



Vibration signal analysis using symbolic dynamics for gearbox fault diagnosis

Ruben Medina¹ · Jean-Carlo Macancela² · Pablo Lucero² · Diego Cabrera² · Mariela Cerrada² · René-Vinicio Sánchez² · Rafael E. Vásquez³

Received: 18 January 2019 / Accepted: 7 May 2019 / Published online: 27 July 2019
© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract

This paper addresses the use of two algorithms based on symbolic dynamics analysis of vibration signal for fault diagnosis in gearboxes. The symbolic dynamics algorithm (SDA) works by subdividing the phase space described by the Poincaré plot into several angular regions; then, a symbol is assigned to each region. The probability distributions generated by the set of symbols are considered as features for classification of faults in a gearbox. The peak symbolic dynamics algorithm (PSDA) is a variant that extracts a sequence of peaks from the vibration signals and then performs the phase-space subdivision and symbol coding. A gearbox vibration signal dataset is analyzed for classifying 10 types of faults. Fault classification is attained using a multi-class support vector machine. The highest accuracy attained using k -fold cross-validation is 100.0% for load L3 with SDA and 100% with load L2 with PSDA. The accuracy considering all signals in the gearbox dataset is 99.2% with SDA and 99.8% with PSDA. The algorithms proposed have the advantage of being simple, accurate, and fast, and they could be adapted for online condition monitoring.

Keywords Symbolic dynamics · Fault diagnosis · Rotating machines · Poincaré plots

1 Introduction

Gearboxes are mechanical components of rotating machines that enable torque transmission. The configuration of such components is complex and they comprise mainly gears, bearings, and shafts. Faults in gearboxes can occur due to several factors. In particular, strong plastic deformation, lack of lubrication, wear, corrosion, and additionally design errors represent another reason involved in fault development. In consequence, several research efforts have been performed with the objective of developing methods that can detect faults in rotating machines with high accuracy and reliability [7, 61]. Detection and diagnosis of faults in gearboxes are tasks usually performed based on acquisition and analysis of vibration signals [30]. The standard method performs fault classification using machine learning techniques [42, 45]. Accomplishing this task requires the extraction of useful features from the recorded signals using classical signal analysis methods either in time [25]

or frequency domains, and time-frequency representation algorithms [1, 28, 57].

Data-driven fault detection and classification are performed by using extracted features as inputs of high-performance machine learning techniques [36]. For instance, Li et al. [29] addressed the problem of fault classification in gearboxes using vibration measurements; they use a Gaussian-Bernoulli deep Boltzmann machine that performs fault classification in gearboxes and bearings using a statistical features set. Such features set is estimated from different representations of the vibration signal including the original time-based representation, their frequency representation, and even time-frequency representations obtained using the Fourier transform or the wavelet transform. He et al. [16] reported an application concerning the diagnosis of faults in a gearbox transmission chain using an unsupervised approach known as deep belief networks, where optimization of structural parameters is attained using a genetic algorithm; the approach shows improvements with respect to other machine learning techniques for classifying faults in gearboxes and bearings. Cerrada et al. [8] developed a system that enables diagnosis of multi-class faults in spur gears; they considered a selected set of features extracted from vibration signal using

✉ Ruben Medina
ruben.djmedina@ieee.org

frequency and time based analysis. Ranking of such features using genetic algorithms enables faults classification with a random forest algorithm. Sanchez et al. [45] reported a methodological framework for multi-fault classification in rotating machinery using K -nearest neighbors and random forest; the ten most relevant features are selected using ranking techniques including chi-square, the ReliefF algorithm, and the information gain algorithm. This method is useful for online implementation as the extraction of features is developed in time domain. Praveenkumar et al. [42] used support vector machines (SVM) to classify faults in an automobile gearbox; they extracted a set of statistical features from the times series representation of the vibration signal and used it for fault classification with SVM. More recently, Medina et al. [32] also applied SVM with a different approach; in this research, Poincaré plots are the source of features enabling the classification of faults in gearboxes with an error correcting output codes (ECOC)-based multi-class SVM [12, 56]. Such application was inspired by the work of Brennan et al. [4] and Piskorski and Guzik [40]. A generalization of the Poincaré plot is considered where the lag between samples took an arbitrary value; this had the advantage of considering a time delay embedding system useful for modeling chaotic time series [55].

The underlying assumption when using classical feature extraction methods is that the source of acquired signals corresponds to a linear system satisfying stationary and periodicity properties [27]. Unfortunately, vibration signals in rotating machinery include sub-harmonics, chaotic features, and super-harmonics [9] due to the non-linear properties of the mechanical system. Nonetheless, some works in literature do use non-linear models; for instance, in [35], the authors reported the non-linear dynamic model for three bevel gears.

In rotating machinery where non-linear models are the most accurate, the non-linear dynamics methods and chaos theory are tools useful for analysis of vibration signals [48, 59]. Previous research within this context has reported investigations incorporating chaos and fractal approaches aimed at the analysis of the recorded signals. As an example, Sun [51] studied the evolution of faults in non-linear rotatory mechanical systems by using the largest Lyapunov exponent (LLE). Janjarasjitt et al. [20] reported the utilization of the correlation dimension (CD) to show the statistical differences of a brand new roller element with respect to another roller element close to failure based on the vibration signal analysis. More recently, Soleimani and Khadem [48] addressed the use of phase space for the analysis of gearboxes and ball bearings status; additionally, authors also proposed to use several features aimed at damage detection. They used features such as the largest Lyapunov exponent, the correlation dimension, and the approximate entropy.

A non-linear dynamical system can be analyzed by studying symbolic dynamics attained by phase-space topological partitions [5, 14]. In particular, symbolic signal processing is a method useful for transforming the recorded signals into sequences of symbols to enhance the features, discarding unnecessary details. Gupta and Ray [15] applied symbolic dynamics techniques for anomalies detection; they proposed a wavelet-based partitioning approach of the phase-space for extracting symbol sequences from the recorded time series. They performed the wavelet analysis after a careful selection of the necessary parameters. The resultant wavelet coefficients are represented as a one-dimensional signal at selected time and shift position. This scale series is used for generating the symbol sequence using a maximum entropy partition approach. The method is known as symbolic dynamic filtering (SDF) and one of the applications is early detection and prognosis of anomalies in several systems. Rao et al. [43] compared SDF with other methods for anomalies detection; in this research, the analytic signal is processed for extracting the symbol sequence and the scheme for space partitioning is extended to bi-dimensional data in order to generate symbolic representations of images. Space partitioning of the analytic signal relies on the application of the Hilbert transform. A D-Markov machine, operating as a finite state automata, is used for representing statistics of the symbol sequence. The 2D methodology for pattern detection using SDF is also applied to 2D wavelet coefficient datasets for anomaly detection. Further details of the approach using SDF are reported in [49]. An additional application of SDF was reported by Subbu et al. [50], where the evolution of fatigue damage is detected and quantified by using 2D SDF on images representing the surface of test specimens made of a polycrystalline alloy; they compared these results with respect to a similar analysis performed on 1D ultrasound signals simultaneously recorded from the same specimen, showing close agreement. Akintayo and Sarkar [2] reported a modification of the SDF method, corresponding to an algorithm for hierarchical symbolic dynamics filtering (HSDF) that is useful for modeling time series that are non-stationary.

Applications of symbolic dynamics have been reported for different research domains, as addressed by Daw et al. [11] which includes works in fields such as astrophysics, chemistry, geophysics, mechanical systems, biology, and telecommunications. For instance, Sanjith et al. [46] addressed fault detection of bearings using vibration signals; the time series vibration signal is converted into a symbolic sequence, and then, by using the generated symbolic data, a dictionary is constructed. The vibration signal for the no-load case is used as a reference for detecting the bearing in healthy or faulty condition. The comparison with respect to the reference

data is estimated using the Common Signal Index. The classification accuracy varies from 91.5 to 97.1%. de Paula et al. [39] used symbolic dynamics for flow patterns identification in a bank of tubes; the symbolic dynamics of the turbulent cross flow over tube banks can be useful for identifying patterns contained within the data. The symbolic dynamics approach was performed on the experimental time series, by using histograms according to a binary alphabet. The turbulent cross flow over a tube bank of five rows of circular cylinders placed in an aerodynamic channel was studied for validating the proposed approach. Considering more complex systems, Xu and Beck [58] performed an investigation of share price returns using a simple symbolic description. The approach considers two and four symbols for coding the signals. The sequence of symbols is analyzed using the spectrum of Rényi entropies. The share price returns showed a non-Markovian evolution in time.

An important number of applications of symbolic analysis have been reported in the neuro-science and cardiovascular physiology literature [41]. Kurths et al. [26] used symbolic dynamics to study the variability of heart rate intervals extracted from electrocardiographic signals; authors used the symbolic dynamics for quantifying the possibility of suffering ventricular arrhythmia and eventually sudden cardiac death. The symbolic description uses four symbols for representing the amplitude of beat-to-beat intervals. The symbols series are analyzed using generalized Rényi entropy [44]. Valencia et al. [54] studied whether symbolic transformations of time series extracted from electrocardiographic signals (RR and QT time series) could be useful for separating ischemic dilated cardiomyopathy (IDC) patients with respect to control (HC) subjects. The method attained an accuracy of 80% for separating both compared groups. Kim et al. [23] used the subdivision of the Poincaré plot in the context of electrocardiogram signal processing aimed at detection of cardiac abnormalities; in this application of symbolic dynamics, the goal is the detection of neurocardiogenic syncope after the analysis of heart rate signals extracted from electrocardiograms acquired during the head-up tilt test. The Poincaré plot is subdivided using lines parallel to the identity line $y = x$. The subdivision considers six regions and each region is assigned a symbol. The symbol sequences generated are analyzed using the approximated entropy showing their usefulness for abnormality detection.

In this work, a gearbox vibration signal dataset is acquired in a test rig under different conditions of load, motor speed, and faults related to the normal class and nine different fault configurations considering tooth breakage, wear, crack, and chafing. Some faults were configured in the pinion, other ones in the gear, and one both combined. Additionally, balanced data is considered, that is, the dataset has the same amount of samples for different loads and

motor speed under normal or fault condition. Then, these data are analyzed with the goal of attaining the multi-class classification of such faults. Two methods for extracting features from vibration signals in gearboxes by using SDA and PDSA on the Poincaré plot are proposed. The extracted features are used for gearbox faults classification with a multi-class SVM [12, 56]. Both methods exploit the non-linear nature of the gearbox mechanical system [35] for representing their symbolic dynamics aimed at faults classification. The Poincaré surface of section [5, 14] is obtained by plotting the measured vibration signal in a phase space of lagged coordinates. We have considered a modification of the method reported by Kim et al. [23]; in particular, the symbol sequence is obtained by partitioning the phase space of lagged coordinates into a set of angular regions. Each of the regions is identified with a symbol extracted from the available alphabet. The probability distribution generated by data from each symbol is used for attaining the classification with the ECOC-SVM algorithm. The proposed methods are simple and highly accurate for classification of faults in gearboxes attaining classification accuracies higher than 99.0%.

2 Theoretical background

2.1 Poincaré plot

The Poincaré plot is formally the intersection of a periodic orbit of a continuous dynamical system, in the state space, with a subspace of lower dimensions [53]. Such Poincaré section is the projection of the data in the plane spanned by the vectors defined by $[x(t), x(t + \tau)]$ that allows to obtain the Poincaré map or plot of the dynamical system [9]. In vibration signal analysis, the Poincaré plot is obtained by plotting each sample versus the next. The Poincaré plot can be generalized by setting an arbitrary lag τ between samples. The Poincaré plot has been mainly used as a graphical tool for describing the mechanical dynamics of chaotic nature in gearboxes and bearings, either in normal operation or in faulty conditions [10, 37, 38]. However, in applications concerning the detection of heart diseases based on electrocardiographic signal analysis, the Poincaré plot has been used for extracting useful features for abnormalities detection [24].

