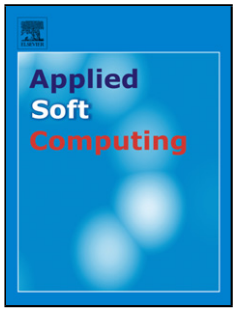


Title: Automatic Feature Extraction of Time-Series applied to Fault Severity Assessment of Helical Gearbox in Stationary and Non-Stationary Speed Operation

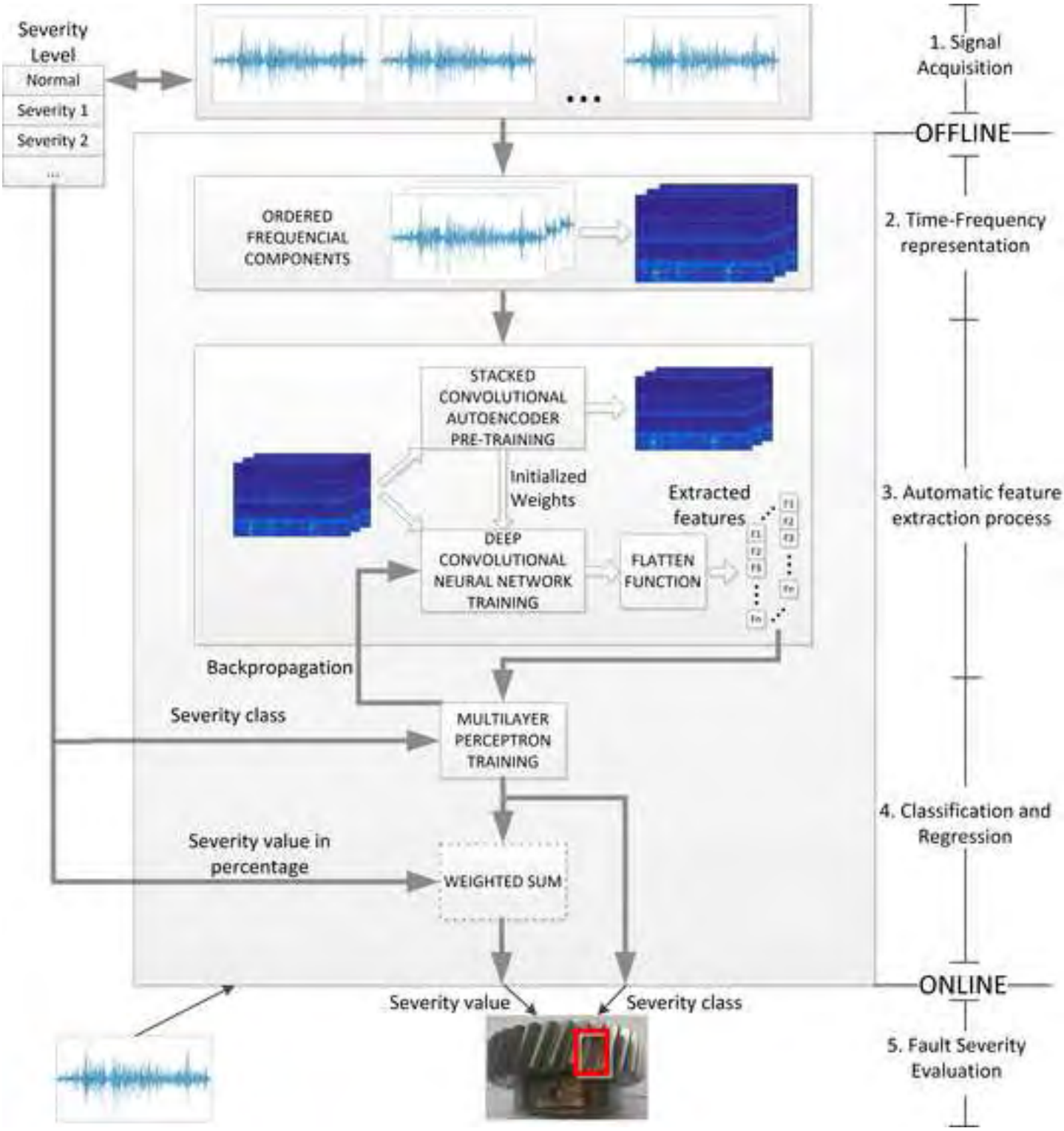
Author: Diego Cabrera Fernando Sancho Chuan Li Mariela Cerrada René-Vinicio Sánchez Fannia Pacheco José Valente de Oliveira

PII: S1568-4946(17)30188-6
DOI: http://dx.doi.org/doi:10.1016/j.asoc.2017.04.016
Reference: ASOC 4149

To appear in: *Applied Soft Computing*

Received date: 6-4-2016
Revised date: 6-12-2016
Accepted date: 7-4-2017

Please cite this article as: Diego Cabrera, Fernando Sancho, Chuan Li, Mariela Cerrada, René-Vinicio Sánchez, Fannia Pacheco, José Valente de Oliveira, Automatic Feature Extraction of Time-Series applied to Fault Severity Assessment of Helical Gearbox in Stationary and Non-Stationary Speed Operation, *<![CDATA[Applied Soft Computing Journal]]>* (2017), http://dx.doi.org/10.1016/j.asoc.2017.04.016



Automatic Feature Extraction of Time-Series applied to Fault Severity Assessment of Helical Gearbox in Stationary and Non-Stationary Speed Operation

Diego Cabrera[a,b,*], Fernando Sancho[b], Chuan Li[c], Mariela Cerrada[d], René-Vinicio Sánchez[a], Fannia Pacheco[a], José Valente de Oliveira[e]

[a] *Department of Mechanical Engineering, Universidad Politécnica Salesiana sede Cuenca, Ecuador*
[b] *Department of Computer Science and Artificial Intelligence, Universidad de Sevilla, España*
[c] *Chongqing Key Laboratory of Manufacturing Equipment Mechanism Design and Control, Chongqing Technology and Business University, Chongqing, China*
[d] *Control Systems Department, Universidad de Los Andes, Mérida, Venezuela*
[e] *CEOT, Universidade do Algarve, Portugal*

Abstract

Signals captured in rotating machines to obtain the status of their components can be considered as a source of massive information. In current methods based on artificial intelligence to fault severity assessment, features are first generated by advanced signal processing techniques. Then feature selection takes place, often requiring human expertise. This approach, besides time-consuming, is highly dependent on the machinery configuration as in general the results obtained for a mechanical system cannot be reused by other systems. Moreover, the information about time events is often lost along the process, preventing the discovery of faulty state patterns in machines operating under time-varying conditions. In this paper a novel method for automatic feature extraction and estimation of fault severity is proposed to overcome the drawbacks of classical techniques. The proposed method employs a Deep Convolutional Neural Network pre-trained by a Stacked Convolutional Autoencoder. The robustness and accuracy of this new method are validated using a dataset with different severity conditions on failure mode in a helical gearbox, working in both constant and

*Corresponding author
Email address: dcabrera@ups.edu.ec (Diego Cabrera)

variable speed of operation. The results show that the proposed unsupervised feature extraction method is effective for the estimation of fault severity in helical gearbox, and it has a consistently better performance in comparison with other reported feature extraction methods.

Keywords: Deep Learning, convolution, auto-encoder, wavelet packets, helical gearbox

1. Introduction

Gearboxes are fundamental components in rotating machines mainly composed by gears, bearings and shafts. These parts interact in a lubricated environment to minimize the friction effects(Kateris et al., 2014). In this context, the most common failure modes in gearboxes can happen either by mechanical components or by lubrication conditions (Nie & Wang, 2013). Gearbox failures can produce undesired machinery stops, causing huge economic losses and even fatal accidents (Chinniah, 2015). Hence, it is important to be able to recognize the condition of each component in an easy way and at a reasonable cost.

