Knowledge extraction from deep convolutional neural networks applied to cyclo-stationary time-series classification

Diego Cabrera [a, b], Fernando Sancho [c], Mariela Cerrada [b], René-Vinicio Sánchez [b], Chuan Li [a, *]

[a] School of Mechanical Engineering, Dongguan University of Technology, China
[b] GIDTEC, Universidad Politécnica Salesiana, Ecuador
[c] Dpt. of Computer Science and Artificial Intelligence, Universidad de Sevilla, España

Keywords:

Deep learning
Convolutional neural network
Cyclo-stationary time-series analysis
Knowledge extraction
Fault diagnosis

A B S T R A C T

Modelling complex processes from raw time series increases the necessity to build Deep Learning (DL) architectures that can manage this type of data structure. However, as DL models become deeper, larger and more diverse datasets are necessary and knowledge extraction will become more difficult. In an attempt to sidestep these issues, in this paper a methodology based on two main steps is presented, the first being to increase size and diversity of time-series datasets for training, and the second to retrieve knowledge from the obtained model. This methodology is compared with other approaches reported in the literature and is tested under two configuration setups of Condition-Based Maintenance problems: fault diagnosis of bearing, and fault severity assessment of a helical gearbox, obtaining not only a performance improvement in comparison, but also in retrieving knowledge about how the signals are being classified.

1. Introduction

The complexity of time series is directly related to the process that produces them, and is usually associated with their variability and the difficulty of extracting patterns that can be used to make inferences about the process.

Dynamical systems, such as engines, gearboxes, and mechanical transmission lines, that are found as key points in industrial production chains, are excellent representative of processes generating complex time series. Identification of the healthy state of their components is a central task from social [10] and economic perspectives. The behavior of signals obtained by measuring some variables, e.g., vibration, acoustic emission, oil state, and temperature, change under different states of the process [29]. However, the number and variety of signals that can be obtained is very limited in practical situations.

Changes in the system components produce a response in the measured signal. For example, vibration and acoustic emission signals have immediate responses to internal events in mechanical systems and have been traditionally used to component diagnosis [21]. These events are cyclic over time with a period that depends on the rotational speed of the component [18].

Usually, techniques from signal processing are used to recognize the presence of repetitive patterns. For example, in [19] the extraction of impulsive transients from noisy signals are proposed using a multiscale clustered gray infogram, while

* Corresponding author.
E-mail addresses: dcabrera@ups.edu.ec (D. Cabrera), chuanli@dgut.edu.cn (C. Li).

spectral kurtosis is proposed as a representation of impulsiveness in signals in [1]. Both approaches are based on the study of cyclostationarity of signals as proposed in [4] (or [8], for a more exhaustive review). However, all these works are focused on obtaining the most informative spectrum from certain patterns of frequency bands to be analyzed in order to identify interesting behaviors. They provide a physical interpretation of the phenomenon, but need a high a priori knowledge of the process configuration.

In recent years the tendency has been changed to address complex system modelling through data-driven approaches emphasizing the power of Machine Learning (ML) techniques. For reference, in [7] and [9] a classical ML approach based on feature extraction, feature selection, and parameter optimization with Random Forest classifiers were proposed. Other works try to avoid feature extraction and feature selection stages by the use of Deep Learning (DL) models, as in [14], in which a modified Convolutional Neural Network (CNN) is used for real-time fault detection in motors. The use of DL techniques allows addressing more complex problems, but the obtained models are closer to black boxes in which no information about the system can be extracted. Furthermore, datasets to train DL models should be sufficiently large to provide the desired generalization capacity in the specific application, and this size is commonly limited in real world applications, where few resources of hardware and accessibility are available.

Additionally, models based on the theory of evolving connectionist systems, which can deal naturally with time series, have been developed as a proposal closer to biological reasoning systems [15] (more detailed surveys can be found in [16] and [25]). For example, in [27] knowledge is extracted from data streams as fuzzy logic sentences by using spiking networks. These approaches present significant advances in terms of knowledge representation and interpretability; however, unlike the aim of this work, they do not seek the global representation of the knowledge captured by the model.

As will be seen, and unlike evolutionary connectionist approaches that learn from temporal changes, the extraction of knowledge from the proposed method is represented in a spectrum of frequencies with higher powers in sectors in which the network will look for the presence of frequency patterns. Therefore, this proposal is aimed at dynamic cyclo-stationary systems. In this context, an informative spectrum refers to the frequency response of the network represented as a filter pointing out the bands that will be attenuated and those that will be amplified at any input. It is *informative* because it provides additional information to be interpreted by experts. For example, in a rotating mechanical system with unknown constructive characteristics, this spectrum would provide information on the internal rotation frequencies of the components that are being diagnosed along with the bands where their faults are evident.

Motivated by the drawbacks of most of the approaches regarding the interpretability/explainability of the proposed models, in this work the extraction of an informative spectrum from a time-series signal by using a one-dimensional (1D) deep CNN (DCNN) is proposed. In this context, the main contributions of this paper are (i) a technique for dynamic batch generation that improves the generalization capacity of the DL model with long time-series as input; (ii) an algorithm to build an informative spectrum from the DL model, independently of specific time-series input; and (iii) a methodology to create diagnosis applications for dynamical systems able to provide additional information to experts.

The rest of the paper is organized as follows. In Section 2, the DL model is formalized and the proposed online dynamical batch-generation technique and informative spectrum extraction algorithm are presented, followed by presentation of a general methodology with which to build a mode identification system. To cross-check this methodology for fault diagnosis problems, two case studies and a comparison with other techniques are shown in Section 3. Finally, Section 4 presents several conclusions and directions for future work.

2. Methodology

Since the origination of CNNs [11], they have been successfully applied in image recognition tasks under a variety of conditions [17].

In the original CNN architecture the input is a 2D data structure to be passed through a layer of 2D convolutional kernels trained to extract patterns from it. This process can be repeated using a chain of convolutional layers to build DCNN models, providing a hierarchical system for pattern extraction.

To apply this model to time series, a dimensional transformation to a suitable 2D data structure must be performed. For example, in [6] a transformation from the time domain to time-frequency domain was proposed using the frequency ordered coefficients of the last level in the decomposition tree of a Wavelet Packet Transform (WPT), and in [23,30] the transformation was performed through a Short-Time Fourier Transform (STFT), in which both the time and frequency resolution of the resulting spectrogram are constants. Regardless of the transformation method, these approaches should be carefully used because they involve implicit loss of information in time or frequency due to the Uncertainty Principle [12].

In addition, the high complexity of inner structure of signals shows that lacking a correct way to measure similarity of signals from different operation modes may lead to identify two signals being captured from the same mode as distant as two signals captured from different modes. For example, Fig. 1 shows six acquired vibratory signals from five different faulty modes (Fig. 1(b) and (e) are under the same faulty mode) of a piece of rotating machinery, and Table 1 shows the distance matrix between pairs of signals using the Dynamic Time Warping (DTW) algorithm [24]. As can be seen, the most similar pairs correspond to signals from (Fault 3, Fault 5), (Fault 5, Fault 6), and (Fault 1, Fault 3), not necessarily from the same faulty mode.