2.2 Support vector machines

The usual approach for data-driven fault classification involves three stages: the first stage is devoted to feature extraction. The second stage is optional and it is related to features selection; the fact is that some features have low sensitivity to fault evolution, and in consequence, these

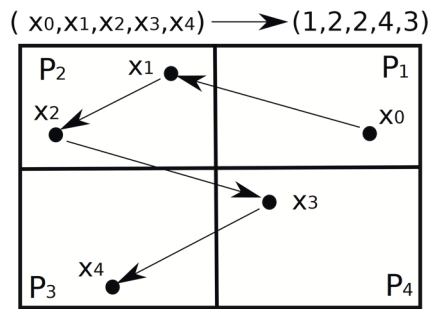


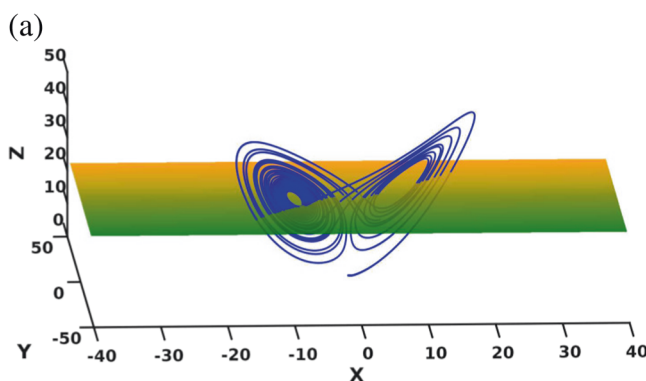
Fig. 1 Graphical representation of the phase space partitioning showing the obtained symbolic dynamics. Given the partitioning $\mathbf{p} = \{P_1, P_2, P_3, P_4\}$, the trajectory $(x_0, x_1, x_2, x_3, x_4, \dots)$ is represented by the symbol sequence $(s_0, s_1, s_1, s_3, s_2, \dots) = (1, 2, 2, 4, 3, \dots)$

features could become irrelevant or redundant for fault detection and diagnosis and features selection could be a necessary stage [34]. The third stage is fault classification using machine learning algorithms such as SVM [21, 60] and artificial neural networks (ANN) [36].

SVMs are machine learning algorithms useful for binary classification where an optimal hyper-plane between classes is estimated [31]. Multi-class versions of SVMs are also available to construct the decision functions simultaneously for all classes [56]. An alternative solution consists in subdividing the problem of multi-class classification in order to solve instead several simple binary classifications. Combination of such binary classifiers using the ECOC approach [12] enables multi-class classification. Discrimination of one class against the rest corresponds to the one-versus-all ECOC strategy. However, other strategies are also available. In particular, the one-versus-one method improves the performance of classification in problems with several classes.

2.3 Symbolic dynamics

Reconstruction of chaotic systems from noisy signals using a symbolic approach is a feasible method as shown in [52].



According to this theory, a dynamical system denoted $f : Q \rightarrow Q$ with time evolution (x_0, x_1, \dots) can be described by a sequence of symbols. This type of dynamical description, requires introducing a partition $\mathbf{p} = \{P_1, P_2, \dots, P_r\}$ subdividing the state space Q into r disjoint sets. Each subset is represented by a symbol $s_i \in \{1, 2, \dots, r\} = \mathbf{S}$. An orbit in the state space is represented by the sequence of visited symbols $\mathbf{s} = \{s_0, s_1, \dots\}$; this is shown in Fig. 1. The symbol sequence is considered as a transform for the input data that is able to retain important temporal information [11]. The symbolic dynamics is related to Poincaré maps where a multi-dimensional phase-space trajectory is intercepted with a plane to represent the simple Poincaré map [17]. The heuristical idea behind the symbolic representation is introducing a coarse re-sampling of the Poincaré map. Therefore, any interception of the phase-space trajectory located within a sub-region of the Poincaré map is known as “a cell” and it is assigned the same symbol; this is illustrated in Fig. 2 where in (a), the phase-space trajectories for the Lorenz attractor [13] intercept a plane for constructing the Poincaré plot. In (b), the Poincaré plot is shown. As the plane space can be arbitrarily discretized into several “cells,” the points of the Poincaré plot located in each “cell” can be represented by the same symbol.

Even when the symbolic dynamics has a well established theoretical basis, in practice, the generation of partitions of the Poincaré plot for attaining a trajectory in phase space associated with a unique symbol sequence is only feasible for certain systems [52]. In consequence, practical applications of symbolization have been mainly heuristic and empirical.

3 Methodology

The test rig for signal acquisition is located at the Rotating Machinery Laboratory of Universidad Politécnica Salesiana

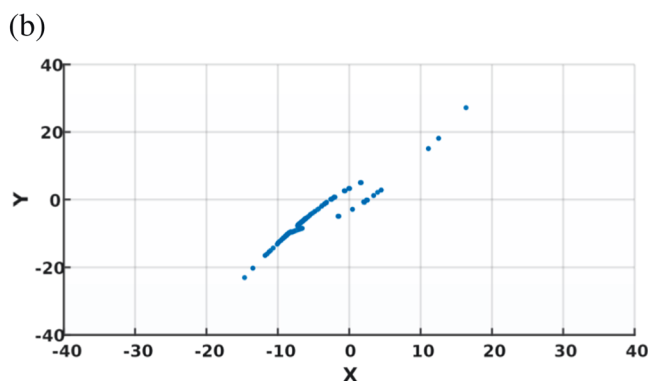
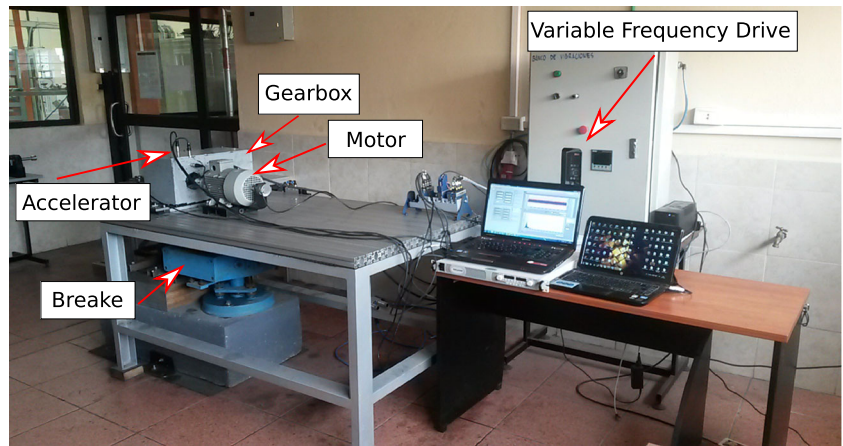


Fig. 2 **a** The phase-space trajectories for the Lorenz attractor are intersected with a plane for obtaining the Poincaré map. **b** The intersection points between the phase-space trajectories and the plane are represented by the discretized Poincaré plot

Fig. 3 Vibration analysis laboratory at Universidad Politécnica Salesiana, Cuenca, Ecuador



from Cuenca, Ecuador. Such a test rig is shown in Fig. 3. An induction motor model Siemens 1LA7 090-4YA60, 2 Hp for 1.1 kW is powered by three-phase electric line of 220 V at 60 Hz. This motor is used for providing rotational motion. The nominal speed of this motor has been set to 1650 rpm. The mechanical motion produced by the motor is transferred to one-stage gearbox where several gear faults are incorporated. A magnetic brake system connected to the output of the gearbox through a pulley is used for loading the motor. A diagram showing the components of the test rig used in this research is presented in Fig. 4.

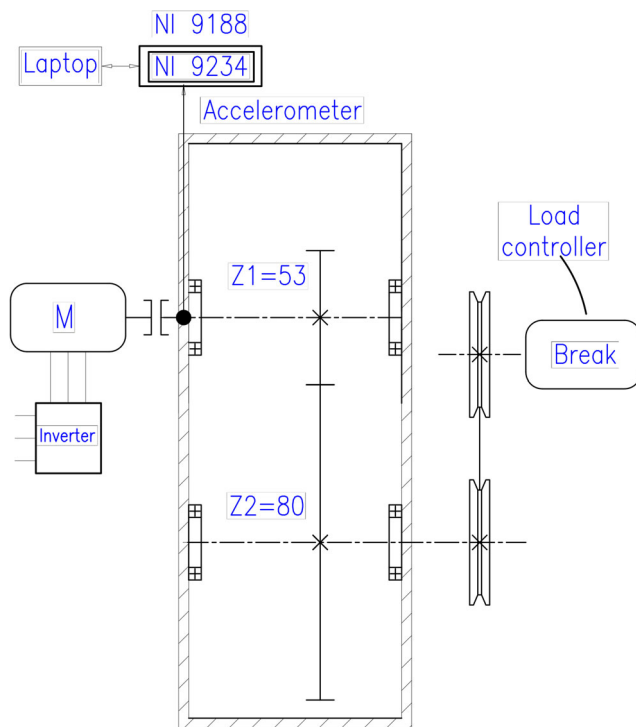


Fig. 4 Schematic of the experimental test bed for simulation of faults in rotating machinery

For this research, different break loads were implemented using the magnetic brake. Such loads were 0.83 Nm (L1), 4.19 Nm (L2), and 7.54 Nm (L3). Hence, a set of three variable speeds (720–1080 rpm, 300–720 rpm, and 480–900 rpm) and three different constant speeds (480, 720, and 900 rpm) were generated using a variable-frequency drive. The length of each recorded signal was acquired during 10 s at a sample frequency of 50 kSamples/second. The acquisition is performed with 24-bit resolution after processing the signals with an anti-alias filter. Ten different fault modes were configured in the gearbox components including the healthy condition as described in Table 1. Each fault mode included up to a pair of faulty elements. The configured faults for the gearbox elements are shown in Fig. 5. The healthy (normal) condition is denoted as P1. Incipient faults are denoted as P2 and P3. A fault is considered incipient when symptoms of malfunction are just starting to appear; the diagnosis of incipient faults is very important in industry for reducing costs and risks for human operators. Four moderate faults are denoted as P4, P5, P7, and P9. Two severe faults are also included: P6 and P8, and finally P10 is a multi-fault. Several combinations of test-rig operation parameters

Table 1 Gear fault conditions

Label	Pinion, Z1	Gear, Z2
P1	Normal	Normal
P2	Chafing tooth	Normal
P3	Tooth wear	Normal
P4	25% tooth breakage	Normal
P5	50% tooth breakage	Normal
P6	100% tooth breakage	Normal
P7	Normal	25% crack
P8	Normal	100% crack
P9	Normal	50% chafing
P10	25% tooth breakage	25% crack