Gear wear is a specific failure mode that can appear in all the stages of the device useful life. This fault mode might start from the beginning of the machine operation and increases over time. The gear wear identification allows detecting incipient faults and facilitates the synchronization with planning process, inventory management and it is close to "on-time" maintenance.

The work in Jardine et al. (2006) shows that historically the main approaches, to diagnose the device conditions in rotating machinery, are: i) waveform data analysis, ii) value type data analysis, and iii) data analysis combining event data and condition monitoring data. In the first case, time-frequency based techniques have out-stood in the representation of information; e.g., Fan & Zuo (2006) have shown that Hilbert transform combined with Wavelet Packet Decomposition are suitable to obtain the fault characteristic features. In the second one, condition indicators are designed to predict the status of machinery devices. Finally, the third approach has emerged from advances in machine

learning techniques over the last years. As an example, Chen et al. (2013) show a Support Vector Machine classifier with a novel wavelet kernel function.

Generally speaking, the approaches mentioned above are conceived to work independently from each other. The link, that usually allows to join waveform data analysis with other techniques is the intermediate stage of feature extraction. The feature extraction process could be accomplished computing statistics metrics from time, frequency, or time-frequency domain of the signal representation. For example in (Vakharia et al., 2015) classical feature extraction and different feature selection techniques are performed for a fault diagnosis application using a Support Vector Machine classifier. However the designing of appropriate features is not a trivial task under different stationary or non-stationary operational parameters (Li et al., 2013), where the frequency spectrum of an acquired vibration signal shows changes in position and shape at different speeds. Historically the feature extraction task is highly dependent on the application problem, and for fault diagnosis in mechanical systems, it is mainly dependent on the machine to be analyzed. Hence, it is necessary to find a novel approach oriented to the original ideas of machine learning which try to merge data processing, feature extraction, knowledge representation, unsupervised and supervised learning in a unique model, where every part interacts directly with each other and where refinement strategies are applied together.

This paper introduces Stacked Convolutional Autoencoders (SCAE) together with Deep Convolutional Neural Network (DCNN) as a method for unsupervised hierarchical feature extraction for fault severity assessment in helical gearbox. The feature extraction process uses raw time series data under stationary and non-stationary operational conditions. The main contribution of the proposed method is the expert-free accuracy improvement of fault severity assessment estimation. This is carried out through unsupervised detection of a hierarchy of time-frequency patterns local or globally related to each other using DCNN, regardless the zone where these patterns occur. DCNN is enhanced with SCAE used for capturing a-priori patterns, and for pre-initializing the parameters of the DCNN. The results show that the proposed method is effective for the es-

timation of fault severity (severity class and fault size in percentage) in helical gearbox.

The remaining of the paper is organized as follows. Section 2 provides the background of convolutional neural networks and convolutional autoenconders. Section 3 describes in detail the proposed method and applications to fault severity diagnosis. Section 4 shows the experimental setup used for the case study of fault severity assessment in helical gearbox. Subsequently, the results of the proposed method and the comparison with similar unsupervised feature extraction approaches, as (Mdlazi et al., 2007) and (ZHOU et al., 2015), and supervised feature extraction approach, as (Vakharia et al., 2015), are performed in Section 5. Finally, Section 6 presents the main conclusions and contributions obtained in this work.

2. Preliminaries

2.1. Convolutional Neural Network

Convolutional Neural Network (CNN) is a learning model bio-inspired by animal visual cortex which tries to learn an optimum set of g kernels relatively to a specific task from a dataset. CNN has been originally used in image recognition tasks, for example the model LeNet-5 in Lecun et al. (1998) is used to character recognition.

Equations 1 and 2 summarize the complete feature extraction process at layer m from I features maps in the layer $m-1$, Figure 1 shows the architecture of the CNN:

$$h_k^m[x,y] = o\left(\sum_{i=0}^{I^{m-1}-1} f_i^{m-1} * g_{i,k}^m + b_k^m\right) \tag{1}$$

$$f_k^m[x,y] = max - pooling(h_k^m) \tag{2}$$

where I is the number of input functions f, g is the kernel function, b is a bias factor, and o is an optional non-linear function.

The parameters of a CNN can be learned using a slightly modified version of the stochastic gradient descent (SGD) algorithm, since it needs to compute