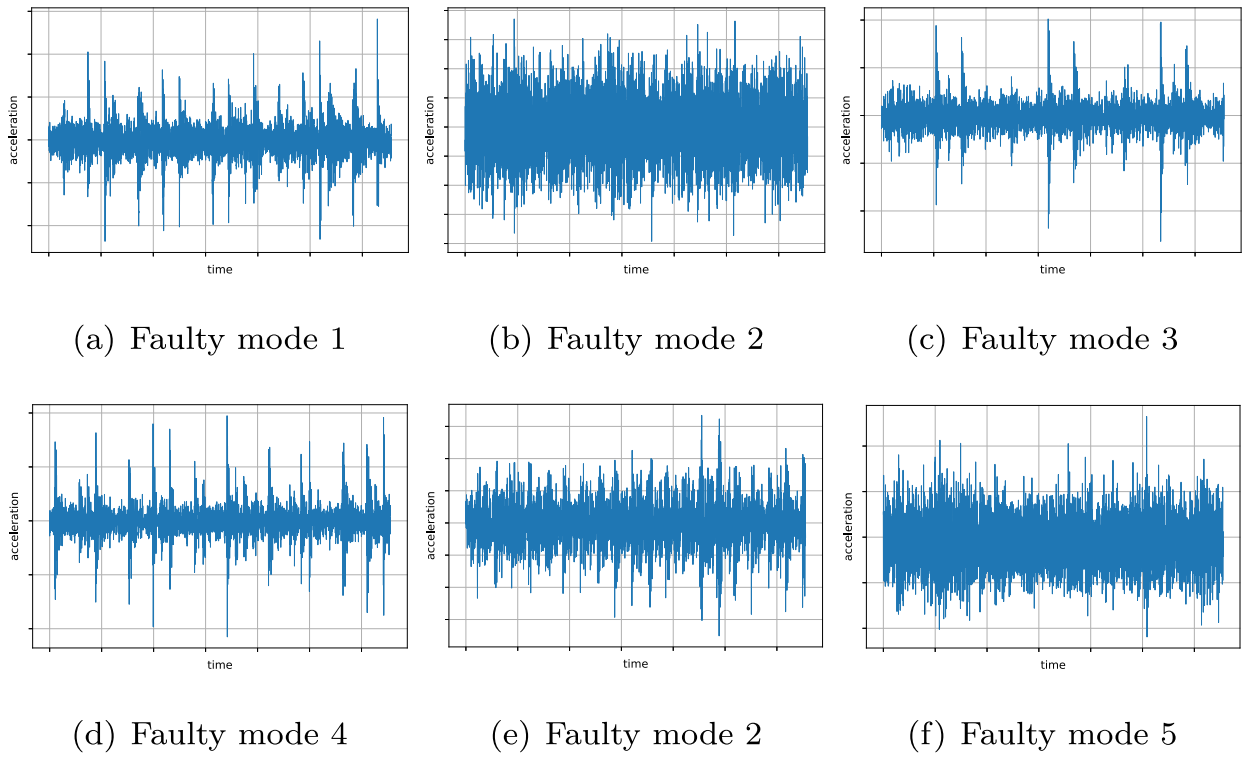


Fig. 1. Samples of acquired vibration signals of a piece of rotating machinery.

Table 1
Matrix distances (using DTW algorithm) between signals from Fig. 1.

Signals	(a)	(b)	(c)	(d)	(e)
(b)	17.355				
(c)	10.868	19.086			
(d)	17.929	19.845	20.715		
(e)	11.459	18.588	9.484	22.493	
(f)	15.885	25.305	12.684	29.59	10.714

2.1. 1D Deep convolutional neural networks

Based on the same principles of the original DCNN, a 1D DCNN, specifically designed to handle time series, is presented here. Several works have used this architecture previously but, to the best of our knowledge, none provided a formalization as follows (see Fig. 2).

In a layer l a set of 1D convolutional kernels, given by $h_l^{k,j}(t)$, are applied to a multi-channel time series $x_l^k(t)$ $(1 \leq k \leq M_l)$ in the following way to obtain the output layer o_l^j:

$$o_l^j(t) = \sum_{k=1}^{M_l} x_l^k(t) * h_l^{k,j}(t) + b_l^j, \qquad 1 \leq j \leq N_l, \tag{1}$$

where M_l and N_l are, respectively, the number of channels and kernels in the layer, b_l^j is a set of free parameters, and $t \in [1, T_l]$ is the time discrete variable, with T_l being the length of the discrete time-series input in the layer.

As usual in neural networks, an element-wise non-linear function, f, is applied to the output of the convolutional process. Several options have been commonly reported for this task, e.g., *tanh, sigmoidal*, and *ReLU* functions, that have demonstrated advantages for accelerating the convergence of deep neural network models in the training process.

Next, a sub-sampling operation is performed to obtain a translation invariance property. A popular sub-sampling is the $max - pooling$ function, which moves a sliding window of a pre-fixed size, ρ, with no overlapping, and then selects the maximum value inside the window:

$$x_{l+1}^j(\tau) = max\_pooling[f(o_l^j(t))] \tag{2}$$

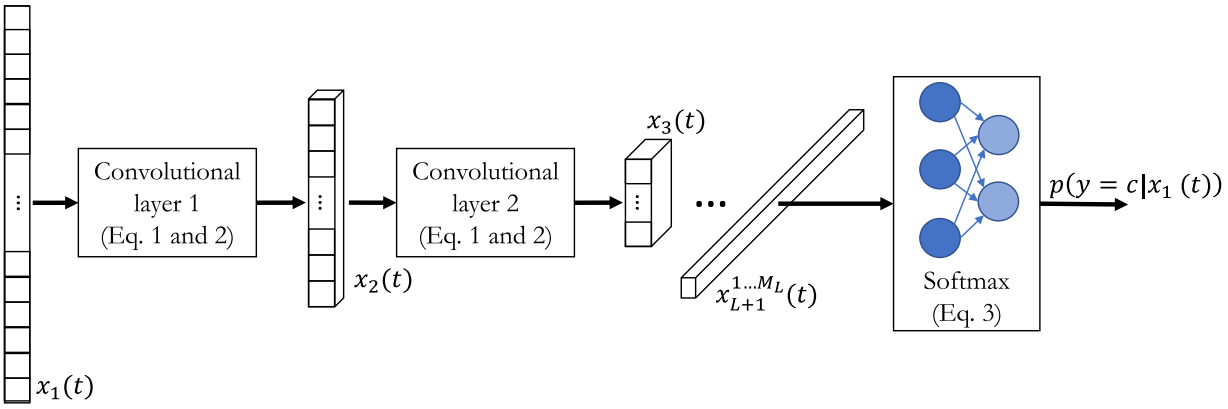


Fig. 2. Architecture of proposed 1D DCNN.

where $T_{l+1} = \frac{T_l}{\rho}$, and $\tau \in [1, T_{l+1}]$.

A model with stacked convolutional layers can be obtained by connecting the output of the lth layer with the input for the $(l+1)$th layer: $x_{l+1}^j = o_l^j$.

After the application of all these layers, the channels obtained from the last one, L, are stacked together, $x_{L+1}^{1\ldots M_1} = (x_{L+1}^1, \ldots, x_{L+1}^{M_L})$, and then an additional *softmax* layer, σ, is applied to model the probability distribution of possible modes, C, conditioned on the input evidence:

$$p_c = p(y = c|x_1^{1\ldots M_1}) = \sigma\,(x_{L+1}^{1\ldots M_L})_c \qquad \forall c \in C. \tag{3}$$

Finally, the predicted mode will be the one with higher probability:

$$\hat{y} = \underset{c\in C}{\text{argmax}}\ \ p_c. \tag{4}$$

2.2. Dynamic batch generation for time series

As usual in supervised ML models for classification, an optimization stage is applied to find the set of parameters, $\Theta = \{h_l^{k,j}(t),\ b_l^j\}_{l,k,j}$, that minimizes the cross-entropy cost function for C process modes:

$$H(\Theta) = -\frac{1}{|C|}\sum_{c\in C} \mathbf{1}_{\{y=c\}} \ln(p_c) + (1 - \mathbf{1}_{\{y=c\}}]) \ln(1 - p_c), \tag{5}$$

and for this, typically, a gradient-based algorithm is used, e.g. Stochastic Gradient Descent (SGD) or Adam, together with back-propagation to compute the gradients and a dataset D of pairs $(\mathbf{x}(t), y)$, where $\mathbf{x}(t)$ is an obtained signal and y the associated state of the system. As usual, D is partitioned into three subsets: train D_{tr}, validation D_v, and test D_t, where D_{tr} and D_v are used during the optimization stage and D_t is used for testing the accuracy and generalization capacity of the resulting model.