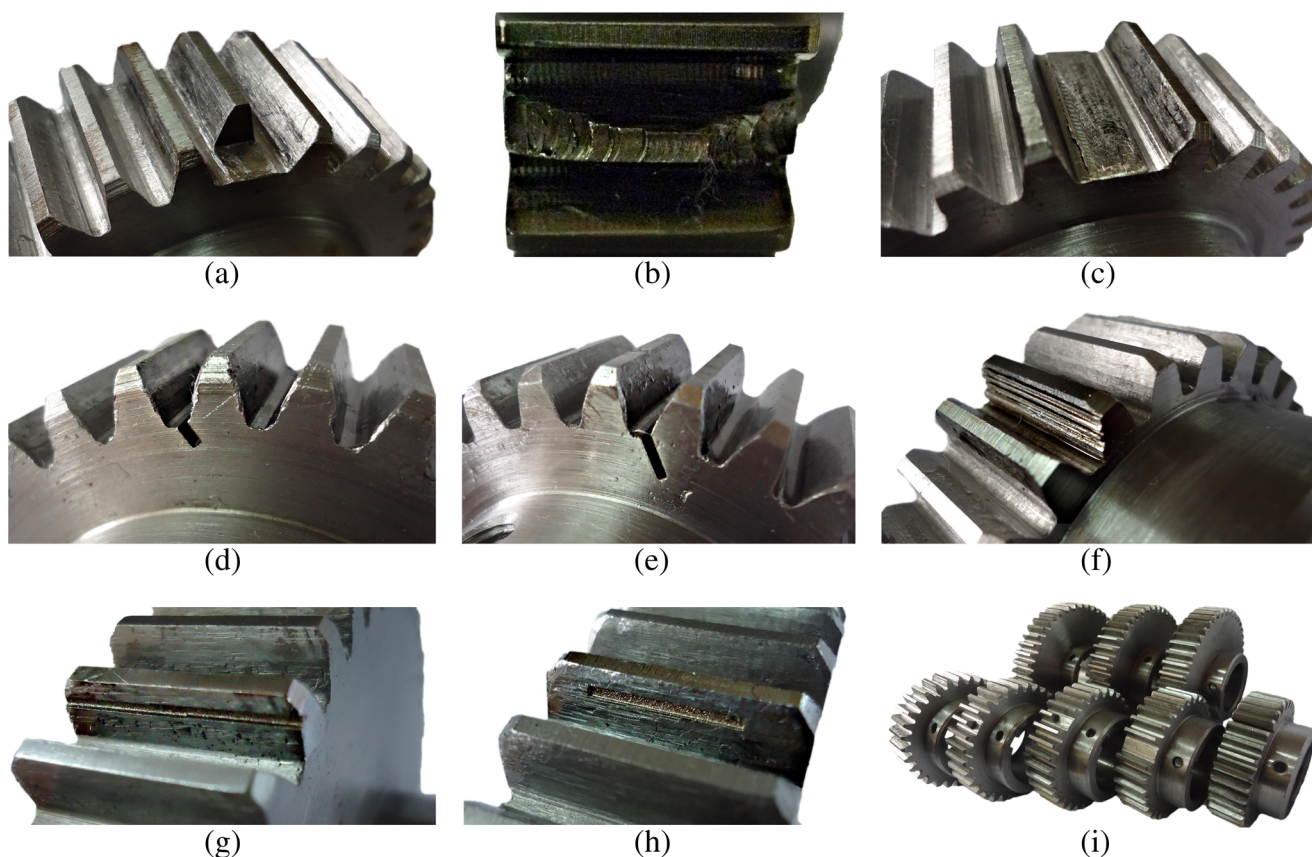


Fig. 5 Simulated faults **a** 25% tooth breakage ($P4$), **b** 50% tooth breakage ($P5$), **c** 100% tooth breakage ($P6$), **d** 25% gear crack ($P7$), **e** 100% gear crack ($P8$), **f** tooth wear ($P3$), **g** tooth chaffing ($P2$), **h** 50% tooth chaffing ($P9$), and **i** gear set with simulated faults

were considered for acquiring 900 vibration signals in total. Each fault condition includes 90 signals where each load is represented by 30 vibration signals.

3.1 Feature extraction

3.1.1 Phase map and lagged Poincaré plot

In non-linear dynamical systems, the intersection of the state space trajectories with the Poincaré section is known as the phase map [9]. In the case of the test rig described previously, the observation measurement is the vibration signal represented by time series $x(t)$ that measures the instantaneous displacement in the horizontal direction. A fragment of the vibration signal for a healthy gearbox ($P1$) is shown in Fig. 6a. The phase map is a plot of the time derivative $dx(t)/dt$ in function of $x(t)$ and the gearbox performs a quasi-periodical motion. In consequence, the phase map forms a closed curve as time evolves. Such a close curve is shown in Fig. 6b representing a fragment of signal shown in Fig. 6a. In Fig. 6c, a fragment of signal extracted from a faulty gearbox from the class $P10$ with faulty pinion Z1 and faulty gear Z2 is shown. In Fig. 6d, the

phase map for the signal representing the faulty class $P10$ is shown.

In several applications, the Poincaré plot is obtained by plotting the sample $x(t + 1)$ versus $x(t)$ [4, 22, 40]. In this research, a generalization of the Poincaré plot is considered; hence, rather than plotting two consecutive samples of $x(t)$, we are plotting the sample $x(t + \text{lag})$ as a function of $x(t)$, where $\text{lag} \geq 1$ is a lag in samples (known as lagged Poincaré plot [24]). In this application, the parameter lag enables a non-linear warp of the phase map. This is shown in Fig. 7a where the phase map shown in Fig. 6b appears rotated and deformed using a lag value $\text{lag} = 11$ samples. As we are interested only in the location of points within the Poincaré plot, the trajectory of points can be ignored and we obtain the Poincaré plot shown in Fig. 7b and d. The Poincaré plot is shown considering $\text{lag} = 11$. The Poincaré plot for the healthy class $P1$ is different with respect to the class $P10$. This suggests that the shape and location of points (Figs. 6 and 7 in the Poincaré plot) can be useful for fault detection. Selection of the lag parameter for optimal separation between fault classes has been previously reported [32]. The lag parameter should be greater than 10 samples and it can be chosen as the lag

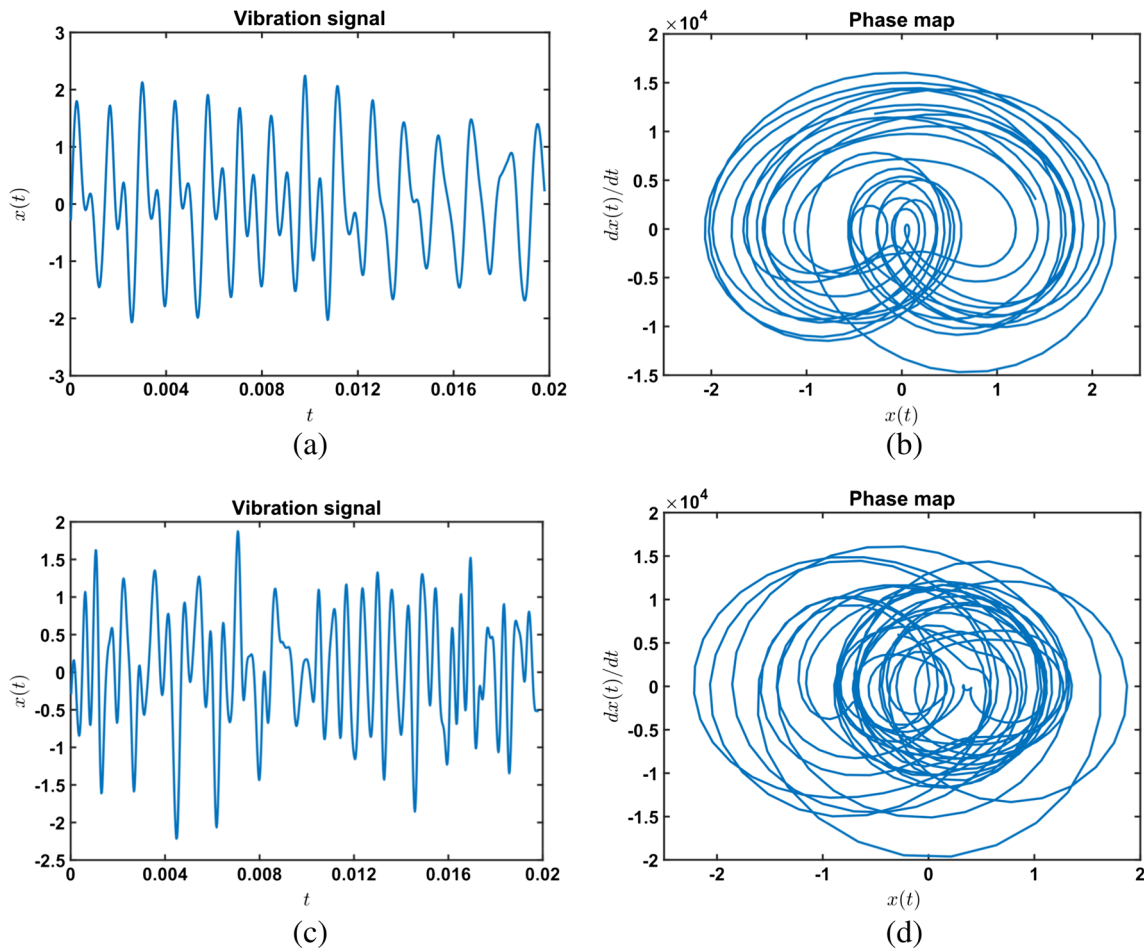


Fig. 6 Phase maps. **a** Vibration signal from a gearbox of the healthy class *P1*. **b** Phase map for the vibration signal from class *P1* obtained by plotting $dx(t)/dt$ versus $x(t)$. **c** Vibration signal from a gearbox of the faulty class *P10*. **d** Phase map for the vibration signal from class *P10*

representing the first zero-crossing for the auto-correlation sequence of $x(t)$. In particular, for this gearbox dataset, the optimal value for this parameter is $\text{lag} = 24$.

3.1.2 Peak detection

One of the stages of the peaks symbolic dynamics algorithm is the peak detection as shown in Fig. 8. The goal is to detect the peaks separated by at least a lag_{peak} samples. With this goal, the following steps are performed:

1. The absolute value for the vibration signal time series $\|x(t)\|$ is analyzed and the peaks satisfying the condition of being separated by at least lag_{peak} samples are retained. This stage generates a sequence of samples denoted as peaks signal $vp(t)$ that has a smaller size than the original signal. The sequences of peaks for signals of the healthy class *P1* and the faulty class *P10* are shown in Fig. 9a and c. The peak signal is shown overlaid on the original signal. A portion of 1000 samples of the original signal is shown. The peaks on

the original signal that are not selected do not satisfy the requirement of being separated from the previous peak by at least lag_{peak} samples. This stage enables the compression of signals by reducing its length to the size of a vector including only the peaks detected. The parameter lag_{peak} defines the minimum distance between peaks extracted from the vibration signal. This parameter can be set as the lag, where the first zero-crossing of the auto-correlation sequence is located. The selected value $\text{lag}_{\text{peak}} = 24$ is used for extracting the sequence of peaks.

2. Construction of the Poincaré plot using the time series $vp(t)$. The Poincaré plot is constructed with the set of points defined by each pair of consecutive peaks detected. The Poincaré plot for class *P1* is shown in Fig. 9b and the Poincaré plot for the faulty class *P10* is shown in Fig. 9d. The lag used for the Poincaré plots is $\text{lag} = 1$. The shape and distribution of points in the 2D space of the Poincaré plot are different for both classes.