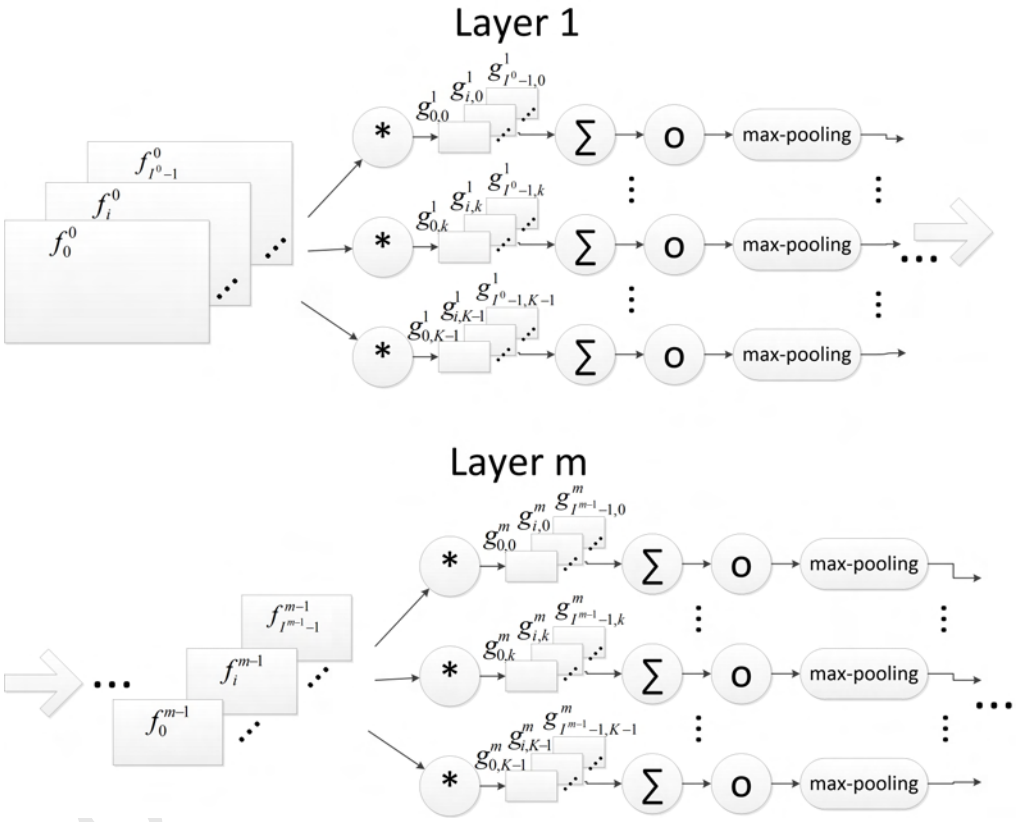


Figure 1: Typical extraction of feature maps in Convolutional Neural Network

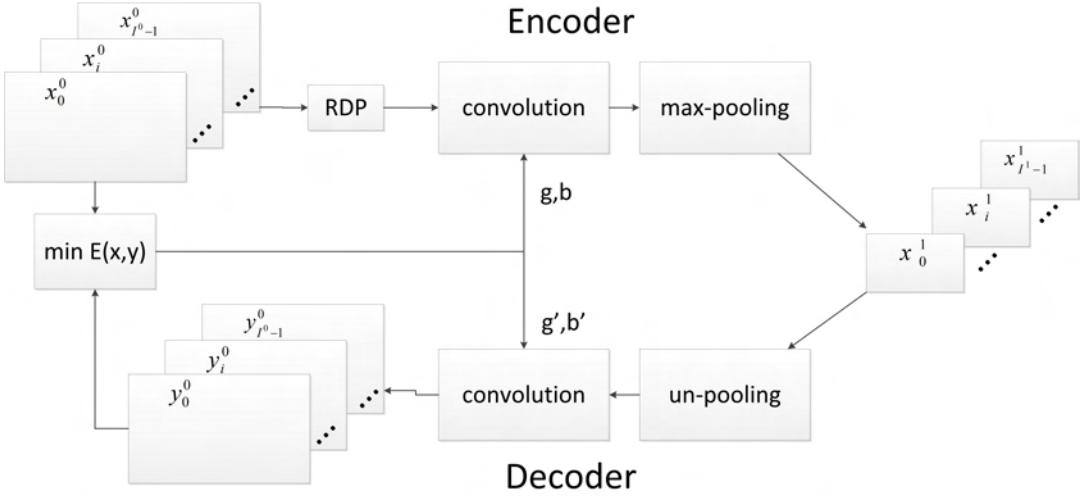


Figure 2: A layer of the Convolutional autoencoder.

the total gradient with respect to $g^m_{i,k}$, and it is applied many times to the input in the feature map extraction process, then the total gradient is obtained considering each individual gradient.

2.2. Convolutional Autoencoder

The Convolutional Autoencoder (CAE) is introduced by Masci et al. (2011), where the model is used to unsupervised feature extraction in the MNIST database of handwritten digits (a recognition task from a big image dataset)(LeCun & Cortes, 2010).

The goal of CAE is the same as that of denoising Auto-Encoder proposed by Vincent et al. (2008), but applied to CNN, to initialize a convolutional layer with the best possible estimates of parameters g and b. This is achieved using a modified autoencoder structure where the encoding process consists of convolution and max-pooling, and the decoding process is composed by un-pooling and convolution, as shown in Figure 2. Max-pooling is a destructive non-reversible operation. Then, the un-pooling operation only consists by horizontal and vertical replication of each (x, y) element using the same max-pooling size factor in all resulting feature maps of encoding process.

The training process for a CAE can be summarized in minimizing the error between $\mathbf{x}$ and $\mathbf{y}$. In this case, we can consider the objective function as:

$$E = ||\mathbf{x} - \mathbf{y}||^2 \tag{3}$$

where $\mathbf{x}$ is the input feature maps. Note that $\mathbf{y}$ is the output of the CAE model with parameters: $\mathbf{W}$, $\mathbf{W}'$ representing the g function, $\mathbf{b}$ and $\mathbf{b}'$ from bias vector. The optimization problem is summarized as:

$$\min_{\mathbf{W},\mathbf{W}',\mathbf{b},\mathbf{b}'} E(\mathbf{W}, \mathbf{W}', \mathbf{b}, \mathbf{b}') \tag{4}$$

From a computational point of view, it can be beneficial to consider $\mathbf{W}' = \mathbf{W}^T$, and the optimization problem result in:

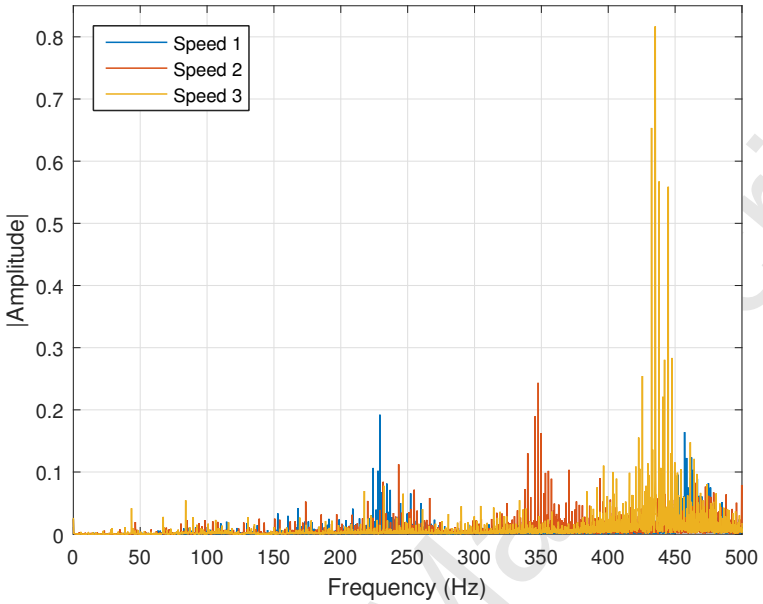
$$\min_{\mathbf{W},\mathbf{b},\mathbf{b}'} E(\mathbf{W}, \mathbf{W}^T, \mathbf{b}, \mathbf{b}') \tag{5}$$

3. Methodology

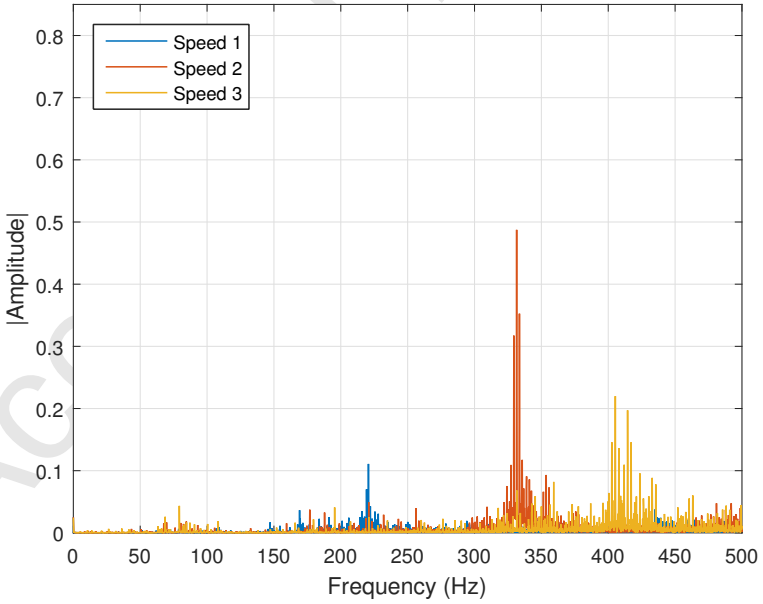
The pattern discovery problem in machinery composed by a helical gearbox has difficulties inherent to dependency on acquired signals with the operational parameters. This is illustrated in Figure 3 with stationary different operational parameters, where the frequency spectrum of an acquired vibration signal shows changes in position and shape at different speeds. Moreover, harmonics at Speed 1 could be easily confused with fundamental frequency at Speed 3. It happens when the operational speeds are multiples of each other. While it is true that the spectrum positions at different loads are almost the same, it is appreciated that the shapes change, although the speeds are the same as in all cases (see Figure 3(a) vs Figure 3(b)).

Under non-stationary operational parameters, pattern identification is even more difficult. Figure 4 illustrates the spectrum at one variable speed and different constant loads. For this case depending on the load, the patterns in frequency domain are not clearly identifiable.

To address the problem of pattern identification in fault diagnosis of machinery composed by gearbox, two main approaches have been used. The first one



(a) Test to Load 1



(b) Test to Load 2

Figure 3: Frequency spectrum of signal vibration acquired from a helical gearbox at 3 different constant speeds.