In several real world applications the number of instances per process mode on the dataset is very limited, which has become a problem for the application of classical ML methodologies and which is accentuated in DL-based models in which the number of parameters is larger and more data are required [28]. Moreover, as the problem becomes more complex, the number of parameters must be increased to ensure a correct discrimination between different modes.

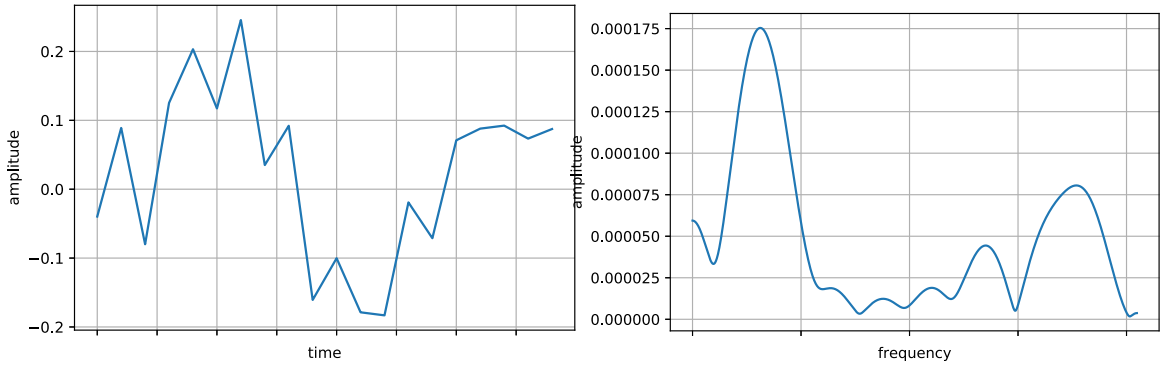
When working with signals, although the number of instances can be limited, it is not the case with their lengths. For example, in measurement-based rotating machinery diagnosis, each time series is composed of a large number of samples obtained from one measurement under some sampling rate (e.g., 10 s measured under 50,000 *samples*/*s* gives a time series of 500,000 *samples*), and this fact will be taken advantage of here to propose a modification in the batch-selection process for the learning algorithm that can be applied to any iterative gradient-based optimization method.

Essentially, our proposal consists of independently applying S times (the desired batch size) a random variable that uniformly takes an element from D_{tr} $((\mathbf{x}, y) \xleftarrow{\text{uniform}} D_{tr})$ and extracts a section of prefixed length L from its domain. This stochastic process can be inserted in every epoch of the training loop (see Algorithm 1).

This technique aims at improving the generalization capacity of the model by increasing the diversity of the samples to be used in the optimization stage. New subsignals obtained from randomly selected instances from the original dataset could overlap, but they show slight differences enough to provide better results because higher diversity batches with no repeated instances between iterations increases the information about each mode of the process.

Data: D_{tr}, L (length of subsignal), S (batch size)
Result: completed epoch on training stage
foreach *iteration in epoch* **do**
  $B = \emptyset$;
  foreach $i \leq S$ **do**
    $(\vec{x}, y) \xleftarrow{\text{uniform}} D_{tr}$;
    $t \xleftarrow{\text{uniform}} [1, T - L]$;
    $B = B \cup \{(\vec{x}_{[t,t+L)}, y)\}$;
  end
  Run training algorithm with B;
end

Algorithm 1: Re-sampling-based batch algorithm.



(a) Time response of kernel-based filter (b) Spectrum of kernel-based filter

Fig. 3. Coefficients and spectrum associated with a kernel.

2.3. Algorithm for explanatory spectrum extraction

In general, a CNN can be seen as a set of filters interacting between them in order to remove or highlight some patterns on the input. After a correct learning process, the highlighted patterns are those that respond to certain properties of the optimized filter for the learning task. This fact has special relevance in the 1D DCNN model, in which both the inputs and filters are time series and it is possible to analyze their specific significance in terms of their frequencies.

Every kernel $h_l^{k,j}(t)$ has a specific behavior as a multi-band pass filter in its frequency domain (see Fig. 3). Every filter is convolved with the input [see Eq. (1)] and highlights the frequency bands having high amplitude on the filter spectrum while reducing frequency bands with low amplitude. This phenomenon is explained by the Convolution Theorem, which states a convolution in the time domain is equivalent to a point-wise product in the frequency domain [2].

Additionally, from Eq. (1), for each filter group $\{h_l^k\}_{1 \leq k \leq M_l}$, the resulting signals of the convolution process are added together. Because of the linearity of the Fourier Transform (FT), in every layer this is equivalent to sum all the signals' frequency spectra.

From this fact, a method is proposed herein that combines these operations in all layers of the 1D DCNN in order to obtain the informative frequency spectrum codified by the network (Algorithm 2).

The input spectrum is the fast FT (FFT) of the δ impulse function, which by definition is 1 in all the frequency components. The previous statement initialize the prior knowledge about the spectrum with a constant probability distribution. In effect, if X_1^k (the spectrum of the kth input channel to the network) is considered a random variable, then its distribution is uniform. This assumption is valid for most of the problems, where some prior knowledge about the spectrum is not available. However, this can be included easily in Algorithm 2 by replacing the term $FFT(\delta)$ with a vector per input channel that would contain the highest values in the most important frequency components.

For every layer, the element-wise multiplication with $FFT(h_l^{k,j})$ produces a restriction in some frequency bands that passes to the next layer according to the filter frequency spectra, while the addition process guarantees to keep the maximum spectrum envelope in the resulting spectrum.

Data: 1D DCNN, R (layer up to which the spectrum will be obtained)
Result: Spectrum stack
$X_1^{1\dots M_1} \leftarrow FFT(\delta)$;
foreach $l < R$ **do**
  foreach $j \leq N_l$ **do**
    $X'_j = \sum_{k \leq M_l} X_l^k \odot FFT(h_l^{k,j})$;
    $X_{l+1} \leftarrow X'_j$;
  end
end
Return: $X_R^{1\dots M_R}$

Algorithm 2: Explanatory Spectrum Extraction Algorithm.

Table 2
Summary of benchmark datasets.

Dataset	Classes	Training size	Testing size	Length	Best reported
FordA	2	3601	1320	500	Shapelet Transform (ST) [13]
ECG5000	5	500	4500	140	Collective of Transformation-Based Ensembles (COTE) [3]
ECG200	2	100	100	96	Bag-of-SFA-Symbols (BOSS) [26]
StarlightCurves	3	1000	8236	1024	COTE

The output spectrum set, $X_R^{1\dots M_R}$, contains as many elements as convolutional filters available in the Rth layer. To summarize all these components in only one informative spectrum, IS_R, they just must be aggregated:

$$IS_R = \sum_{j=1}^{M_R} X_R^j. \tag{6}$$

2.4. Informative spectrum envelope

From the preceding sections, a general four-step methodology can be given to obtain an informative spectrum envelope from a set of time-series measurements obtained from a signal-generating system. Fig. 4 provides a visual understanding of the entire process:

1. **Signals Acquisition:** Obtain a set of representative signals of every interesting mode of the system (in fault diagnosis, each mode is a faulty condition, including the healthy one). The signals must contain all presented frequencies in the selected variable (from the Nyquist Theorem [22], this is achieved by ensuring an acquisition rate of, at least, twice the maximum frequency generated by the process).
2. **Model Construction:** Train and validate the 1D DCNN model with signals from the previous step. Use Algorithm 1 to maximize the diversity of signals along with the learning iterations.
3. **Layer-wise Spectrum Extraction:** Apply Algorithm 2 to extract a set of informative general spectra providing a global view of the system.
4. **Spectrum Aggregation:** Combine the spectrum set obtained from the preceding step through the sum of all its elements, obtaining the general informative spectrum.