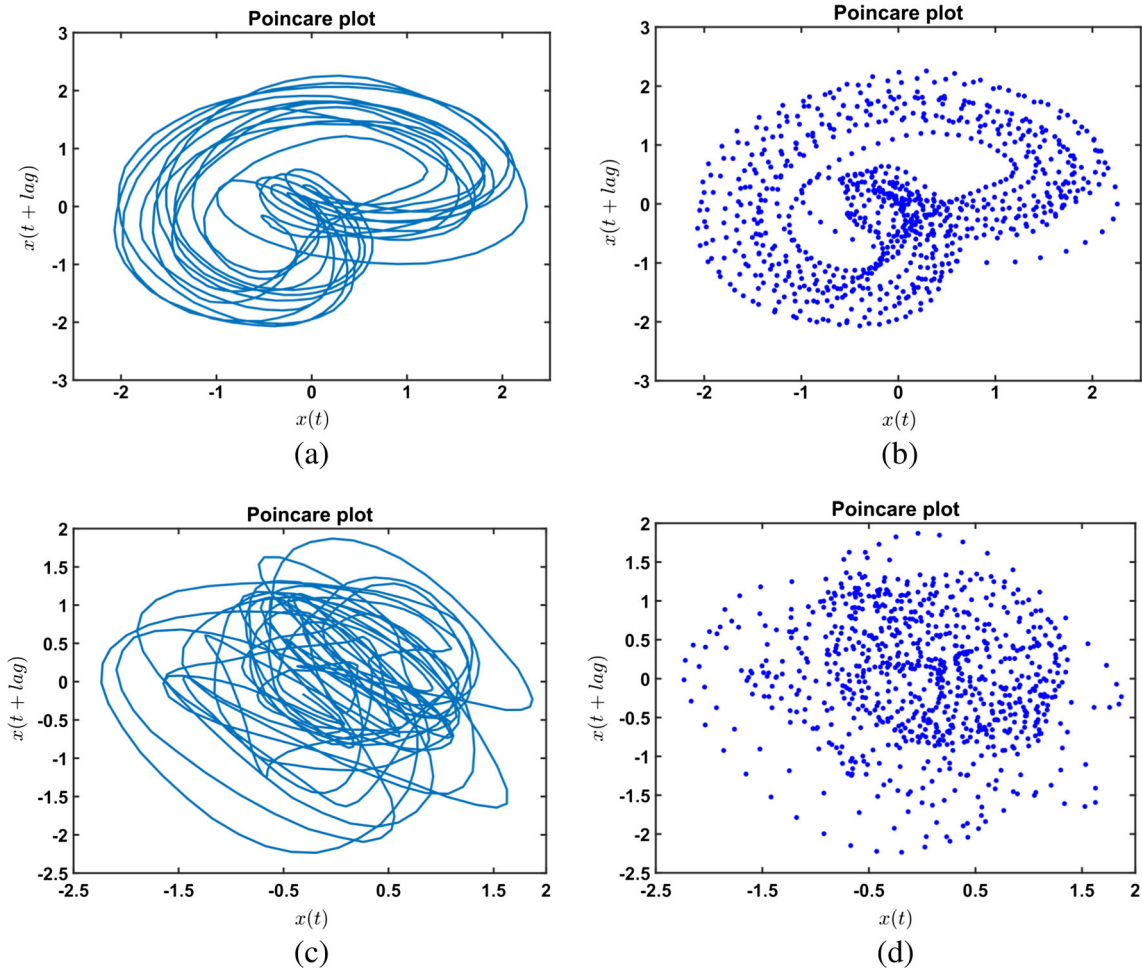


Fig. 7 Lagged Poincaré plots. **a** Poincaré plot obtained by plotting $x(t + \text{lag})$ versus $x(t)$ for a lag value set as lag = 11. The plot represents a vibration signal from a gearbox of the healthy class *P1*. The plot also shows the trajectories followed by points as time evolves. **b** Poincaré

plot showing only the location of points. **c** Lagged Poincaré plot for the vibration signal extracted from class *P10* showing the temporal trajectories of points. **d** Lagged Poincaré plot for signal from class *P10* showing only the location of points

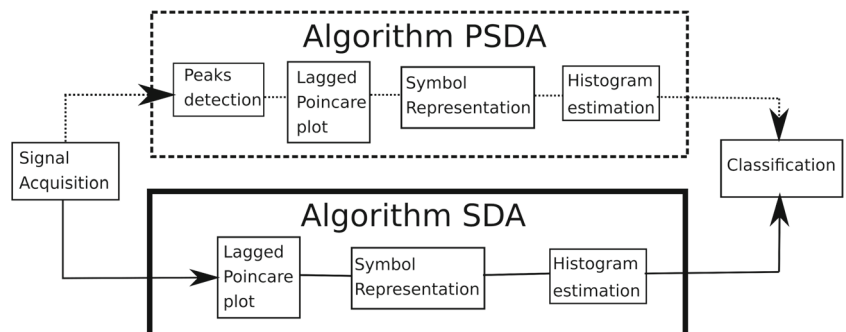
3.1.3 Symbolic dynamics algorithm

A simple method known as symbolic dynamics algorithm (SDA) for feature estimation is proposed (see Fig. 8). The algorithm includes the following steps:

1. The first step consists in plotting the Poincaré map using a lag value of 24 as shown in [32].

2. The phase-space spanned by the Poincaré plot is subdivided into 12 angular regions as shown in Fig. 10.
3. Each region is assigned a symbol represented by integer numbers {0, 1, 2, 3, ..., 11} as shown in Fig. 11.
4. The symbolic representation of the vibration signal is processed for obtaining the histogram. This is simply done by counting the number of points on each region.

Fig. 8 Two proposed algorithms are shown. The upper part shows the peaks symbolic dynamics algorithm. The lower part in solid line shows the symbolic dynamics algorithm. The PSDA includes an additional stage corresponding to the peaks detection



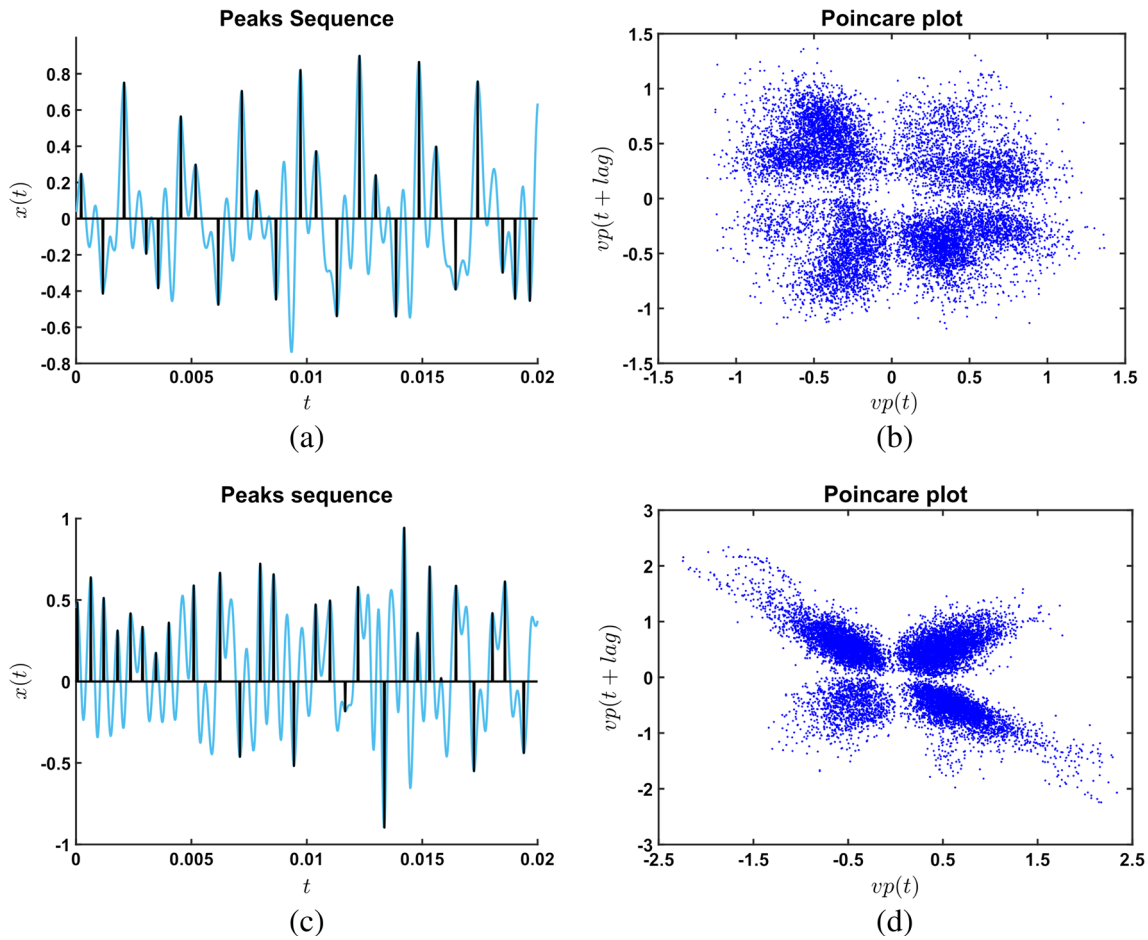


Fig. 9 Peaks extraction and Poincaré plots for the peaks sequences. **a** Peaks sequence overlaid on the original vibration signal extracted from healthy class *P1*. **b** Poincaré plot for the peaks sequence extracted from a signal of the class *P1*. The lag used for the Poincaré plot is

$\text{lag}_{2\text{map}} = 1$. **c** Peaks sequence overlaid on the original vibration signal extracted from faulty class *P10*. **d** Poincaré plot for the peaks sequence extracted from a signal of the class *P10*. The lag used for the Poincaré plot is $\text{lag}_{2\text{map}} = 1$

Fig. 10 Poincaré plot. The phase-space is subdivided into 12 angular regions

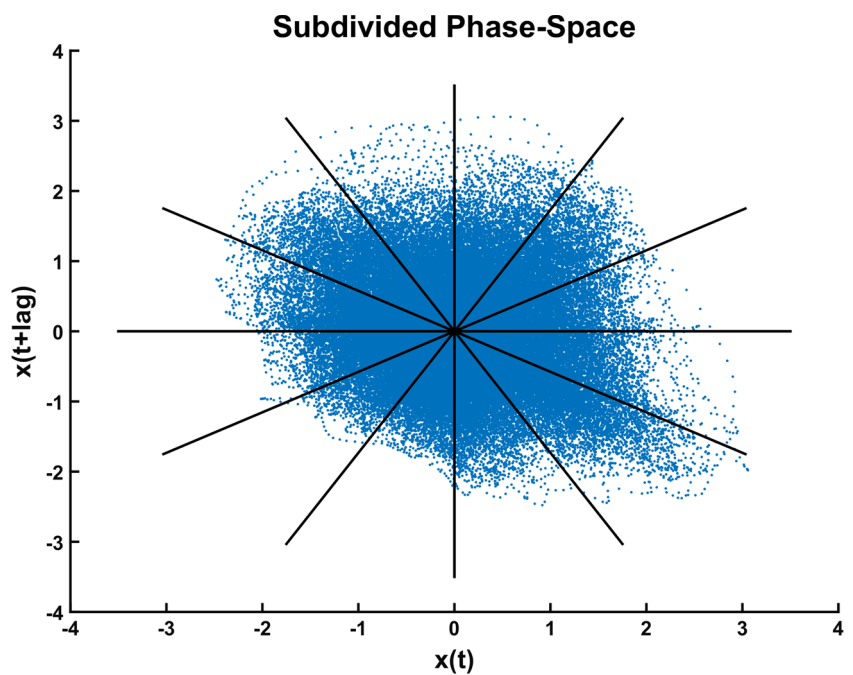
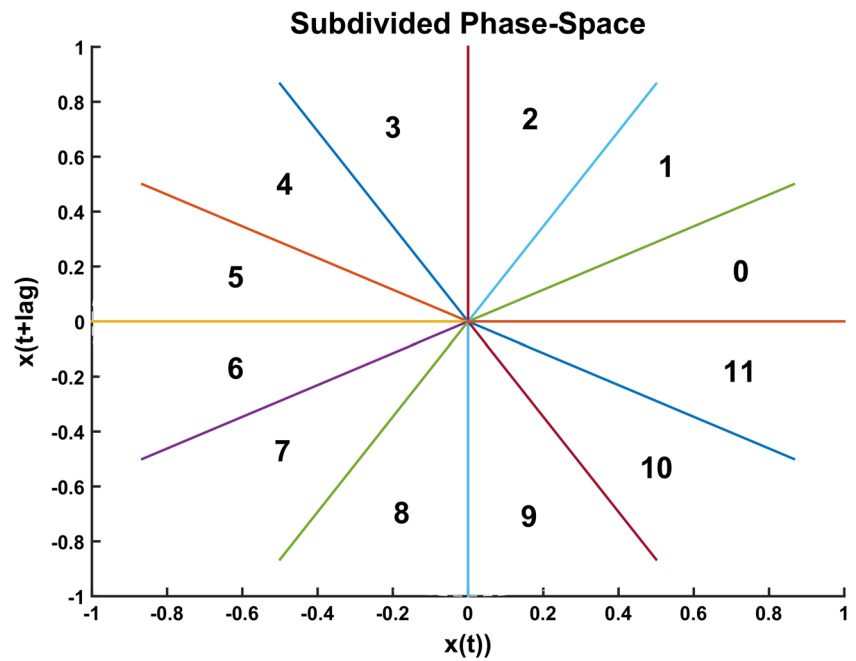


Fig. 11 Discretization of the phase-space. Each of the angular regions is assigned a symbol between 0 and 11



An example of the symbol histogram for a vibration signal is shown in Fig. 12.