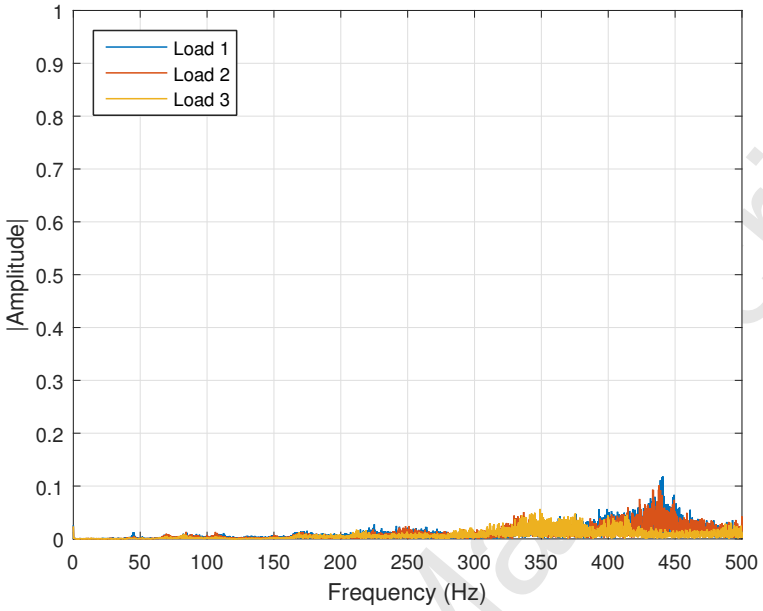


Figure 4: Frequency spectrum of signal vibration acquired from a helical gearbox to variable speed at 3 different constant loads.

is based on signal analysis and processing methodologies, that has as purpose identifying characteristic frequency bands in the spectrum of vibration signals (signatures) (Feng & Zuo, 2013), for the stationary case where the band location typically indicates the fault type and its amplitude shows the fault severity. In the non-stationary case, ad-hoc filtering based techniques has been applied to extract a more informative signal to be analyzed (Li & Liang, 2012).

The second approach is based on pattern classification of classically extracted features from vibration signals in time domain, frequency domain, and/or time-frequency domain as Figure 5 shows. Each feature is carefully designed depending on the case study to extract only the most robust, invariant to changes in the process and informative, condition parameters. After the feature extraction process, classical shallow learning models are used to the pattern classification process, as in Cabrera et al. (2015) and Pacheco et al. (2016). Some deep learning models have also been applied (see, for example, Li et al. (2015)), but in these cases the proposal works on classical pattern classification of previously

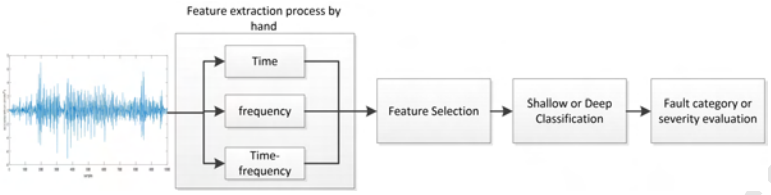


Figure 5: Classical methodology for fault diagnosis with conventional feature extraction.

extracted features.

As we noted before, feature extraction and selection are not trivial tasks and are highly dependent on the operational parameters. Due to its complexity, in this work we propose a method for automatic feature extraction to severity assessment of time-series (vibration signals) extracted from a machinery under stationary and non-stationary operational conditions. The proposed method is presented in the Figure 6, where it is composed by the following steps:

1. **Signal Acquisition:** Acquire vibration signals under multiple fault severity values with stationary and non-stationary operational conditions.
2. **Time-Frequency Representation:** Transform the vibration signals in time to their equivalent time-frequency representation.
3. **Automatic Feature Extraction:** Construct the DCNN model initialized by SCAE, to feature extraction using the time-frequency representation of the signals.
4. **Classification and Regression:** With the extracted features, build the classification model to estimate the discrete severity level, using the extracted features. Optionally, add a regression layer of weighted sum to obtain the continuous severity value.
5. **Online Fault Severity Evaluation:** Evaluate the fault severity assessment of a new vibration signal using the obtained models. The output can be the categorical severity class or the severity value in percentage.

These stages are detailed in next subsections.

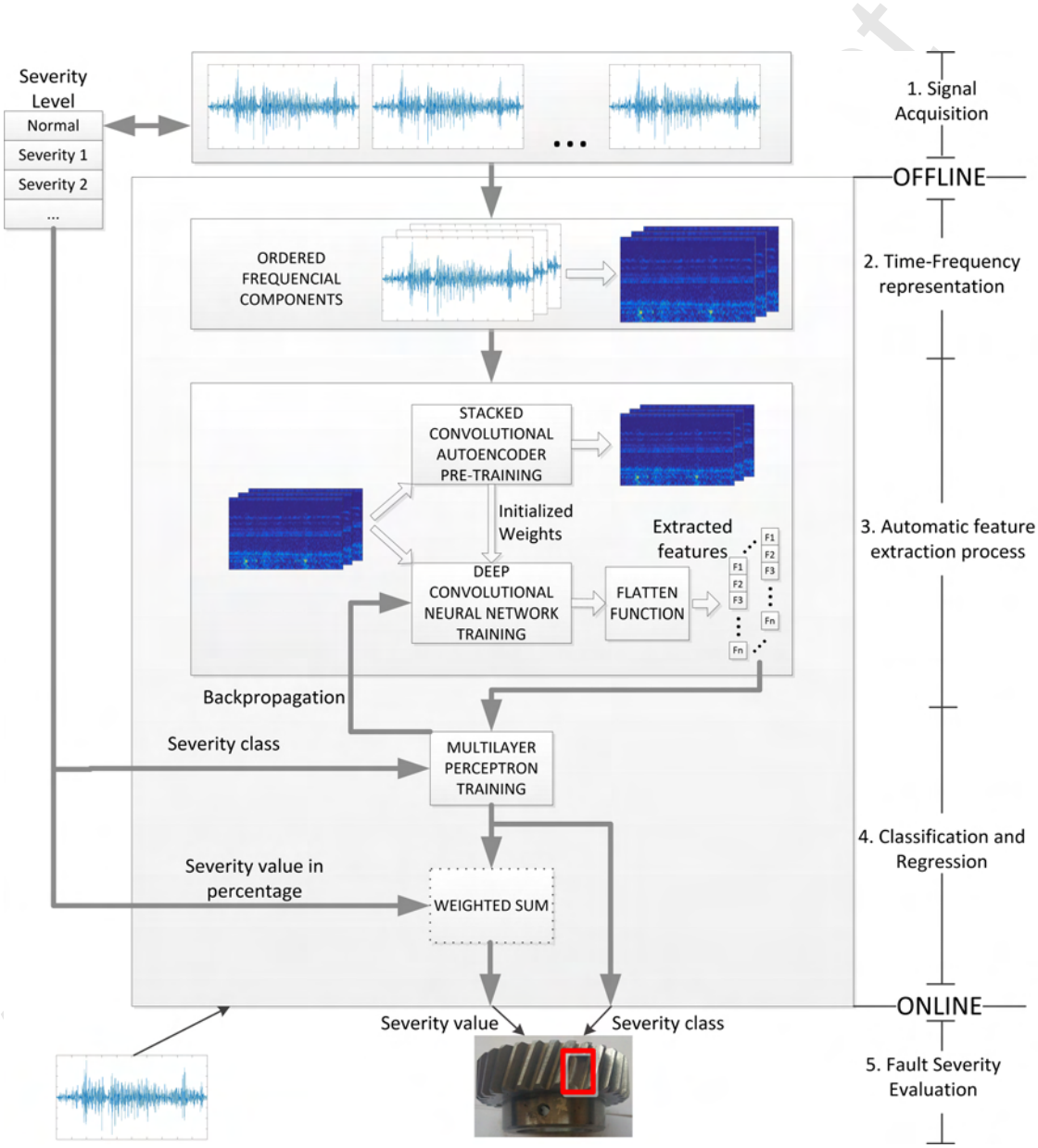


Figure 6: Proposed method for fault severity assessment

3.1. Signal Acquisition

The vibration signals are obtained by direct measuring of this variable from the machinery and its discretization through a Analog to Digital process. Each vibration signal is normalized to $[0, 1]$ interval and grouped with other vibration signals acquired under either different or no-stationary speed and load operational parameters.