The source code implementation of the preceding algorithms and the entire methodology is available in [5].

3. Evaluation

3.1. Experiments

To evaluate our methodology in general time-series classification tasks, it was applied to the following four benchmark datasets:

1. FordA: Diagnose whether an automotive system has a symptom through the measures of the engine noise.
2. ECG5000: Classify the congestive heart failure in five categories using the electrocardiogram (ECG) signal.
3. ECG200: Determine whether a patient has myocardial infarction using the ECG signal.
4. StarlightCurves: Classify celestial objects in one of three types through the signals obtained from their brightness changes.

Table 2 summarizes the characteristics of each one of the benchmark datasets, as well as the name of the best model applied to each case reported in the literature.

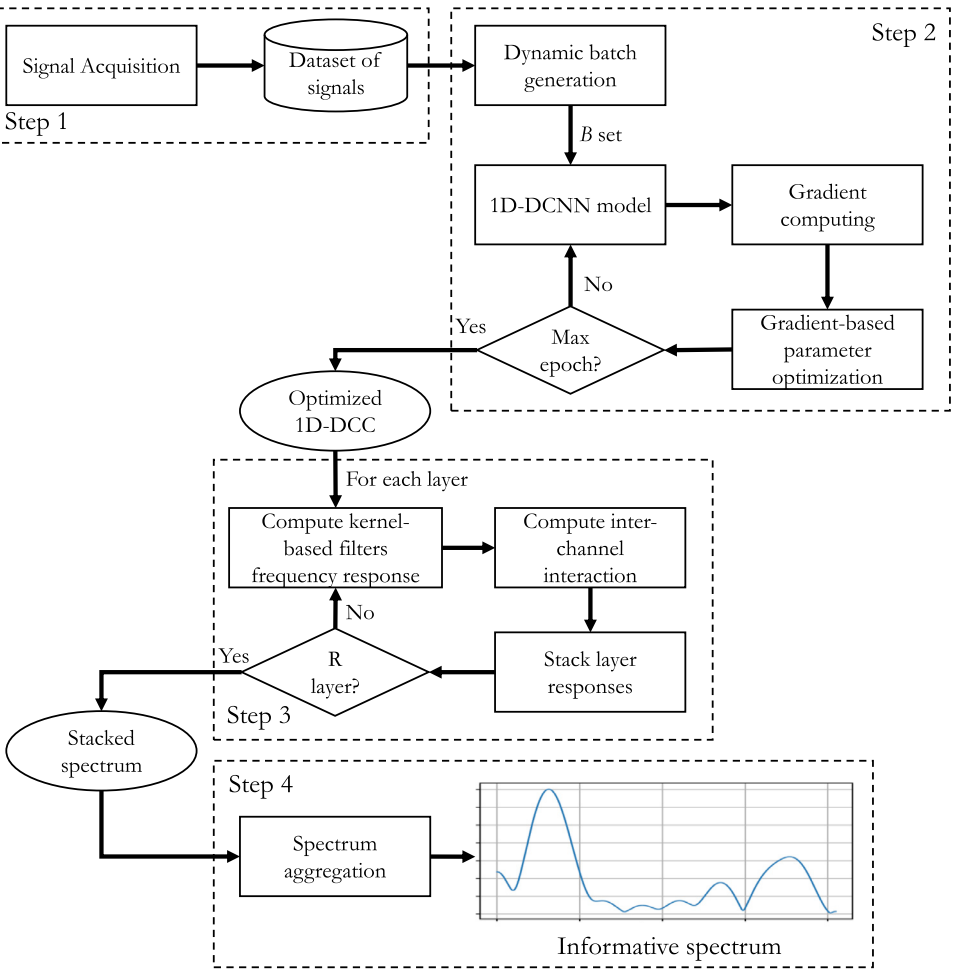


Fig. 4. Four-step methodology.

Additionally, to test our methodology, two fault diagnosis tasks in real industrial cases have been addressed: (a) mixed single and multi-component fault diagnosis on a bearing, and (b) fault severity assessment on a helical gearbox. The two experimental setups were assembled in the Vibratory Analysis Lab of the Universidad Politécnica Salesiana, Ecuador, under the support of the GIDTEC group.

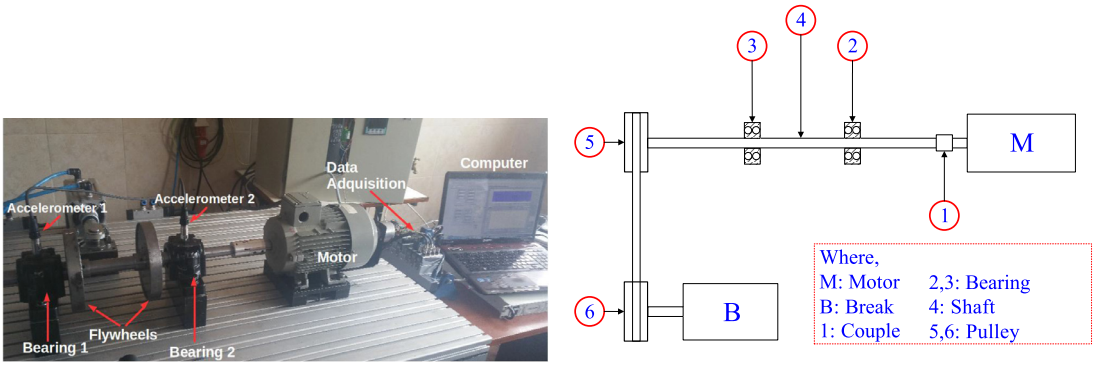
In both cases, a motor *Siemens 1LA7 090-4YA60*, with 1.49 kW and four poles, was coupled to the input of the machine to provide rotation force (driven by a speed controller *Danfoss VLT 1.5 kW*). At output, a magnetic break was coupled by a belt to simulate load [driven by a power source *TDK Lambda GEN 150-10*,$0 - 150V,\ 10A$]. The signal acquisition was performed through a Data Acquisition Card *DAQ6212* (National Instruments). The measured variable is the vibration in the vertical axis of the damaged component that was acquired by a *unidirectional accelerometer IMI Sensor 603C01*, $100mV/g$. The electrical signals returned by the accelerometer were converted from analog to digital using a *cDAQ6234* card (National Instruments).

For every faulty mode, and every operation speed and load, a pre-fixed number of signals were acquired, obtaining a labelled dataset of signals. Details for every experimental setup are presented in the following subsections.

3.1.1. Single and multi-component fault diagnosis on bearing

For this experiment, two *SKF 1207 Ektn9/C3* bearings were mounted in a $30 - mm$ shaft, each one using a housing *SKF Snl 507-606* (Fig. 5).

A total of $7(faulty\ modes) \times 3(speeds) \times 3(loads) \times 5(signals) = 315$ vibration signals were acquired. Faulty modes are described in Table 3. Speeds correspond to 8, 12, and 15 Hz of rotational speed of the motor. Loads correspond to 0, 10, and $20V$ of voltage levels of the source power of the magnetic break. The sampling rate of the acquisition card was set to 50, 000 $samples/s$ with an acquisition time of $20s$, obtaining time series of 1, 000, 000 $samples$ length.

