Phys. **50**, 69 (1976).

[31] K. Ikeda, Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system, Opt. Commun. **30**, 257 (1979).

[32] E. N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. **20**, 130 (1963).

[33] O. E. Rössler, An equation for continuous chaos, Phys. Lett. A **57**, 397 (1976).

[34] A. M. Fraser and H. L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A **33**, 1134 (1986).

[35] J. Zheng, H. Guo, and H. Chu, A large scale 12-lead electrocardiogram database for arrhythmia study, PhysioNet (2022), doi:10.13026/wgex-er52.

[36] J. Zheng, H. Chu, D. Struppa, J. Zhang, S. M. Yacoub, H. El-Askary, A. Chang, L. Ehwerhemuepha, I. Abudayyeh, A. S. Barrett, G. Fu, H. Yao, D. Li, H. Guo, and C. Rakovski, Optimal multi-stage arrhythmia classification approach, Sci. Rep. **10**, 2898 (2020).

[37] B. R. R. Boaretto, A. C. Andreani, S. R. Lopes, T. L. Prado, and E. E. N. Macau, The use of entropy of recurrence microstates and artificial intelligence to detect cardiac arrhythmia in ECG records, Appl. Math. Comput. **475**, 128738 (2024).

[38] O. Yildirim, M. Talo, E. J. Ciaccio, R. S. San Tan, and U. R. Acharya, Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ECG records, Comput. Methods Programs Biomed. **197**, 105740 (2020).

[39] Z. Li, D. Zhou, L. Wan, J. Li, and W. Mou, Heartbeat classification using deep residual convolutional neural network from 2-lead electrocardiogram, J. Electrocardiol. **58**, 105 (2020).

[40] U. R. Acharya, O. Faust, V. Sree, G. Swapna, R. J. Martis, N. A. Kadri, and J. S. Suri, Linear and nonlinear analysis of normal and cad-affected heart rate signals, Comput. Methods Programs Biomed. **113**, 55 (2014).

[41] R. J. Martis, U. R. Acharya, H. Prasad, C. K. Chua, C. M. Lim, and J. S. Suri, Application of higher order statistics for atrial arrhythmia classification, Biomed. Signal Process. Control **8**, 888 (2013).

[42] https://github.com/JorgeMalosti/RM-SD-Data.