5. All 12 elements in the histogram are considered as features for fault classification.

3.1.4 Peak symbolic dynamics algorithm

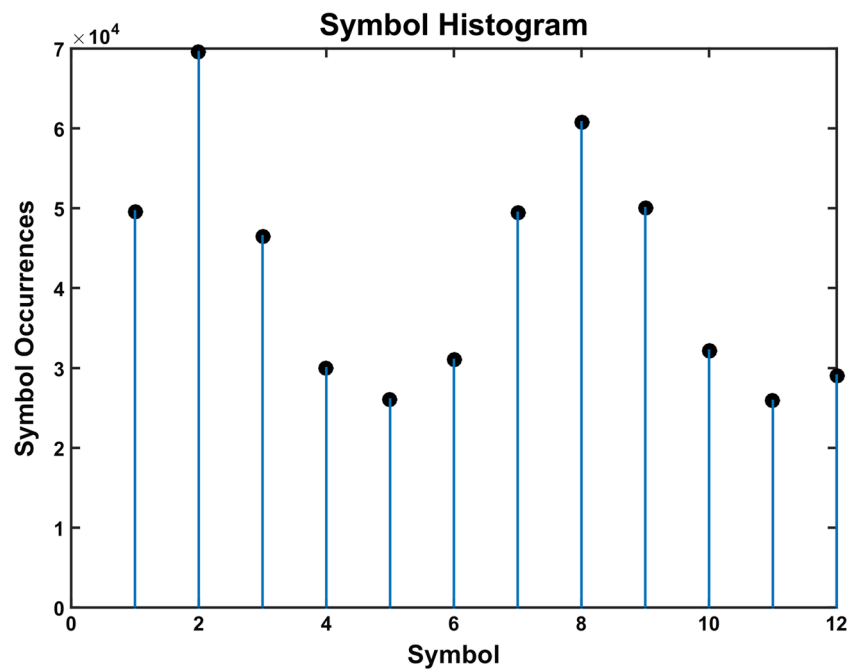
A variant of this algorithm is also implemented based on the Poincaré plot of the sequence of peaks extracted from the

vibration signal. The algorithm is known as peak symbolic dynamics algorithm (PSDA) (see Fig. 8). The rest of this algorithm includes the same steps of SDA.

3.2 Gearbox faults classification

The classification was performed in Matlab using multi-class ECOC-SVM. The features considered for each signal are the 12 elements in the histogram obtained using either

Fig. 12 Symbol histogram for a gearbox vibration signal



the SDA or the PSDA. Selection of SVM parameters was performed using an empirical method. First, the kernel type was selected as the one that provided higher classification accuracy considering the rest of parameters as default. The Gaussian or RBF kernel was selected and for this type of kernel, the scale parameter (σ) is necessary. For selecting this parameter, an interval centered at the default scale parameter calculated by the Matlab software was considered. The scale parameter was varied and the value retained provided the maximum classification accuracy within this interval. Parameter C also known as *Box Constraint* is a regularization coefficient that controls the maximum penalty imposed on margin-violating observations, and prevent overfitting. The default value for this parameter is 1.0 and several values were considered; however, there was not any improvement in classification accuracy with respect to the default value. The *Standardize* option was set to *true* such that the software function centers and scales each column of the predictor data.

The *k-fold* cross-validation procedure was used to accomplish the phases of training and testing of the SVM based model. The use of this procedure is well proved to obtain less biased models and, in consequence, more robust models regarding the generalization capability, than other methods such as a simple train/test split.

Two experiments for faults classification were performed:

1. Experiment 1: the first experiment was performed considering all vibration signals available in the dataset. During the first classification experiment, the group of 900 vibration signals was randomly partitioned into

10 equal sized subsets (10-fold) including 90 signals. From the 10 subsets, one is considered as the test-set for validating the trained model. The training set is constructed with the rest of subsets. This training and validation process is repeated 10 times performing the validation each time with a different subset. The multi-class SVM was trained using a Gaussian kernel with $\sigma = 0.87$.

2. Experiment 2: in the second experiment, the classification was performed considering the signals acquired at a constant load: $L1, L2,$ and $L3$. In this case, the dataset at constant load included 300 vibration signals, including 30 signals for each class. A 10-fold cross-validation was performed.

Evaluation of the classification ability of the trained classifier considered several metrics [18]. The discriminator metrics were based on the confusion matrix [47]; in particular, the accuracy or error rate is the evaluation metric commonly used in practice for evaluating the generalization capability of the classifier. In the case of binary classification, the correctness of a classification is evaluated by calculating the number of instances correctly classified (true positives (TP)), the number of instances that do not belong to the class and were correctly recognized (true negatives (TN)), and the instances that either were incorrectly assigned to the class (false positives (FP)) or were not recognized as instances belonging to the class (false negatives (FN)). These four calculated quantities are the elements of the confusion matrix for the binary classifier. In this research, however, we are dealing with multi-class classifiers. The confusion matrix for the multi-class

Table 2 Metrics calculated from the multi-class confusion matrix

Metrics	Formula	Metrics description
Accuracy (acc)	$acc_i = \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}$	Ratio of correct predictions over the total number of instances evaluated
Error rate (err)	$err_i = \frac{fp_i + fn_i}{tp_i + tn_i + fp_i + fn_i}$	Ratio of incorrect predictions over the total number of instances evaluated
Sensitivity (sn)	$sn_i = \frac{tp_i}{tp_i + fn_i}$	Fraction of positive instances that are classified correctly
Specificity (sp)	$sp_i = \frac{tn_i}{tn_i + fp_i}$	Fraction of negative instances that are classified correctly
False negative rate (fnr)	$fnr_i = \frac{fn_i}{fn_i + tp_i}$	Fraction of positive instances that are classified as negatives
False positive rate (fpr)	$fpr_i = \frac{fp_i}{fp_i + tn_i}$	Fraction of negative instances that are classified as positives
Precision (p)	$p_i = \frac{tp_i}{tp_i + fp_i}$	Ratio of positive instances correctly predicted over the total predicted instances of the positive class
F1-score ($F1 - s$)	$F1 - s = \frac{2 * p_M * sn_M}{p_M + sn_M}$	Average harmonic mean between precision and sensitivity

classifier is a generalization of the binary case [47]. In the case of M classes, each of them is denoted as C_i where $i \in \{1, 2, \dots, M\}$ and it has the associated true positive denoted tp_i , false positive fp_i , false negative fn_i , and true negative tn_i counts. Details of the calculation are presented in [47]. In particular, from the multi-class confusion matrix, several measures are calculated for each class. A set of calculated metrics is presented in Table 2. Calculation of the average metrics considering all classes can also be estimated as shown in [18, 47].

The receiver operator characteristic (ROC) curve has been used for evaluation of machine learning algorithms [3, 19]. In particular, the area under the ROC curve denoted AUC is the metric commonly used as a measure of classifier performance. The AUC has been shown by Huang and Ling [19] that is a better metric than accuracy. The ROC was also considered as a metric for analyzing the SVM performance as a classifier.

4 Results

4.1 Cluster structure for the symbolic dynamic algorithm

The analysis of the cluster structure, generated with the estimated features extracted directly from the symbolic representation obtained from the input signal, was developed using the SDA, considering all vibration signals and also for each of the recorded loads. The best results were obtained considering each of the loads recorded separately. In particular, the best results are obtained with loads $L2$ or $L3$. In this case, the load $L3$ was chosen for analyzing the cluster structure. Visualization of the cluster structure is attained by selecting only three features from the set of 12 features. Considering $n = 12$ and taking combinations of three elements provides 220 possible combinations; testing all these combinations is very time consuming. However, as the

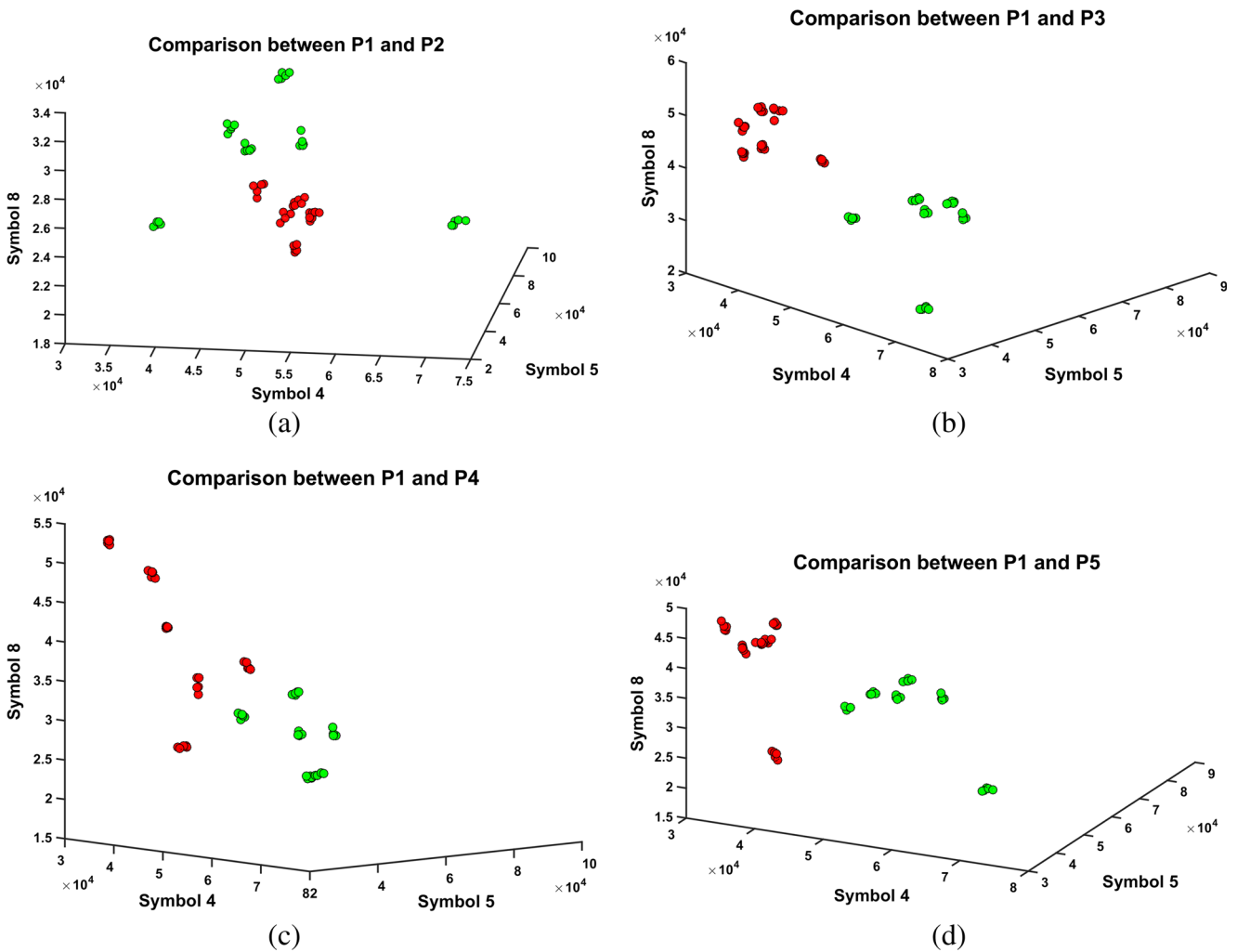


Fig. 13 Cluster structures comparison for load $L3$ of the gearbox dataset. Class $P1$ (the healthy class) is shown in light dots and faulty class is represented using dark dots. **a** Cluster structures for classes $P1$

and $P2$. **b** Cluster structures for classes $P1$ and $P3$. **c** Cluster structures for classes $P1$ and $P4$. **d** Cluster structures for classes $P1$ and $P5$