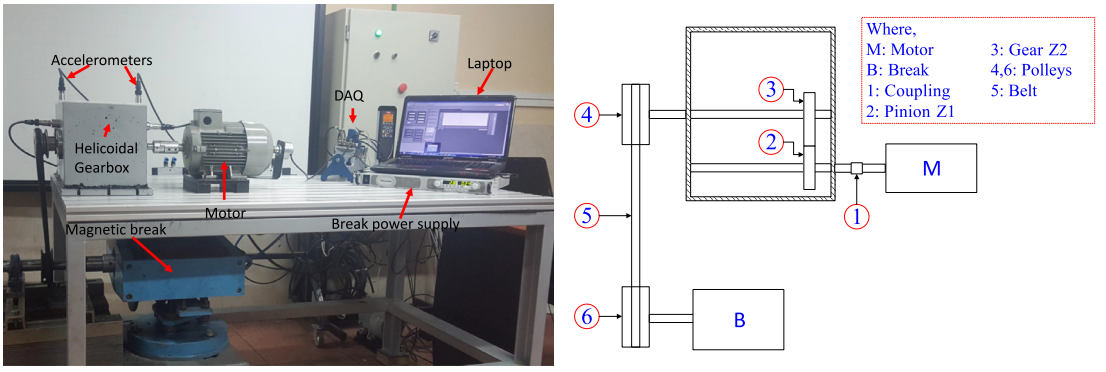


(a) Real setup (b) Schematic setup

Fig. 5. Experimental setup for single and multi-component fault diagnosis.

Table 3
Faulty modes in bearing configuration.

Faulty mode	Bearing 1	Bearing 2
P1	healthy	healthy
P2	inner race fault	healthy
P3	outer race fault	healthy
P4	rolling element fault	healthy
P5	inner race fault	outer race fault
P6	inner race fault	rolling element fault
P7	outer race fault	rolling element fault



(a) Real setup (b) Schematic setup

Fig. 6. Experimental setup for fault severity assessment on helical gearbox.

3.1.2. Fault severity assessment on helical gearbox

For this experiment, a one-stage gearbox was composed by a helical input pinion of 35 teeth coupled with an output gear of 45 teeth, both of 2.25 module and mounted on independent shafts. Both elements are built of steel *E410* and employ a Lovejoy motor as the type of coupling (Fig. 6).

A total of $10(damage\ levels) \times 5(speeds) \times 3(loads) \times 5(signals) = 750$ vibration signals were acquired. The 10 severity levels of damage are only in the input pinion (see Table 4), while the output gear is maintained in healthy condition. The five speeds correspond to 8 Hz, 12 Hz, 15 Hz, 8–15 Hz sine profile and 8–15 Hz square profile of rotational speed of the motor. In the case of the variable-speed profile, the period is $2s$. The three loads are the voltage levels of the source power of the magnetic break: 0, 10, and $30V$. The sampling rate of the acquisition card is set to 50, 000 $samples/s$ with an acquisition time of $10s$, obtaining time series of 500, 000 $samples$ length.

Table 4
Damage levels of helical gear tooth breakage fault.

Code	Description	Damage (mm)	Percentage of tooth (%)
P1	Level 1 (Normal)	0.0	100.0
P2	Level 2	2.37	88.42
P3	Level 3	4.0	80.42
P4	Level 4	5.73	71.94
P5	Level 5	7.6	62.81
P6	Level 6	10.57	48.29
P7	Level 7	12.37	39.48
P8	Level 8	14.33	29.85
P9	Level 9	17.5	14.36
P10	Level 10 (without tooth)	20.43	0.0

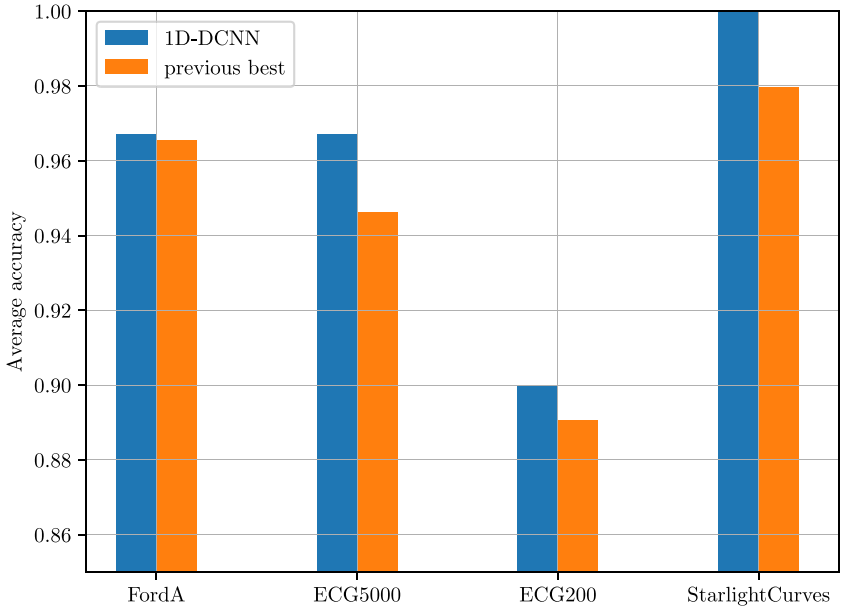


Fig. 7. Comparative classification results in benchmarks.

Table 5
1D DCNN architectures.

ID	Number of layers	Number of kernels	Length of kernels
1D DCNN-1	1	18	20
1D DCNN-2	2	18×36	20
1D DCNN-3	3	$18 \times 36 \times 72$	20
1D DCNN-4	4	$18 \times 36 \times 72 \times 144$	20

3.2. Results and discussion

First, the comparison results of the proposal with the other methods reported in the literature for the benchmark cases in Table 2 are presented in Fig. 7. The results show positive improvements obtained with the 1D DCNN in all cases. Specifically, improvements of 0.16%, 2.09%, 0.95%, and 2.04% were obtained for FordA, ECG5000, ECG200, and StarlightCurves, respectively.

As in the reported models, length of time-series and size of training dataset affect the 1D DCNN performance. However, due to the re-sampling algorithm, this effect is decreased as the signal length is allowed to increment the size of the dataset (as the results for ECG5000 show). In the same way, the obtained results for ECG200 suggest that when dataset size and signal length are too low the resulting model shows poor performance because re-sampling cannot be correctly applied.

Second, following the proposed methodology, four model architectures (Table 5) were analyzed for each fault diagnosis problem. Each model was trained using Algorithm 1 with a learning rate $= 0.0005$, no decay rate, 40 epochs, 100 batches, and batch size $= 300$. The length of the signals was 16, 384 *samples*, equivalent to 0.32768*s*. In each layer a sub-sampling

Table 6
Case 1. Accuracy of 1D DCNN and PFSC models.

ID	Accuracy
1D DCNN-1	0.771
1D DCNN-2	0.974
1D DCNN-3	0.993
1D DCNN-4	**0.999**
PFSC	0.975