3.2. Time-frequency representation

Time-frequency domain signals can be obtained using several tools: Short Fourier Transform, Fractional Fourier Transform, Wavelet Packet Transform (WPT), etc. We will use WPT because it offers adaptive resolution adjust (multi-resolution) in time and frequency. In our experiments we will work with no periodic signals due to variable speed conditions in the input.

WPT is obtained by recursive decomposition of the input signal. This decomposition is carried out by two parallel processes:

$$A = s * g \tag{6}$$

$$D = s * h \tag{7}$$

where s is the input signal, g is the impulse response of ϕ wavelet father function which is characteristic of low pass filter, h is the impulse response of ψ wavelet mother function which is characteristic of high pass filter, and A and D are the resulting signals of the decomposition process (called Approximation and Details signals, respectively). The two filters g and h are restricted to be mutually complementary. Wavelet functions can be any function with finite energy.

A and D signals are decomposed recursively in order to achieve a desired level of decomposition, obtaining a binary tree decomposition of signals as Figure 7 shows.

The last level of nodes has all the information in time and frequency domains of the input signal, and each node at this level represents a specific range in the spectrum. Signals inside these nodes can be normalized in time scale and after

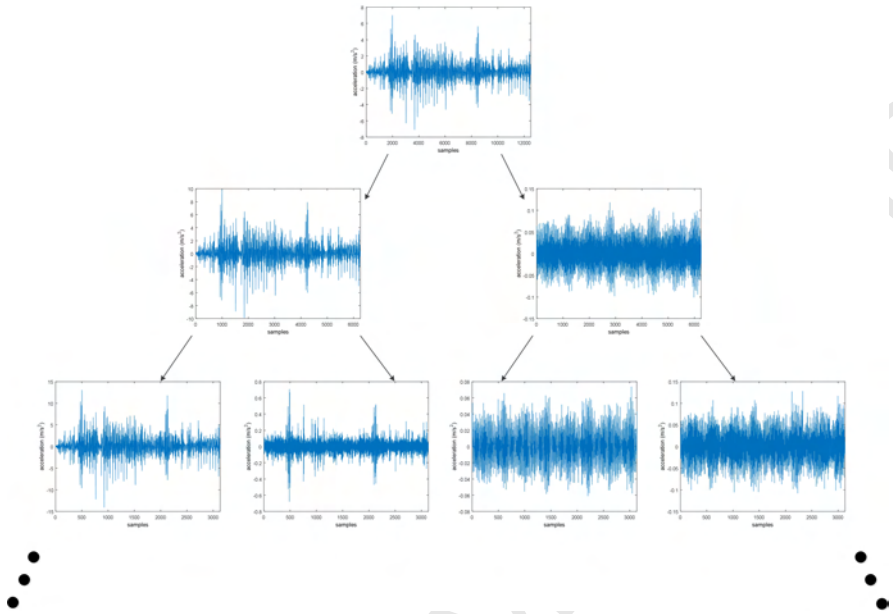


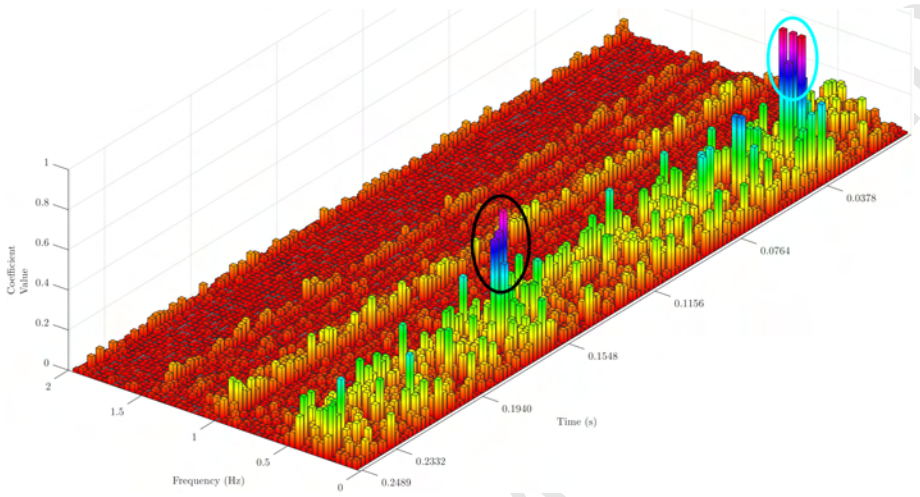
Figure 7: Wavelet Packet Decomposition tree of a time-domain vibration signal.

stack them from lower frequency to the higher one, we obtain the time-frequency representation showed in Figure 8(a) and its equivalent 2D in the Figure 8(b). As an example of possible patterns for gear fault severity assessment, the coefficients with the highest values are marked. In a real world case, the patterns will be much more complex and more difficult to be identified.

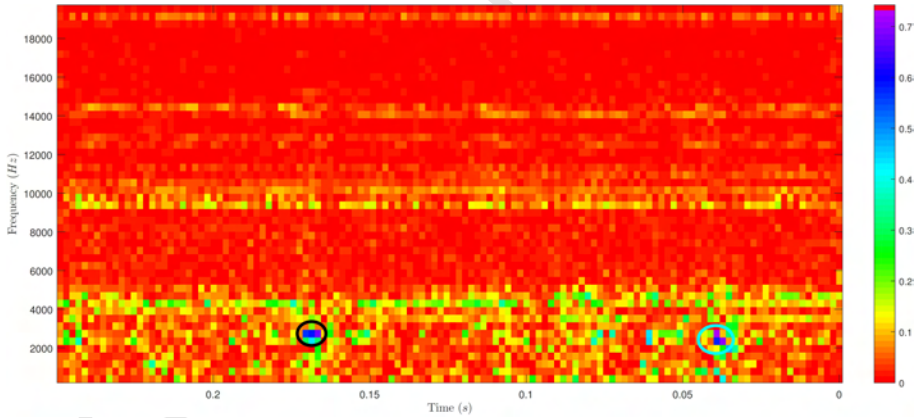
In the works introduced by Cerrada et al. (2016) and Cerrada et al. (2015) a genetic algorithm has been used to evaluate the importance of each wavelet family in gearbox fault diagnosis. Their results show daubechies family as the best choice for this task, and we will use it to obtain the time-frequency plane for the signals in our experiments.

3.3. Automatic Feature Extraction Process

The features in the time-frequency representation of a vibration signal exhibit highly moving patterns due to the lack of an external synchronization signal per turn of motor. This characteristic requires the model to recognize features in any location of the time-frequency representation.



(a) 3D time-frequency representation



(b) Equivalent 2D time-frequency representation

Figure 8: Time-frequency representation of vibration signal using db5 wavelet. Figure 8(a) illustrate the stacked coefficient signals, where the amplitude of each bar represent the value of coefficient at specific time and frequency. Figure 8(b) present the 2D version where the color indicate the value of specific coefficient. The ovals mark equivalent patterns in 3D and 2D visualization with the colors according to its position.

Deep Convolutional Neural Network (DCNN) model works for detecting patterns locally related to others in a bi-dimensional input, regardless the zone where these events occur. Therefore DCNN is robust to shifts and could be a good candidate as feature extraction model from patterns with local occurrence in time-frequency representation. But, since it has to be trained by Stochastic Gradient Descent from randomly initialized parameters, it could suffer from premature convergence to a local minimum.