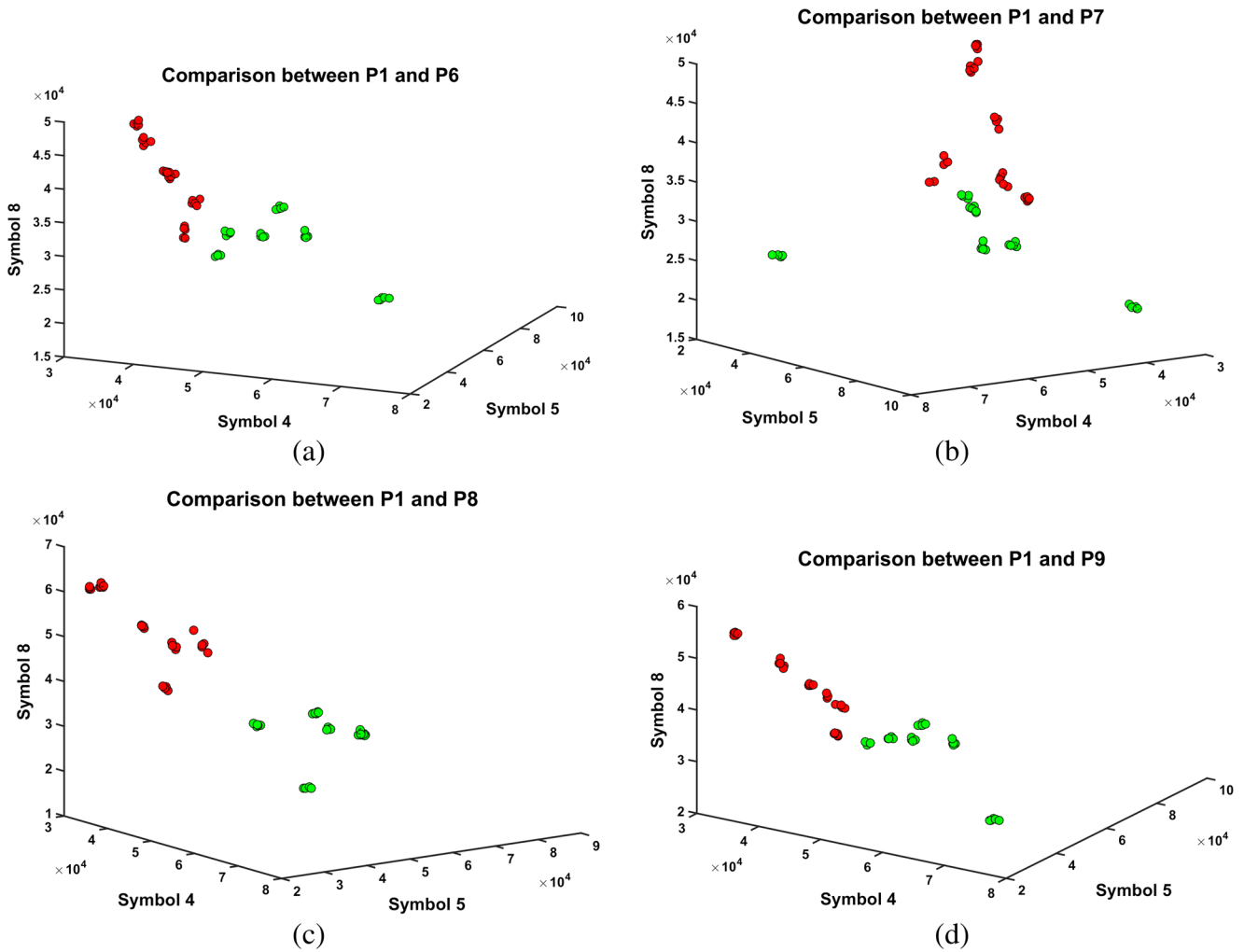


Fig. 14 Comparison of cluster structures for load *L3* of the gearbox dataset. The class *P1* (healthy class) is shown in light dots and the faulty class is shown in dark dots. **a** Cluster structures for classes *P1*

and *P6*. **b** Cluster structures for classes *P1* and *P7*. **c** Cluster structures for classes *P1* and *P8*. **d** Cluster structures for classes *P1* and *P9*

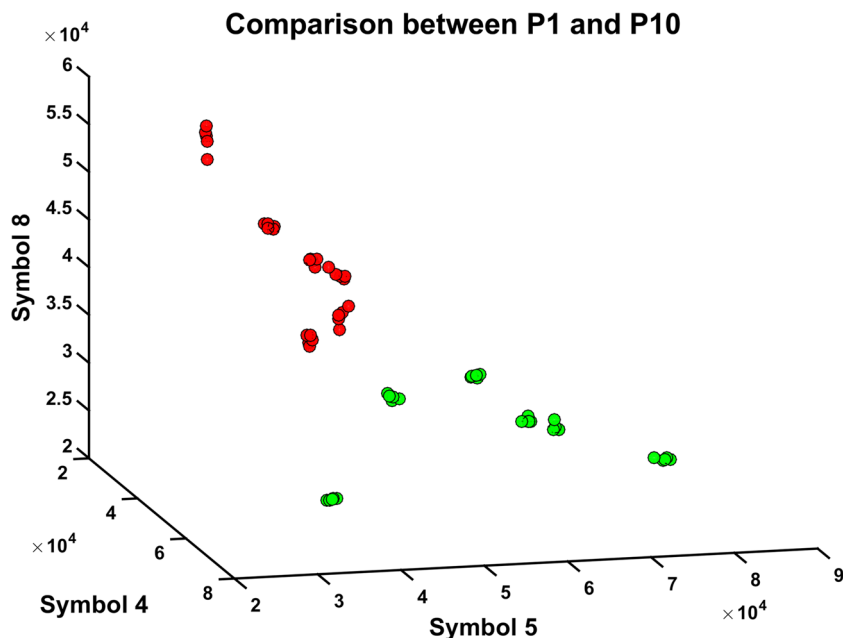
primary goal is to provide clues about the generated cluster structure, in this case, only a few combinations were tested for showing the cluster structure. Features 4, 5, and 8 were selected and the scatter plot was used for comparing the faulty classes *P2* to *P10* with respect to healthy class *P1*. The faulty classes are represented using dark dots while the class *P1* is represented using light dots. The comparison between class *P1* and *P2* is shown in Fig. 13a; the cluster structure shows that class *P2* can be separated from the normal class in the space generated for only three features. The separation could be attained with a single surface. Similarly, in Fig. 13b, the comparison between classes *P1* and *P3* is shown; clearly, the class *P3* can be separated from normal class *P1*. Comparison of class *P1* with respect to classes *P4* and *P5* is shown in Fig. 13c and d. The separation of clusters in the parameter space for both faulty classes is apparent. However, separation is better between class *P5* and the normal class *P1*.

In Fig. 14a and b, the comparison between class *P1* and classes *P6* and *P7* is shown. The normal class has a cluster well separated with respect to clusters for the faulty classes. Similarly, the comparison between the normal class *P1* and classes *P8* and *P9* is shown in Fig. 14c and d. Separation between the clusters *P8* and *P9* with respect to the normal cluster *P1* is evident. Finally, comparison of a combined fault corresponding to class *P10* with respect to normal class *P1* is shown in Fig. 15. A single surface could be useful for separating the faulty cluster *P10* with respect to the normal class *P1*.

4.2 Results of experiment 1 with SDA

The first experiment concerning the fault classification was performed considering all vibration signals. The average accuracy obtained during the cross-validation was 99.2% using a Gaussian kernel with $\sigma = 0.87$. The resultant

Fig. 15 Comparison between classes *P1* and *P10* using three features considering load *L3* of the gearbox dataset



confusion matrix for this experiment, considering all signals in the dataset, is presented in Fig. 16. Table 3 shows the metrics values measuring the classification performance associated to each class. The formulas for calculating these metrics were presented in Table 2. The column 1 represents the false negative rate (*fnr*); classes *P3*, *P4*, *P5*, *P6*, and *P9* attain the higher values for *fnr* that are, in all cases lower than 3.7%. The *fpr* metric is shown in column 2;

classes *P5*, *P6*, and *P9* attain the higher values for this metric. However, they are lower than 3.3%. The sensitivity (*sn*) is shown in the third column; class *P6* attained the lowest sensitivity with a value of 96.7%. The highest sensitivity is 100%. The specificity (*sp*) is also shown. Class *P5* attained the lowest specificity with a value of 99.6%.

The ROC curve is useful for visualizing classifiers performance where sensitivity is plotted against 1-specificity [3]. The ROC for the experiment using SDA with all vibration signals in the dataset is shown in Fig. 17. As we are dealing with a multi-class classification problem and the classification accuracy is very high and similar for different classes, the curves are overlaid. In this plot, the accuracy attained by class *P5* is 96.7%. The accuracy attained by classes *P3*, *P4*,

Confusion Matrix

Output Class	1	90 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	2	0 0.0%	90 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	3	0 0.0%	0 0.0%	89 9.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	4	0 0.0%	0 0.0%	0 0.0%	89 9.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	5	0 0.0%	0 0.0%	0 0.0%	1 0.1%	87 9.7%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	97.8% 2.2%
	6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.2%	89 9.9%	0 0.0%	0 0.0%	1 0.1%	0 0.0%	96.7% 3.3%
	7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	0 0.0%	100% 0.0%
	9	0 0.0%	0 0.0%	1 0.1%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	89 9.9%	0 0.0%	97.8% 2.2%
	10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	100% 0.0%
		100% 0.0%	100% 0.0%	98.9% 1.1%	98.9% 1.1%	96.7% 3.3%	98.9% 1.1%	100% 0.0%	100% 0.0%	98.9% 1.1%	100% 0.0%	99.2% 0.8%
	1	2	3	4	5	6	7	8	9	10		
	Target Class											

Fig. 16 Confusion matrix representing the results of faults classification using SDA considering all signals in the dataset

Table 3 Confusion matrix results considering all signals and SDA for feature extraction

Class	<i>fnr</i>	<i>fpr</i>	Sensitivity	Specificity
P1	0.000	0.000	1.000	1.000
P2	0.000	0.000	1.000	1.000
P3	0.012	0.000	1.000	0.999
P4	0.012	0.000	1.000	0.999
P5	0.037	0.022	0.978	0.996
P6	0.012	0.033	0.967	0.999
P7	0.000	0.000	1.000	1.000
P8	0.000	0.000	1.000	1.000
P9	0.012	0.022	0.978	0.999
P10	0.000	0.000	1.000	1.000

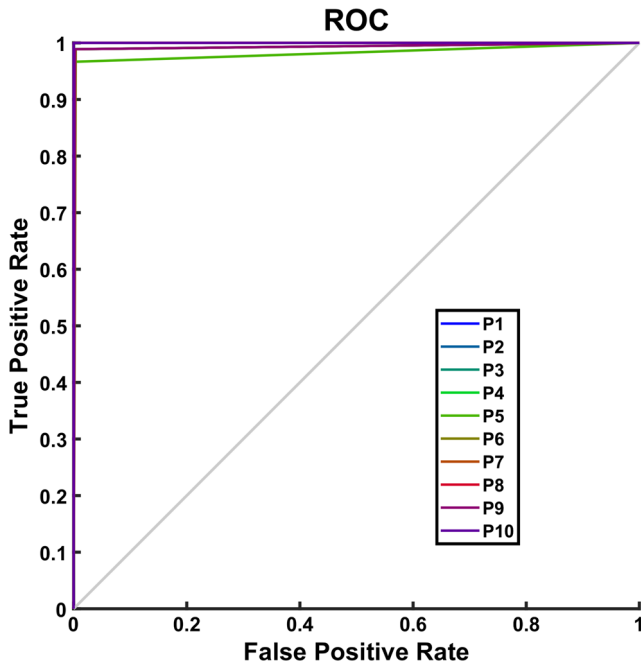


Fig. 17 ROC curve for SDA considering all signals in the dataset

P6, and *P9* is 98.9% and these classes are overlaid. The remaining classes *P1*, *P2*, *P5*, *P8*, and *P10* attain a perfect classification accuracy of 100% and they appear overlaid. In