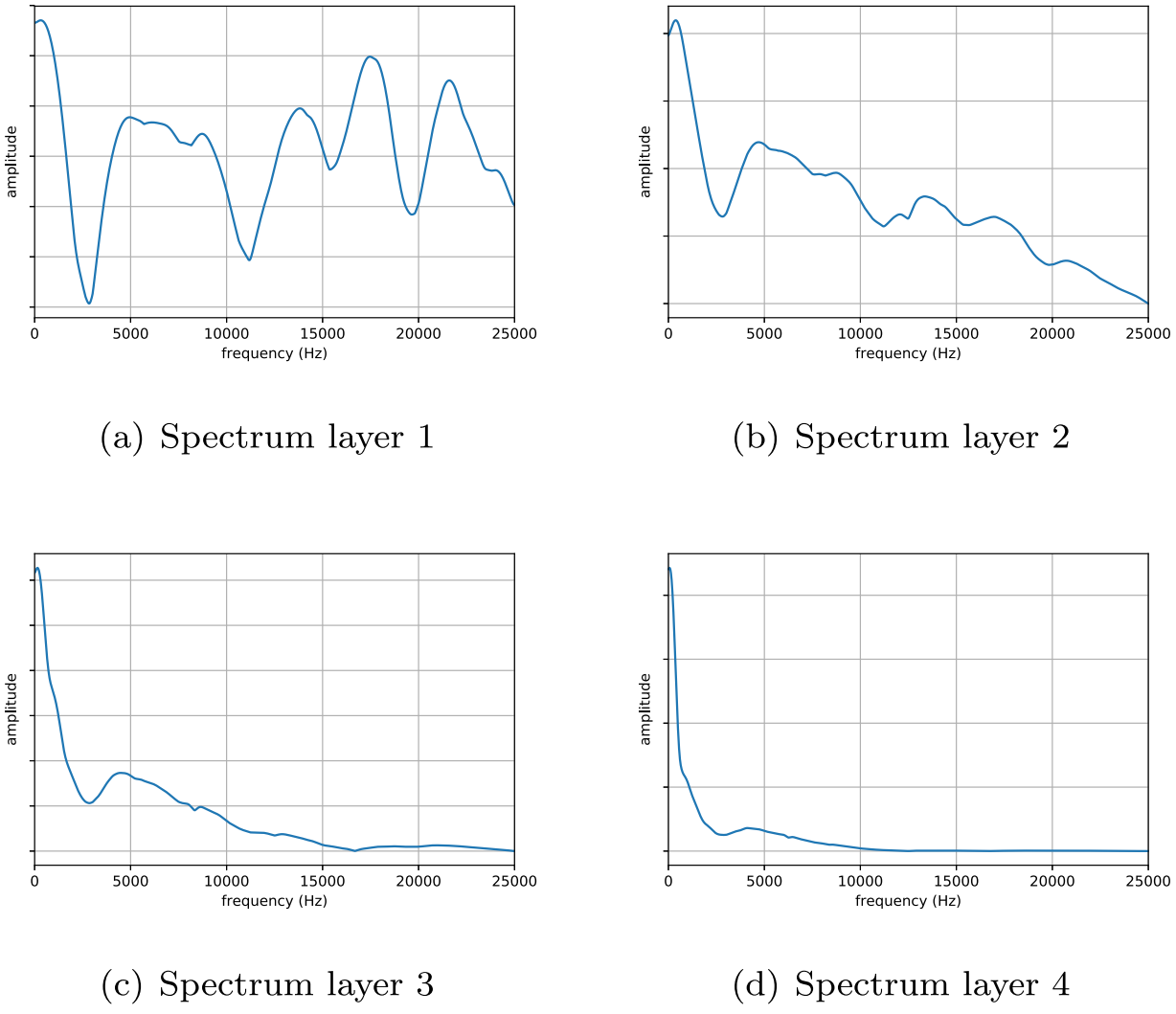
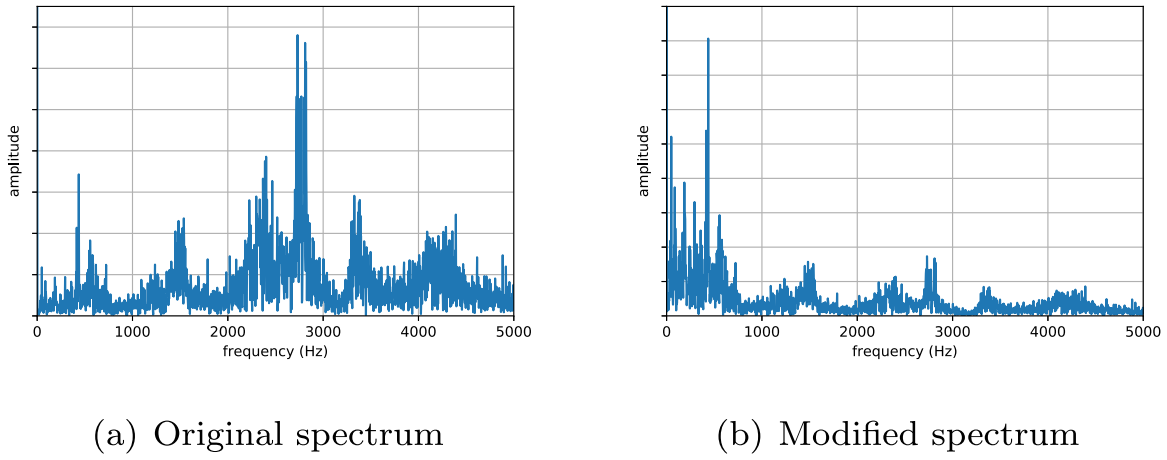


Fig. 8. Case 1. Layer-wise spectrum of 1D DCNN-4.

by a factor of 2 was carried out, and hence the number of kernels is duplicated in order to maintain constant information through layers.

Table 6 shows the accuracy of each model for the first experimental setup. Additionally, results of the Probabilistic Fuzzy System Classifier (PFSC) (in which the authors added an additional feature-extraction stage before building the classifier [20]) are shown as a comparison with other models reported in the literature applied to the same dataset. Signals from the test set (1, 000, 000 *samples* length) are split, without overlapping, in sub-signals of 16, 384 *samples* length to be similar to those from the training set, maintaining the class label.

The informative spectrum obtained from Algorithm 2 for the best model, 1D DCNN-4, is presented in Fig. 8. In the spectrum of layer 1, the inclusion of all the frequencies from 0 up to the Nyquist value can be seen. As depth of layers is increased, high frequencies are discarded, resulting in a frequency range from 0 up to 10,000 Hz in layer 4. One can deduce



(a) Original spectrum (b) Modified spectrum

Fig. 9. Case 1.Information extraction from P3 faulty mode test signal.

Table 7
Case 2. Accuracy of 1D DCNN and SCAE models.

ID	Accuracy
1D DCNN-1	0.471
1D DCNN-2	0.924
1D DCNN-3	0.949
1D DCNN-4	**0.975**
SCAE	0.947

that, for this task, reducing the sampling rate to 20, 000 *samples*/*s* is possible with the inclusion of an analog low-pass pre-filter with a cutoff frequency of 10,000 Hz to avoid aliasing in the captured signal. Furthermore, the spectra show that the most relevant information is in the low-frequency range between 0 and 2000 Hz.

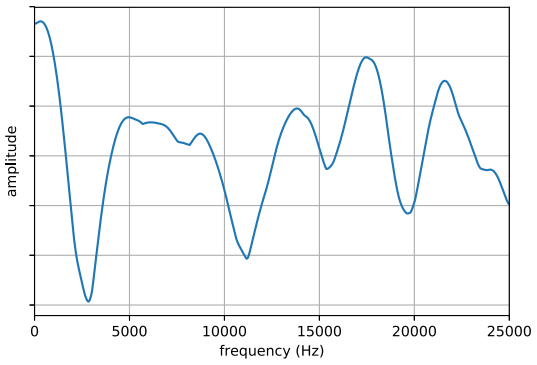
Fig. 9 is included in order to analyze the incidence of the above informative spectrum over a specific P3 fault mode test signal, showing the original and modified spectra, which were computed by element-wise multiplication of the original signal spectrum and the 1D DCNN-4 informative spectrum (limited to 5000 Hz because over that frequency its amplitude is negligible). The result shows a significant reduction of the maximum peak frequency in the original spectrum and highlights a frequency peak at 436.45 Hz. Since the P3 faulty mode is related to outer race damage, the input frequency to the machine is approximately 15 Hz and it has 15 rolling elements; then, the frequency to the rolling element put in contact with the fault is 225 Hz. Consequently, the highlighted frequency is its first harmonic. The minimal difference in the calculus is due to an input frequency deviating from 15 Hz; in fact, the measured input frequency is 14.5 Hz, obtaining a perfect coincidence with the highlighted first harmonic and showing perfect physical equivalence with the information provided by the spectrum generated by the proposed algorithm.