The weakness of DCNN model can be addressed with a better method for initializing parameters. We propose to use Stacked Convolutional Autoencoders (SCAE) architecture which is composed by CAE stacked, similar to Stacked denoising Autoencoder (SdAE) architecture, in order to extract *a priori* knowledge regarding the model capacity to reconstruct the time-frequency plane of the input. Later the parameters obtained with SCAE are used as initial point for fine-tuning of the DCNN.

Each element in the set of feature maps at the output of DCNN is passed through a flatten function and stacked to construct the vector of extracted features. The size of this vector is dependent on the input size and the DCNN architecture, where the output of each layer are k feature maps of size $(\frac{M-m+1}{2} \times \frac{N-n+1}{2})$, given input feature maps of size $(M \times N)$ and k kernels of size $(m \times n)$.

3.4. Classification and Regression

The vector of features obtained with the automatic feature extraction process from time-frequency representation is used as input into a multi-layer perceptron (MLP), built to evaluate the severity fault. Taking into account that commonly the decision space created by the vector of features is high-dimensional, the number of neurons in the hidden layer of the MLP should be large enough. The number of outputs is equal to the number of available severity values established in the training phase.

The classification error is computed using the correct level severity class and the output of the DCNN. Then the error is back-propagated from the output through hidden layer to DCNN to complete the cycle of fine-tuning in

the automatic feature extraction process. Thereby the features are adjusted to obtain a better accuracy according to the prediction of the severity class. Additionally, a regression layer composed by weighted sum operation (linear neuron) of the inputs could be added to obtain a real valued output. It could be done by using the discretization layer after the logistic regressor or not. In the first case, the predicted severity class is directly weighted by its equivalent percentage of damage. In the second case, each probability of the severity class is weighted by the corresponding value of severity in percentage and summing with others to give a real value. Both cases were tested in our experiments.

3.5. Online Fault Severity Evaluation

After the models have been adjusted following the previous method, on-line tests could be performed. A new vibration signal acquired from the machinery is entered into the black-box method without changes in the internal parameters. The process can generate two outputs: The first one with a categorical severity class and the second one with the percentage of damage present in the evaluated gear of the machinery.

4. Experiments

In this section we present the set of experiments carried out in helical gearboxes to obtain the vibration signals. Also we present the different configurations used by the proposed method.

4.1. Experimental Setup

The experiments were performed following the diagram displayed in Figure 9. The one stage helical gearbox (GB) is composed by two gears (Z1 and Z2) mounted in independent shafts. The input shaft is connected to a motor, that transforms electrical energy to rotational movement in order to be transmitted to the mechanical system. The output shaft is linked to the break system (B), which has a belt connected to a magnetic break, and transforms the electrical energy into mechanical force opposed to the rotational movement of the output

Component	Description
M	Motor Siemens 1LA7 090-4YA60 1.49 kW, 4 poles, 28.33 Hz
Z1	Pinion, 76 mm, 30 tooth, pressure angle = 20°, helix angle = 20°
Z2	Gear, 112 mm, 45 tooth, pressure angle = 20°, helix angle = 20°
B	Magnetic break, proportional force to input voltage, belt coupled

Table 1: Data of mechanical system components

Component	Description
Speed drive	Danfoss VLT 1.5 kW
Source power	TDK Lambda GEN 150-10, 0 V-150 V, 10 A

Table 2: Data of source system components

shaft. The most important information about these mechanical components is extended in Table 1.

The motor is driven by a speed drive configured as follows: it uses the mode control of estimate speed, 1 s ramp time from min to max speed, automatic identification of the motor parameters, and set point configured in analog voltage input mode. The magnetic break is driven by a variable voltage source power. The technical descriptions of these components are shown in Table 2.

Two signals were obtained from the system. The first one is the vibration acquired from an unidirectional accelerometer device transforming the mechanical acceleration in one axis to an analog voltage signal. Since the contact between gears mainly occurs vertically the accelerometer is mounted in this direction. The second signal is the angular position of the motor, acquired from an encoder device giving a pulse each 0.036°. The technical descriptions of these components are shown in Table 3.

The analog signals are digitalized using Data Acquisition (DAQ) cards. We use a NI9234 card (from National Instrument) for the accelerometer signal measure, specific for piezo-electric sensors. It is configured in pseudo-differential

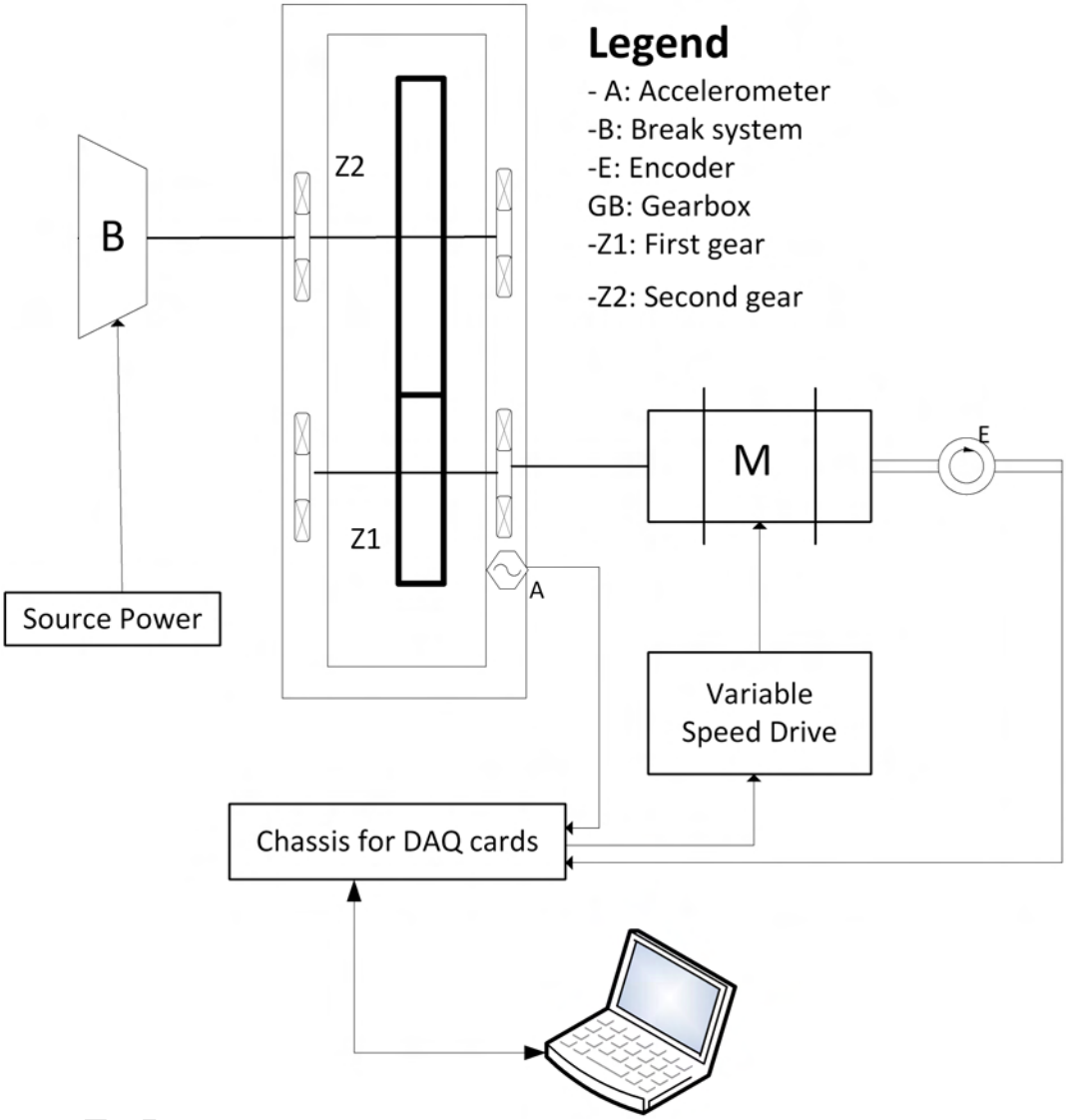


Figure 9: Block diagram of the experimental setup

Component	Description
A	Accelerometer unidirectional IMI Sensor 603C01, 100 mV g^{-1}
E	Encoder SICK DFS60-S4PL10000, 10 000 ppr

Table 3: Data of acquisition system components

mode to noise immunity measure, integrates source power at the same cable, has 24 bits of resolution at 50 ksamples/s which permits to detecting fault characteristic frequencies up to 25 kHz. For the encoder pulses, we use a NI9401 card, with high speed digital inputs, and it is configured to acquire pulses at 1 Msample/s, which is able to detect up to 50 Hz of motor rotation speed. The speed drive and power source of the break system are connected to analog outputs of other DAQ card.