Fig. 18 Confusion matrix representing the results of the experiment for faults classification using the symbol sequence extracted using PSDA considering all signals in the gearbox dataset

		Confusion Matrix										
		1	2	3	4	5	6	7	8	9	10	
Output Class	1	90 10.0%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	98.9% 1.1%
	2	0 0.0%	90 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.1%	98.9% 1.1%
	3	0 0.0%	0 0.0%	89 9.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	4	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	0 0.0%	100% 0.0%
	9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90 10.0%	0 0.0%	100% 0.0%
	10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	89 9.9%	100% 0.0%
		100% 0.0%	100% 0.0%	98.9% 1.1%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	98.9% 1.1%	99.8% 0.2%
		1	2	3	4	5	6	7	8	9	10	

Table 4 Confusion Matrix results considering all the signals and PSDA for feature extraction

Class	<i>fnr</i>	<i>fpr</i>	Sensitivity	Specificity
P1	0.000	0.011	0.989	1.000
P2	0.000	0.011	0.989	1.000
P3	0.012	0.000	1.000	0.999
P4	0.000	0.000	1.000	1.000
P5	0.000	0.000	1.000	1.000
P6	0.000	0.000	1.000	1.000
P7	0.000	0.000	1.000	1.000
P8	0.000	0.000	1.000	1.000
P9	0.000	0.000	1.000	1.000
P10	0.012	0.000	1.000	0.999

general, the accuracy attained for all the classes is high and the lower performance is attained by class *P5*.

4.3 Results of the experiment 2 with SDA

The experiment concerning the faults classification using SDA for vibration signals acquired at a particular load provided a classification accuracy for load *L1* of 97.7%, 100.0% for load *L2*, and finally 100.0% for load *L3*. The

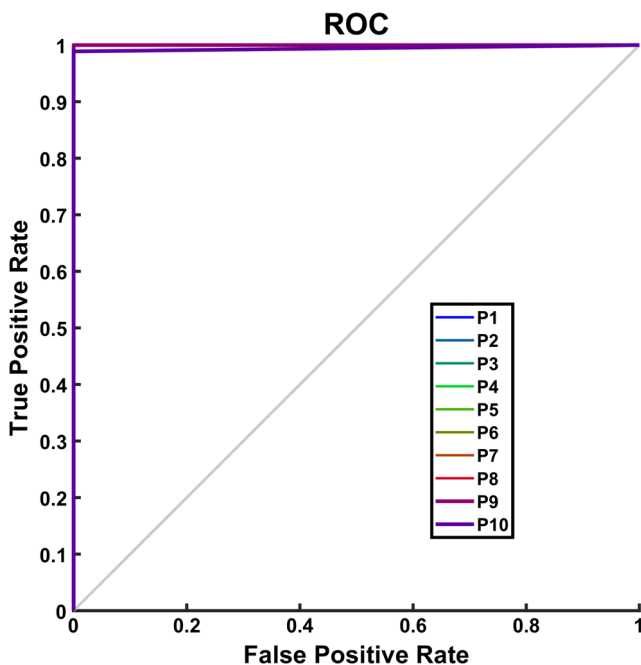


Fig. 19 ROC curve for gearbox faults classification using PSDA considering all signals in the gearbox dataset

best classification accuracy is attained for load $L2$ and $L3$. A list of quantitative metrics for classification with SDA is presented in Table 5; this algorithm shows excellent accuracy either in Experiment 1 or Experiment 2.

4.4 Results of experiment 1 with PSDA

The PSDA was used for fault classification considering all vibration signals in the dataset. The average accuracy obtained during the validation was 99.8% using a Gaussian kernel with a value of $\sigma = 0.87$. Figure 18 shows the resultant confusion matrix considering all signals in the dataset. The performance metrics associated to each class are shown in Table 4. The false negative rate (fnr) is presented in the first column. The classes $P3$ and $P10$ attained the higher values for this parameter which is lower than 1.2%. The false positive rate (fpr) is presented in the

second column. Classes $P1$ and $P2$ attained the higher value of 1.1%. The sensitivity (sn) or recall is presented in the third column. In this experiment, classes $P1$ and $P2$ attained the lowest sensitivity with a value of 98.9% while the rest of classes attained a value of 100%. The specificity (sp) is presented in the fourth column and classes $P3$ and $P10$ attained the lowest value of 99.9%.

The ROC for PSDA, considering all signals in the gearbox dataset, is shown in Fig. 19. In this multi-class classification problem with classification accuracies very high and similar for different classes, the curves are overlaid. Classes $P3$ and $P10$ attained the lower performance in terms of accuracy with a value of 98.9% and both curves are overlaid. The remaining curves for $P1$, $P2$, $P4$, $P5$, $P6$, $P7$, $P8$, and $P9$ attain a perfect classification accuracy of 100% and they are overlaid in this plot. However, in general, the PSDA is highly accurate for gearbox fault classification.

4.5 Results of experiment 2 with PSDA

An experiment was performed and the features used for classification were extracted with PSDA from the sequence of peaks that comes from vibration signal. The accuracy of classification using PSDA was calculated considering a particular load. A classification accuracy of 98.7% for load $L1$ is attained as a result for this experiment. The accuracy obtained with load $L2$ was 100.0%, while the accuracy for load $L3$ was 99.7%. Classification considering signals acquired at load $L2$ provided the higher classification accuracy. A list of quantitative metrics for classification results with PSDA is presented in Table 5 where it is shown that this algorithm is highly accurate.

4.6 Results comparison

4.6.1 Comparison of algorithms proposed

Both algorithms (SDA and PSDA) are highly accurate for fault classification in gearboxes. The classification accuracy is 99.78% for the tests performed using all vibration signals

Table 5 Gearbox classification accuracy for the proposed algorithms (in percents)

Algorithm	Load	Accuracy	Error rate	Specificity	Precision	F1-score	fpr
Algorithm SDA	All	99.22	0.78	99.91	99.23	99.22	0.08
	$L1$	98.33	1.67	99.81	98.43	98.33	0.19
	$L2$	100.00	0.00	100.00	100.00	100.00	0.00
	$L3$	100.00	0.00	100.00	100.00	100.00	0.00
Algorithm PSDA	All	99.78	0.22	99.98	99.78	99.78	0.02
	$L1$	98.33	1.67	99.81	98.45	98.34	0.19
	$L2$	100.00	0.00	100.00	100.00	100.00	0.00
	$L3$	99.67	0.33	99.96	99.68	99.67	0.04

Table 6 Comparison with respect to other methods using the same dataset

	Random	MDSVC	Poincaré	Sparse	Feature	Symbolic
	Forest [6]	[30]	SVM [32]	Representations [33]	Ranking [45]	Dynamics
Accuracy (%)	98.0	97.1	95.3	95.1	99.3	99.8

in the dataset. An accuracy of 100% can be obtained working at a particular load such as load $L2$ and $L3$. The results for the faults classification experiments are presented in Table 5. These results show that combination of the lagged Poincaré plot with symbolic dynamics is very efficient for extracting the features representing the mechanical status of the gearbox.

4.6.2 Comparison with respect to other methods using the same dataset

Comparison with other classification methods applied to the same gearbox dataset shows that accuracy results are higher considering both proposed algorithms. Results of the comparison are shown in Table 6. The maximum accuracy attained for each algorithm is presented. A method using the same gearbox dataset has been reported in [6]. A set of features is composed of time domain statistical features, as well as frequency domain and time-frequency domain features extracted from vibration signals. A hierarchical unsupervised algorithm is used for extracting representative features useful for fault classification. In particular, the classification was performed based on 12% of 811 features using several machine learning techniques, and the algorithm attained a classification accuracy of 98% using random forest classifiers. The same dataset was also analyzed in [30] using multimodal deep support vector classification (MDSVC). In this research, nine features were calculated from the time representation of the vibration signal; these features were also calculated from the frequency representation of the vibration signal. Similarly, a set of features was also extracted from the wavelet packet representation. A MDSVC composed by three Gaussian-Bernoulli deep Boltzmann machine (GDBM) was trained for classifying the faults of the gearbox dataset. The algorithm attained a classification rate of 97.08%. Another research with this dataset [32] performed the classification using only three features extracted from the Poincaré plot for feeding support vector machines. The highest classification accuracy was 95.3% considering vibration signals at constant load. Dictionary sparse-based representations of vibration signals were also used for extracting features for fault classification in this gearbox dataset [33]. In that case, the best classification accuracy was 95.1% and it was obtained at constant load. Further results using this gearbox dataset were reported

in [45]. A total of 30 features are extracted from the time-domain representation of the vibration signal. The extracted features are ranked using ReliefF, chi-square, and information gain methods. After applying feature ranking methods, a selection of ten features was considered for classification. The machine learning algorithms used were K -nearest neighbors (KNN) and random forests. The highest classification accuracy obtained was 99.3%. In contrast, the maximum percent classification accuracy attained by the proposed symbolic dynamics algorithms is 99.8%. Additionally, the proposed algorithm considers only 12 features that are extracted easily from the time-domain representation of the vibration signal.

5 Conclusions

The problem of fault classification in gearboxes is solved using two symbolic dynamics algorithms. The SDA extracts the symbols histogram directly from the Poincaré plot constructed considering the time samples from the vibration signal, and the PSDA extracts the symbols histogram from a Poincaré plot constructed with a peaks sequence obtained from the vibration signal. The feature analysis showed a cluster structure that could be useful for classification of faults in gearboxes.

The result concerning the total accuracy was 99.2% considering features extracted directly from the vibration signal. Alternatively, when the symbol histogram is extracted from the sequence of peaks from the vibration signal, the accuracy obtained was 99.8%. In both cases, all the signals in the dataset are considered.

Performing the classification of faults using only signals acquired at a particular load could reduce slightly the accuracy of classification for loads $L1$ using both algorithms. However, for load $L2$, the accuracy of classification is increased to 100%.

The analysis of the generated cluster structures shows that even considering a reduction in the number of features to only three elements of the histogram produces configuration of instances in the feature space that could be separated using the appropriate machine learning technique.

The experiments also showed that high accuracy is attained when the classification considers features from vibration signals acquired at a constant load. The proposed feature extraction approach has the advantage

of being simple because only 12 features are extracted from the vibration signal. Moreover, their calculation requires a low computational cost. Additionally, the fault classification algorithms proposed are more accurate than other previously reported methods using the same dataset. This approach is also general as it could be useful for extracting features for classification of faults, considering other types of signals recorded from gearboxes or roller bearings.

The proposed method is fast and can be adapted for real-time condition monitoring. A trained SVM could be used for classifying with low delay, a recorded portion of a vibration signal as only 12 features are necessary and their calculation and classification require a low computational cost. No other variant of the SVM algorithm was considered as the obtained performance was adequate. Future works could address the comparison between different SVM algorithms under the particular context of the available number of features and samples of this case study.

The gearbox vibration signals dataset used in this research includes a number of signals that enables the machine learning technique to solve a multi-classification problem that is balanced, since the number of instances in each of the classes is similar. Further research is still necessary for evaluating the efficacy of the feature extraction algorithms in situations where there is no balance among the considered classes. Furthermore, even when the test rig considers classes which comprise combined faults in the pinion and gear, there are more complex combination of faults that would be necessary to study using the proposed algorithms.