Table 7 shows the accuracy of each 1D DCNN model for the second experimental setup. The SCAE model from [6] is included as a comparison with other models reported in the literature applied to the same dataset. This model uses an unsupervised feature-extraction process through a 2D Deep Stacked Convolutional AutoEncoder together with a CNN working on a representation in time-frequency domain of the time-series obtained using the Wavelet Packet Transform. Again, 1D DCNN-4 is the best model, overcoming the performance of the SCAE model by 2.8%.

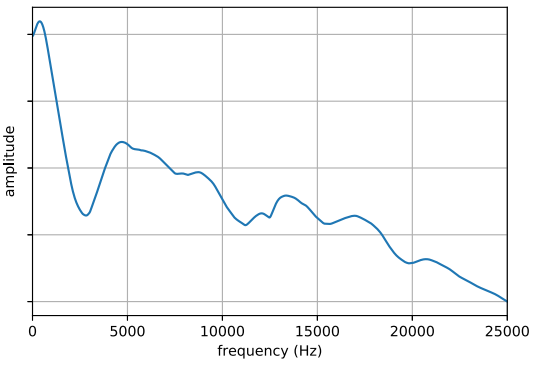
Table 8 presents more detailed comparison between the 1D DCNN-4 and SCAE models, showing the difference between Precision, Recall, and *F*-value metrics obtained from the evaluation of both models. 1D DCNN-4 shows an improvement in Precision and Recall on all faulty modes with an exception in Precision for P4 (but its effect is vanishing in comparison with the increments in the other faulty modes). In some modes, e.g., P6 where Recall is very similar (0.915 in 1D DCNN-4 and 0.913 in SCAE), the increment of 0.071 in Precision shows an improvement in the performance of 1D DCNN-4.

Fig. 10 shows the informative spectrum obtained from Algorithm 2 for the best model, i.e., 1D-DCNN-4. As in the previous case, the spectrum of layer-1 includes a large group of frequencies from 0 up to the Nyquist value and, as the depth is increased, high frequencies are discarded, resulting in an informative frequency range from 0 up to 6000 Hz in layer-4. This means that for this task reducing the sampling rate to 12, 000 *samples*/*s* is possible with the inclusion of analog low-pass pre-filter with a cutoff frequency of 6000 Hz in order to avoid aliasing in the captured signal. Furthermore, the spectrum shows that the most relevant information is in the low-frequency range between 0 and 3000 Hz.

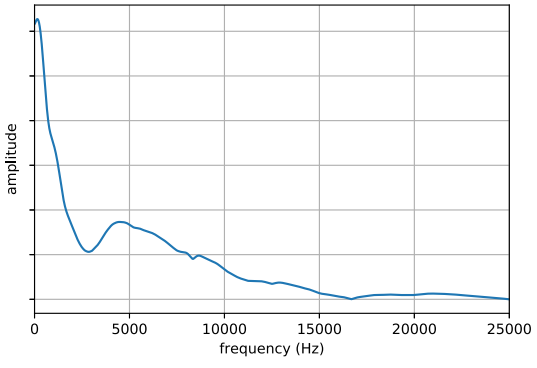
As in the previous case, Fig. 11 shows the analysis of the incidence of the informative spectrum over a specific test signal labelled in the P3 fault mode. The spectrum has been limited to 10,000 Hz because over that frequency its amplitude is



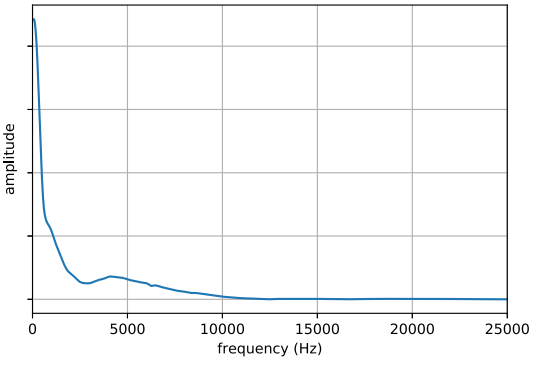
(a) Spectrum layer 1



(b) Spectrum layer 2

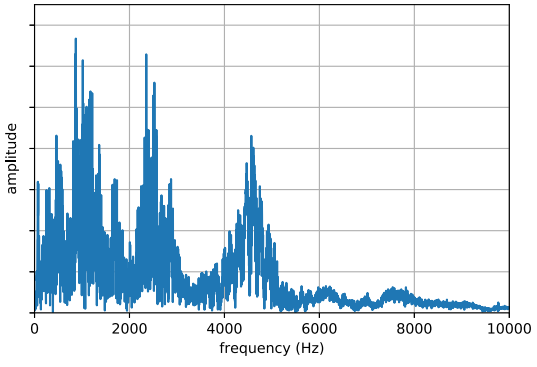


(c) Spectrum layer 3

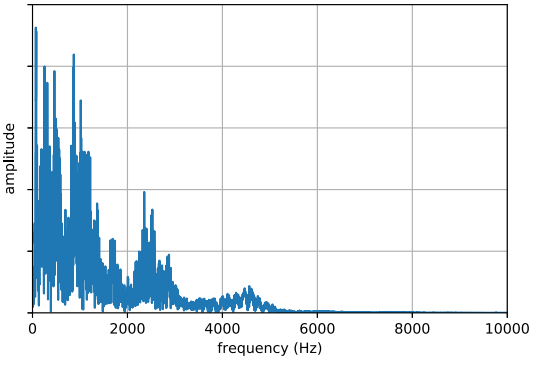


(d) Spectrum layer 4

Fig. 10. Case 2. Layer-wise spectrum of 1D DCNN-4.



(a) Original spectrum



(b) Modified spectrum

Fig. 11. Case 2. Information extraction from P3 faulty mode test signal.

Table 8
Case 2. Detailed differences for faulty modes between 1D DCNN-4 and SCAE models.

Faulty mode	Difference precision	Difference recall	Difference *F*-value
P1	0.023	0.009	0.014
P2	0.026	0.014	0.020
P3	0.025	0.060	0.042
P4	−0.010	0.049	0.018
P5	0.050	0.049	0.049
P6	0.071	0.002	0.035
P7	0.037	0.068	0.052
P8	0.036	0.004	0.020
P9	0.003	0.024	0.014
P10	0.032	0.003	0.017
Avg.	0.029	0.028	0.028

negligible. One can see a significant reduction of the peak between 4000 and 5000 Hz in the original spectrum and the highlight of the range is 0–1000 Hz, showing that it as a noise source that hides the critical information.

4. Conclusions

In this paper, three main goals were attained to manage a deep learning process on time-series data. The first, not fully new, was to provide a complete adaptation of Deep Convolutional Neural Networks for 1D cases in order to use them for multi-channel signal learning. In the second, an algorithm was provided to increase the diversity of time series to be used during the batch-selection stage in iterated learning processes that can be applied together with several gradient-based training algorithms. Finally, the last goal was to provide an algorithm that can extract an informative spectrum of the process from the deep-layer model, adding an explainable layer to the usual black box nature of deep learning models.

Additionally, a general methodology was built and applied to fault diagnosis tasks in rotating machinery, as well as to four benchmark datasets, showing an improvement in accuracy in comparison to state-of-the-art techniques.

The models built by this methodology can successfully discriminate captured signals from different health states of machinery under multiple operating conditions of speed and load, showing that the models capture aspects from the inner states of the systems. Moreover, most importantly, the proposed methodology provides the extraction of information as frequency patterns that can be interpreted for a signal processing expert (not necessarily an expert in Deep Learning), increasing the knowledge about the system/process that generates the signals. In the specific applications shown in the paper, the obtained spectra show that low frequencies store the most discriminative information, and consequently the maximum informative frequency that allows us to reduce the acquisition hardware requirements for the task.

In contrast to other learning-based models, the proposed methodology allows one to extract knowledge of the system from the resulting model, transforming a learning process into a knowledge-extraction process.