All previous DAQ cards are plug in a NI cDAQ-9188 chassis, that permits the temporal synchronization between all signals from different modules with the subsequent real time measure and control. The chassis groups all signals and sends them to a buffer located in the RAM of a laptop (Figure 9). A program developed in Labview takes the buffer data and stores them to the hard drive. On the other hand, data from encoder is used to adjust the real speed and compensate the rotation speed. The source power voltage of B works in open loop, hence the voltage is used to quantify the force opposed to the movement of the gearbox.

4.2. Experimental plan

Based on the previous setup, an experimental plan was designed using the controlled variables in the entire system. The main goal in this plan was the assessment of severity in helical gear tooth breakage fault, and for this purpose different damages were produced by cutting the tooth of Z1 gear at 9 levels while Z2 is not modified. Table 4 presents a summary of the damage configurations and Figure 10 shows the physical results.

The rotation speed was considered in 5 variants. Three of them were at constant speed and two more at variable speed, changing in sine and square profiles. The periods of these profiles were constant and repeated until the end of the experiment. Table 5 summarizes the different speed profiles.

The load had three constant levels. Each level was achieved by varying the voltage supplied to the braking system. They are coded with L1, L2, L3 and represent 0 V (without load), 10 V and 30 V, respectively.

Code	Description	Damage (mm)	Percentage of tooth (%)
P1	Level 1 or 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 or without tooth	20.43	0.0

Table 4: Damage levels of helical gear tooth breakage fault

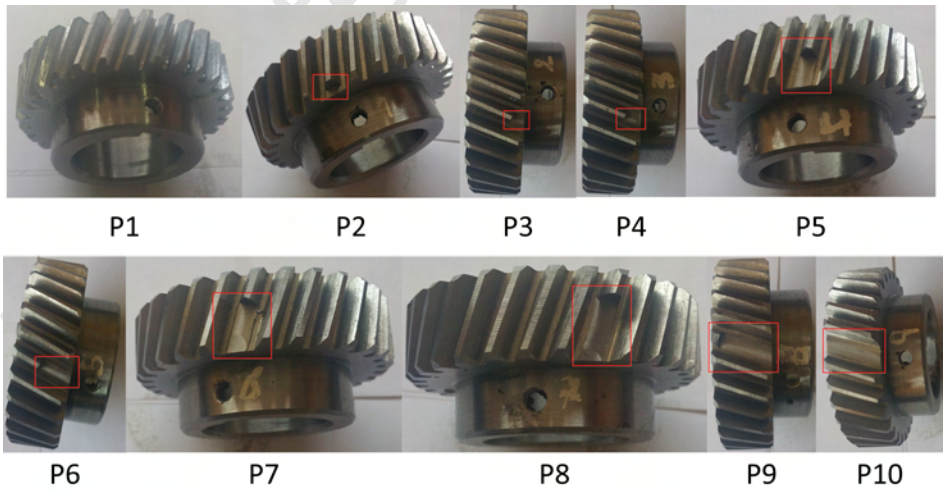


Figure 10: Damage in helical gear tooth according to code from Table 4

Code	Type	Rotation Frequency (Hz)	Profile
F1	Constant	8	-
F2	Constant	12	-
F3	Constant	15	-
F4	Variable	8-15	sine-period=2 s
F5	Variable	8-15	square-period=2 s

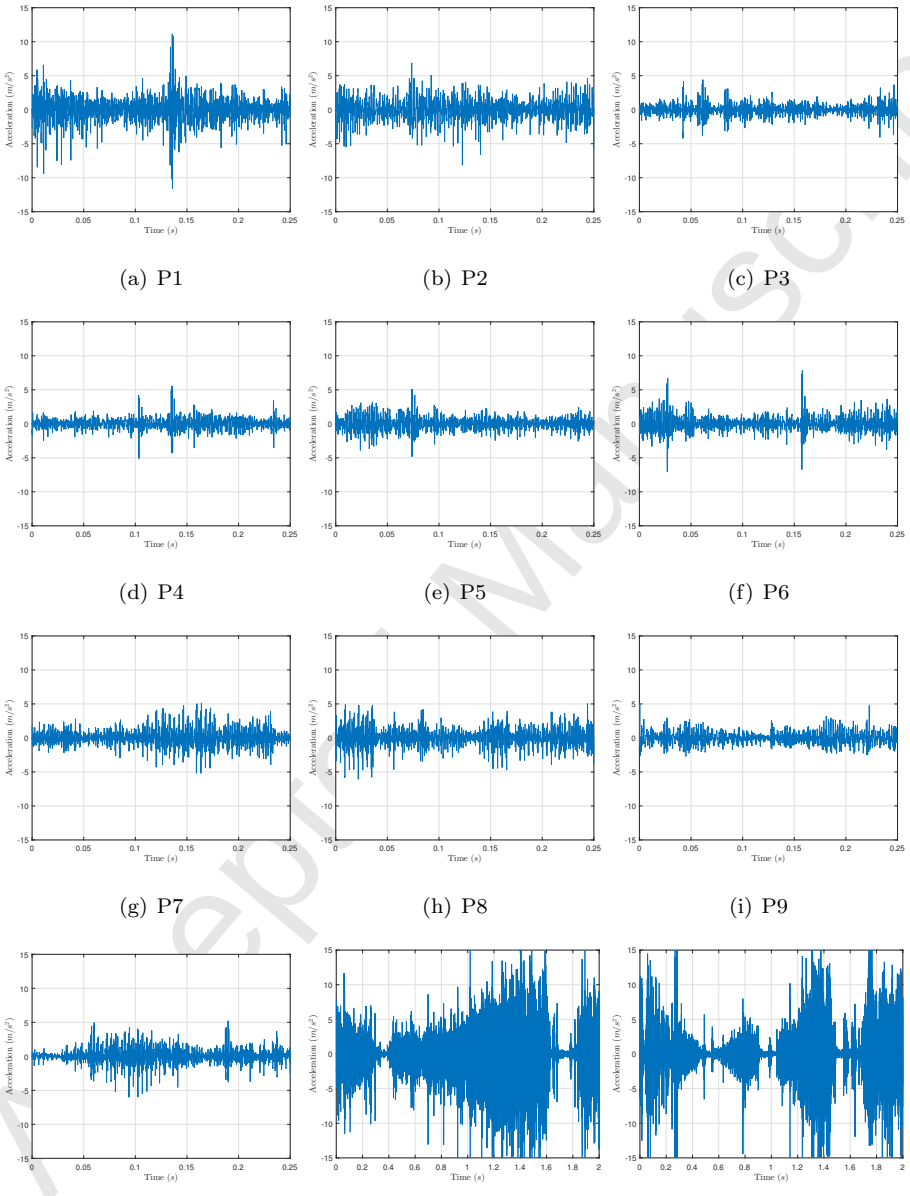
Table 5: Constant and variable speed profiles

For each level of severity P, each speed F was evaluated in each load L. Hence a set of experiments under all combinations of P, F and L could be achieved. For each element of this set, 5 repetitions of 10 s were performed, obtaining an experimental dataset of $10 \cdot 5 \cdot 3 \cdot 5 = 750$ signals, (a subsample of acquired signals is shown in the Figure 11). Then, each signal of this dataset was normalized to the $[0, 1]$ interval and it was cut into 40 sub-signals using a rectangular window of 0.25 s with no overlapping, resulting in a new dataset of 30000 elements. These sub-signals are equivalent to vibrations during 2 turns of the input shaft at the lower rotation frequency of 8 Hz. The time series dataset was transformed into time-frequency dataset using WPT with wavelet mother daubechies 5 at decomposition level 6. As a result, a time-frequency image of 136×64 pixels was obtained for every element of the dataset.