References

- Aherwar A (2012) An investigation on gearbox fault detection using vibration analysis techniques: a review. *Aust J Mech Eng* 10(2):169–183
- Akintayo A, Sarkar S (2018) Hierarchical symbolic dynamic filtering of streaming non-stationary time series data. *Signal Process* 151:76–88
- Bradley AP (1997) The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn* 30(7):1145–1159
- Brennan M, Palaniswami M, Kamen P (2001) Do existing measures of poicare plot geometry reflect nonlinear features of heart rate variability? *IEEE Trans Biomed Eng* 48(11):1342–1347
- Cafaro C, Lord WM, Sun J, Bollt EM (2015) Causation entropy from symbolic representations of dynamical systems. *Chaos: Interdiscip J Nonlinear Sci* 25(4):043106
- Cerrada M, Sánchez RV, Pacheco F, Cabrera D, Zurita G, Li C (2016) Hierarchical feature selection based on relative dependency for gear fault diagnosis. *Appl Intell* 44(3):687–703
- Cerrada M, Sánchez RV, Li C, Pacheco F, Cabrera D, de Oliveira JV, Vásquez RE (2018) A review on data-driven fault severity assessment in rolling bearings. *Mechan Syst Signal Process* 99:169–196
- Cerrada M, Zurita G, Cabrera D, Sánchez RV, Artés M, Li C (2016) Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mech Syst Signal Process* 70:87–103
- Chang-Jian CW (2010) Strong nonlinearity analysis for gear-bearing system under nonlinear suspension—bifurcation and chaos. *Nonlinear Anal: Real World Appl* 11(3):1760–1774
- Chen G (2009) Study on nonlinear dynamic response of an unbalanced rotor supported on ball bearing. *J Vibr Acoust* 131(6):061001
- Daw CS, Finney CEA, Tracy E (2003) A review of symbolic analysis of experimental data. *Rev Sci Instrum* 74(2):915–930
- Escalera S, Pujol O, Radeva P (2010) On the decoding process in ternary error-correcting output codes. *IEEE Trans Pattern Anal Mach Intell* 32(1):120–134
- Ghys É (2013) *The Lorenz Attractor, a Paradigm for Chaos*. Springer, Basel, pp 1–54
- Grebogi C, Ott E, Yorke JA (1987) Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics. *Science* 238(4827):632–638
- Gupta S, Ray A (2007) Symbolic dynamic filtering for data-driven pattern recognition. In: Zoeller EA (ed) *Pattern recognition: theory and application*, p 17–71
- He J, Yang S, Gan C (2017) Unsupervised fault diagnosis of a gear transmission chain using a deep belief network. *Sensors* 17(7):1564
- Holmes P (1990) Poincaré, celestial mechanics, dynamical-systems theory and "chaos". *Phys Rep* 193(3):137–163
- Hossin M, Sulaiman M (2015) A review on evaluation metrics for data classification evaluations. *Int J Data Min Knowl Manag Process* 5(2):1
- Huang J, Ling CX (2005) Using auc and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng* 17(3):299–310
- Janjarasjitt S, Ocaik H, Loparo K (2008) Bearing condition diagnosis and prognosis using applied nonlinear dynamical analysis of machine vibration signal. *J Sound Vib* 317(1):112–126
- Jedliński Ł, Jonak J (2015) Early fault detection in gearboxes based on support vector machines and multilayer perceptron with a continuous wavelet transform. *Appl Soft Comput* 30:636–641
- Karmakar CK, Gubbi J, Khandoker AH, Palaniswami M (2010) Analyzing temporal variability of standard descriptors of poicare plots. *J Electrocardiol* 43(6):719–724
- Kim JS, Park JE, Seo JD, Lee WR, Kim HS, Noh JI, Kim NS, Yum MK (2000) Decreased entropy of symbolic heart rate dynamics during daily activity as a predictor of positive head-up tilt test in patients with alleged neurocardiogenic syncope. *Phys Med Biol* 45(11):3403
- Koichubekov B, Riklifs V, Sorokina M, Korshukov I, Turgunova L, Laryushina Y, Bakirova R, Muldaeva G, Bekov E, Kultenova M (2017) Informative nature and nonlinearity of lagged poicare plots indices in analysis of heart rate variability. *Entropy* 19(10):523
- Krishnakumari A, Elayaperumal A, Saravanan M, Arvindan C (2017) Fault diagnostics of spur gear using decision tree and fuzzy classifier. *Int J Adv Manuf Technol* 89(9-12):3487–3494
- Kurths J, Voss A, Saperin P, Witt A, Kleiner H, Wessel N (1995) Quantitative analysis of heart rate variability. *Chaos: Interdiscip J Nonlinear Sci* 5(1):88–94
- Lei Y, Lin J, Zuo MJ, He Z (2014) Condition monitoring and fault diagnosis of planetary gearboxes: a review. *Measurement* 48:292–305
- Li C, de Oliveira JV, Cerrada M, Pacheco F, Cabrera D, Sanchez V, Zurita G (2016) Observer-biased bearing condition monitoring: From fault detection to multi-fault classification. *Eng Appl Artif Intell* 50:287–301

29. Li C, Sánchez RV, Zurita G, Cerrada M, Cabrera D (2016) Fault diagnosis for rotating machinery using vibration measurement deep statistical feature learning. *Sensors* 16(6):895
30. Li C, Sanchez RV, Zurita G, Cerrada M, Cabrera D, Vásquez RE (2015) Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis. *Neurocomputing* 168:119–127
31. Luts J, Ojeda F, Van de Plas R, De Moor B, Van Huffel S, Suykens JA (2010) A tutorial on support vector machine-based methods for classification problems in chemometrics. *Anal Chim Acta* 665(2):129–145
32. Medina R, Alvarez X, Jadán D, Macancela JC, Sánchez RV, Cerrada M (2017) Poincaré plot features from vibration signal for gearbox fault diagnosis. In: *Proceedings of the ETCM 2017*. Institute of Electrical and Electronics Engineers, IEEE, pp 1–6
33. Medina R, Alvarez X, Jadán D., Macancela JC, Sánchez RV, Cerrada M (2018) Gearbox fault classification using dictionary sparse based representations of vibration signals. *J Intell Fuzzy Syst* 34:3605–3618
34. Mosallam A, Medjaher K, Zerhouni N (2013) Nonparametric time series modelling for industrial prognostics and health management. *Int J Adv Manuf Technol* 69(5-8):1685–1699
35. Motahar H, Samani FS, Molaie M (2016) Nonlinear vibration of the bevel gear with teeth profile modification. *Nonlinear Dyn* 83(4):1875–1884
36. Pacheco F, de Oliveira JV, Sánchez RV, Cerrada M, Cabrera D, Li C, Zurita G, Artés M (2016) A statistical comparison of neuroclassifiers and feature selection methods for gearbox fault diagnosis under realistic conditions. *Neurocomputing* 194:192–206
37. Patel UA, Naik BS (2017) Nonlinear vibration prediction of cylindrical roller bearing rotor system modeling for localized defect at inner race with finite element approach. *Proc Instit Mech Eng, Part K: J Multi-body Dyn* 231(4):647–657
38. Patra P, Saran VH, Harsha S (2018) Non-linear dynamic response analysis of cylindrical roller bearings due to rotational speed. *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics* 232:1464419318762678
39. de Paula AV, Endres LAM, Möller SV (2016) Symbolic dynamics applied to the identification of flow patterns inside tube banks. *Nucl Sci Eng* 184(3):334–345
40. Piskorski J, Guzik P (2007) Geometry of the poincaré plot of rr intervals and its asymmetry in healthy adults. *Physiol Measur* 28(3):287
41. Porta A, Baumert M, Cysarz D, Wessel N (2015) Enhancing dynamical signatures of complex systems through symbolic computation. *Philosophical Transactions of the Royal Society of London A: Mathematical Physical and Engineering Sciences* 373(2034):1–16
42. Praveenkumar T, Saimurugan M, Krishnakumar P, Ramachandran K (2014) Fault diagnosis of automobile gearbox based on machine learning techniques. *Procedia Eng* 97:2092–2098
43. Rao C, Sarkar S, Ray A, Yasar M (2008) Comparative evaluation of symbolic dynamic filtering for detection of anomaly patterns. In: *American control conference*. IEEE, pp 3052–3057
44. Rosso O, Martin M, Figliola A, Keller K, Plastino A (2006) Eeg analysis using wavelet-based information tools. *J Neurosci Methods* 153(2):163–182
45. Sánchez RV, Lucero P, Vásquez RE, Cerrada M, Macancela JC, Cabrera D (2018) Feature ranking for multi-fault diagnosis of rotating machinery by using random forest and knn. *J Intell Fuzzy Syst* 34(6):3463–3473
46. Sanjith BMM, Sujatha C, Jayakumar T (2012) Symbolic dynamics based bearing fault detection. In: *2012 IEEE 5th India international conference on Power electronics (IICPE)*. IEEE, pp 1–5
47. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 45(4):427–437
48. Soleimani A, Khadem S (2015) Early fault detection of rotating machinery through chaotic vibration feature extraction of experimental data sets. *Chaos, Solitons Fractals* 78:61–75
49. Subbu A (2009) Pattern recognition using symbolic dynamic filtering. Ph.D. thesis, The Pennsylvania State University, Electrical Engineering, PA
50. Subbu A, Srivastav A, Ray A, Keller E (2010) Symbolic dynamic filtering for image analysis: theory and experimental validation. *Signal Image Video Process* 4(3):319–329
51. Sun Y (2012) Fault detection in dynamic systems using the largest lyapunov exponent. Ph.D. thesis, Texas A & M University, USA
52. Tang X, Tracy E, Boozer A, Brown R et al (1995) Symbol sequence statistics in noisy chaotic signal reconstruction. *Phys Rev E* 51(5):3871
53. Tucker W (2002) Computing accurate poincaré maps. *Physica D: Nonlinear Phenom* 171(3):127–137
54. Valencia JF, Vallverdu M, Rivero I, Voss A, de Luna AB, Porta A, Caminal P (2015) Symbolic dynamics to discriminate healthy and ischaemic dilated cardiomyopathy populations: an application to the variability of heart period and qt interval. *Philosophical Transactions of the Royal Society of London A: Mathematical Physical and Engineering Sciences* 373(2034):1–20
55. Von Oertzen T, Boker SM (2010) Time delay embedding increases estimation precision of models of intraindividual variability. *Psychometrika* 75(1):158–175
56. Wang Z, Xue X (2014) *Multi-class Support Vector Machine*. Springer International Publishing, Cham, pp 23–48
57. Worden K, Staszewski WJ, Hensman JJ (2011) Natural computing for mechanical systems research: a tutorial overview. *Mech Syst Signal Process* 25(1):4–111
58. Xu D, Beck C (2017) Symbolic dynamics techniques for complex systems: Application to share price dynamics. *EPL (Europhys Lett)* 118(3):30001
59. Yan R, Gao RX (2004) Complexity as a measure for machine health evaluation. *IEEE Trans Instrum Measur* 53(4):1327–1334
60. Yin Z, Hou J (2016) Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes. *Neurocomputing* 174:643–650
61. Zhang W, Jia MP, Zhu L, Yan XA (2017) Comprehensive overview on computational intelligence techniques for machinery condition monitoring and fault diagnosis. *Chin J Mech Eng* 30(4):782–795

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Ruben Medina¹  · Jean-Carlo Macancela² · Pablo Lucero² · Diego Cabrera² · Mariela Cerrada² · René-Vinicio Sánchez² · Rafael E. Vásquez³

Jean-Carlo Macancela
jmacancelap@est.ups.edu.ec

Pablo Lucero
pluceroo@est.ups.edu.ec

Diego Cabrera
dcabrera@ups.edu.ec

Mariela Cerrada
mcerrada@ups.edu.ec

René-Vinicio Sánchez
rsanchezl@ups.edu.ec

Rafael E. Vásquez
rafael.vasquez@upb.edu.co

- ¹ CIByTEL, Engineering School, Universidad de Los Andes, Mérida, Venezuela
- ² GIDTEC, Universidad Politécnica Salesiana, Cuenca, Ecuador
- ³ Department of Mechanical Engineering, Universidad Pontificia Bolivariana, Medellín, Colombia