Specifically, the algorithms and methodology presented here show additional advantages compared with other machine learning and signal processing approaches when applied to fault diagnosis in rotating machinery: the performance obtained by the models built for single and multi-fault diagnosis and fault severity evaluation tasks overcome those reported by other state-of-the-art approaches, while adding an explainable layer.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Diego Cabrera: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft. **Fernando Sancho:** Supervision, Writing - review & editing, Conceptualization. **Mariela Cerrada:** Resources, Funding acquisition, Writing - review & editing, Project administration. **René-Vinicio Sánchez:** Resources, Funding acquisition, Writing - review & editing, Project administration. **Chuan Li:** Supervision, Writing - review & editing, Resources, Funding acquisition.

Acknowledgements

The work was sponsored in part by GIDTEC Project No. 003-002-2016-03-03 and the National Natural Science Foundation of China (Grant No. 51775112). The experimental work was developed at the GIDTEC Research Group Lab of the Universidad Politcnica Salesiana, Cuenca, Ecuador. This research was also partially supported by the projects TIN2017-82113-C2-1-R (VICTORY Project) and TIN2013-41086-P (LOCOCIDA Project), both from Ministerio de Economía, Industria y Competitividad of Spain, with FEDER funds from the European Union.

References

[1] J. Antoni, R. Randall, The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines, Mech. Syst. Signal Process. 20 (2) (2006) 308–331, doi:10.1016/j.ymssp.2004.09.002.

[2] G.B. Arfken, H.J. Weber, Mathematical Methods for Physicists, 6th Edition, Academic Press, 2005.

[3] A. Bagnall, J. Lines, J. Hills, A. Bostrom, Time-series classification with COTE: the collective of transformation-based ensembles, IEEE Trans. Knowl. Data Eng. 27 (9) (2015) 2522–2535, doi:10.1109/tkde.2015.2416723.

[4] S. Braun, B. Seth, Analysis of repetitive mechanism signatures, J. Sound Vib. 70 (4) (1980) 513–526, doi:10.1016/0022-460x(80)90321-1.

[5] D. Cabrera, Cnn1d, 2018, (https://github.com/diegoroman17/cnn1d).

[6] D. Cabrera, F. Sancho, C. Li, M. Cerrada, R.-V. Sánchez, F. Pacheco, J.V. de Oliveira, Automatic feature extraction of time-series applied to fault severity assessment of helical gearbox in stationary and non-stationary speed operation, Appl. Soft Comput. 58 (2017) 53–64, doi:10.1016/j.asoc.2017.04.016.

[7] D. Cabrera, F. Sancho, R.-V. Sánchez, G. Zurita, M. Cerrada, C. Li, R.E. Vásquez, Fault diagnosis of spur gearbox based on random forest and wavelet packet decomposition, Front. Mech. Eng. 10 (3) (2015) 277–286, doi:10.1007/s11465-015-0348-8.

[8] M. Cerrada, R.-V. Sánchez, C. Li, F. Pacheco, D. Cabrera, J.V. de Oliveira, R.E. Vásquez, A review on data-driven fault severity assessment in rolling bearings, Mech. Syst. Signal Process. 99 (2018) 169–196, doi:10.1016/j.ymssp.2017.06.012.

[9] M. Cerrada, G. Zurita, D. Cabrera, R.-V. Sánchez, M. Artés, C. Li, Fault diagnosis in spur gears based on genetic algorithm and random forest, Mech. Syst. Signal Process. 70–71 (2016) 87–103, doi:10.1016/j.ymssp.2015.08.030.

[10] Y. Chinniah, Analysis and prevention of serious and fatal accidents related to moving parts of machinery, Saf. Sci. 75 (2015) 163–173, doi:10.1016/j.ssci.2015.02.004.

[11] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (4) (1980) 193–202, doi:10.1007/bf00344251.

[12] K. Gröchenig, Uncertainty principles for time-frequency representations, in: Advances in Gabor Analysis, Birkhäuser Boston, 2003, pp. 11–30, doi:10.1007/978-1-4612-0133-5_2.

[13] J. Hills, J. Lines, E. Baranauskas, J. Mapp, A. Bagnall, Classification of time series by shapelet transformation, Data Min. Knowl. Discov. 28 (4) (2014) 851–881.

[14] T. Ince, S. Kiranyaz, L. Eren, M. Askar, M. Gabbouj, Real-time motor fault detection by 1-d convolutional neural networks, IEEE Trans. Ind. Electron. 63 (11) (2016) 7067–7075, doi:10.1109/tie.2016.2582729.

[15] N. Kasabov, Evolving Connectionist Systems, Springer London, 2007, doi:10.1007/978-1-84628-347-5.

[16] , Springer Handbook of Bio-/Neuroinformatics, N. Kasabov (Ed.), Springer Berlin Heidelberg, 2014, doi:10.1007/978-3-642-30574-0.

[17] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, doi:10.1109/5.726791.

[18] Y. Lei, D. Kong, J. Lin, M.J. Zuo, Fault detection of planetary gearboxes using new diagnostic parameters, Meas. Sci. Technol. 23 (5) (2012) 55605.

[19] C. Li, D. Cabrera, J.V. de Oliveira, R.-V. Sanchez, M. Cerrada, G. Zurita, Extracting repetitive transients for rotating machinery diagnosis using multiscale clustered grey infogram, Mech. Syst. Signal Process. 76–77 (2016) 157–173, doi:10.1016/j.ymssp.2016.02.064.

[20] C. Li, L. Ledo, M. Delgado, M. Cerrada, F. Pacheco, D. Cabrera, R.-V. Sánchez, J.V. de Oliveira, A bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification, Knowl. Based Syst. 129 (2017) 39–60, doi:10.1016/j.knosys.2017.05.007.

[21] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, R.E. Vásquez, Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals, Mech. Syst. Signal Process. 76–77 (2016) 283–293, doi:10.1016/j.ymssp.2016.02.007.

[22] D. Linden, N. Abramson, A generalization of the sampling theorem, Inform. Control 3 (1) (1960) 26–31, doi:10.1016/s0019-9958(60)90242-4.

[23] R.P. Monteiro, M. Cerrada, D.R. Cabrera, R.V. Sánchez, C.J.A. Bastos-Filho, Using a support vector machine based decision stage to improve the fault diagnosis on gearboxes, Comput. Intell. Neurosci. 2019 (2019) 1–13, doi:10.1155/2019/1383752.

[24] S. Salvador, P. Chan, Toward accurate dynamic time warping in linear time and space, Intell. Data Anal. 11 (5) (2007) 561–580.

[25] S. Schliebs, N. Kasabov, Evolving spiking neural network—a survey, Evol. Syst. 4 (2) (2013) 87–98, doi:10.1007/s12530-013-9074-9.

[26] P. Schäfer, The BOSS is concerned with time series classification in the presence of noise, Data Min. Knowl. Discov. 29 (6) (2014) 1505–1530, doi:10.1007/s10618-014-0377-7.

[27] S. Soltic, N. Kasabov, Knowledge extraction from evolving spiking neural networks with rank order population coding, Int. J. Neural Syst. 20 (6) (2010) 437–445, doi:10.1142/s012906571000253x.

[28] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era (2017).

[29] C.K. Tan, P. Irving, D. Mba, A comparative experimental study on the diagnostic and prognostic capabilities of acoustics emission, vibration and spectrometric oil analysis for spur gears, Mech. Syst. Signal Process. 21 (1) (2007) 208–233, doi:10.1016/j.ymssp.2005.09.015.

[30] L.-H. Wang, X.-P. Zhao, J.-X. Wu, Y.-Y. Xie, Y.-H. Zhang, Motor fault diagnosis based on short-time fourier transform and convolutional neural network, Chin. J. Mech. Eng. 30 (6) (2017) 1357–1368, doi:10.1007/s10033-017-0190-5.