After data acquisition, training and testing of SCAE network was performed, where the dataset was divided into three subsets: training (60%), validation (20%) and testing (20%). In order to obtain the same percentage of elements of every severity level in each subset, a process of stratification was made. The validation subset was used as a method for avoiding overfitting and guarantee the results.

5. Results and Discussions

Three groups of tests were performed for specific purposes. The first group allows to compare between DCNN with randomly initialized kernels and the



(a) P1 (b) P2 (c) P3

(d) P4 (e) P5 (f) P6

(g) P7 (h) P8 (i) P9

(j) P10 (k) F4 (l) F5

Figure 11: Examples of signal vibration. From Figure 11(a) to Figure 11(j) present a signal acquired at each severity level (P1-P10) with F1 speed and L1 load. Figures 11(k) and 11(l) show a signal at P1 severity level, L1 load, F4 and F5 speed, respectively.

proposed DCNN with initialized kernels using SCAE, evaluating the advantage of starting the training process from a pre-optimized solution. Then the best DCNN with SCAE was evaluated in each severity level to analyze if the majority of the errors occur between adjacent classes.

The second group was performed to evaluate the incidence of using the linear regressor with the discretization layer after the logistic regressor and without it. The goal was to measure if the set of probabilities gives information in concordance with the real valued level severity and if each probability is close to the others which is proved with the MSE and MedAE metrics.

The third group provided a comparison with other methods, supervised and unsupervised, that have been previously applied to fault diagnosis cases. The selected methods use classical techniques of feature extraction without the conventional design of features for obtaining a comparison between methods looking for the same goal: the automation in the process of feature extraction from time series. In the next subsections the details about these tests are summarizes.

5.1. Comparison of DCNN vs SCAE-DCNN

The first tests were performed using DCNN with random initialization of parameters, uniformly sampled from a bounded range. The architectures with 2 and 3 convolutional layers were tested. The common factors in architectures and training parameters are:

- Presence of a max-pooling layer in each convolutional layer.
- Unique hidden layer fully connected to last convolutional layer with 1000 neurons.
- Final layer composed of logistic regressors.
- Mini-batch size of 200 elements.
- Learning rate for the supervised training at 0.1.
- Training epochs set at 400.

Code	kernels layer 1	kernels layer 2	kernels layer 3
CNN1	100 × (5,5)	200 × (3,3)	-
CNN2	100 × (5,5)	40 × (5,5)	-
CNN3	100 × (5,5)	200 × (3,3)	400 × (3,3)

Table 6: Configuration of CNNs with number of kernels and shape of each kernel

Code	kernels layer 1	kernels layer 2	kernels layer 3
SCAE1	100 × (5,5)	200 × (3,3)	-
SCAE2	100 × (5,5)	40 × (5,5)	-
SCAE3	100 × (5,5)	200 × (3,3)	400 × (3,3)

Table 7: Configuration of SCAEs with number of kernels and shape of each kernel

The different configurations (changing the number of kernels in each convolutional layer) are summarized in Table 6.

Furthermore, other tests using the same architecture and training parameters were performed with SCAEs. In these new tests, the pretraining stage was added with 0.1, 0.2, and 0.3 of corruption level in layer 1, layer 2 and layer 3, respectively; and corruption level of 0.3 was used in the fully connected layer in all cases. The pretraining learning rate was 0.01 and the number of epochs adjusted to 100. Table 7 summarizes the architecture of each SCAE.

We used the accuracy, which is a classical metric to quantify the global performance of classifiers. In Table 8, CNN1 y CNN2 compared with SCAE1 y SCAE2 show a little difference of 0.0018 between CNN1 and SCAE1, and 0.0012 between CNN2 and SCAE2. Both show a little better performance in CNNs than SCAEs but could be considered negligible. Moreover, SCAE3 shows a significant increase of 0.0223 in performance compared to CNN3, which have the same architecture, and an increase of 0.0083 compared to CNN2, which is the best of all CNNs.

SCAE3, which presents the best accuracy, is evaluated in its performance

Code	CNN	SCAE
1	0.9208	0.919
2	0.9382	0.937
3	0.9242	0.9465

Table 8: Accuracy of each classifier over test set.

code severity	precision	recall	f-measure
P1	0.9680	0.9583	0.9631
P2	0.9533	0.9533	0.9533
P3	0.9538	0.9283	0.9409
P4	0.9400	0.9400	0.9400
P5	0.9218	0.9233	0.9226
P6	0.9103	0.9133	0.9118
P7	0.9422	0.9233	0.9327
P8	0.9532	0.9850	0.9689
P9	0.9635	0.9667	0.9651
P10	0.9589	0.9733	0.9661

Table 9: Metrics of performance in each level of severity

with each severity level using Precision, Recall and f-measure, classical metrics, as Table 9 shows. All metrics are calculated from the data given by the confusion matrix showed in Table 10. The confusion matrix is obtained from the test set with 600 samples for each level of severity.

The best precision, which indicates the capacity of the model to avoid the inclusion of samples from any other classes in the analyzed class, is obtained for P1 (normal condition). It shows that signatures P2 to P10 have the best characteristics of separability with respect to P1. The worst case is obtained with the confusion of signatures P3 (10 samples), P5 (11 samples), and P7 (14 samples) with P6.

Class	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	575	6	3	1	5	1	1	3	0	5
P2	6	572	1	2	3	5	0	4	2	5
P3	2	3	557	5	14	10	2	3	2	2
P4	0	4	3	564	7	8	7	3	4	0
P5	3	2	11	5	554	11	6	1	3	4
P6	4	6	6	5	7	548	11	4	6	3
P7	3	0	1	12	3	14	554	6	4	3
P8	0	2	0	0	0	2	2	591	0	3
P9	0	5	0	6	4	1	3	1	580	0
P10	1	0	2	0	4	2	2	4	1	584

Table 10: Confusion matrix obtained from evaluation of model with the test set

On other hand, the best recall, which shows the capacity of the model to include all the samples that, in fact, are inside a class, is obtained for P8. It shows that signature P8 is distinguished from other signatures. Newly, the worst case is obtained for P6, which shows the major error with P5 (7 samples) and P7 (11 samples). This is justified by the closeness in the severity level of P6 to those from P5 and P7.

Additionally, f-measure shows the general performance of model, considering both precision and recall. In this case, the best f-measure is obtained for P8 and the worst for P6.

5.2. Incidence of discretization layer

This test is performed with discretization layer after logistic regressor and without it. The results are shown in Table 11. These results show a better general performance of SCAE without discrete layer with a lower Mean Squared Error (MSE) that the other case.

The results indicate that the output probability of the logistic regressors gives information related to the fault severity. There is a large difference of

Code	MSE	MedAE
SCAE3 with discrete layer	$7.985 \cdot 10^{-3}$	$5.332 \cdot 10^{-9}$
SCAE3 without discrete layer	$5.853 \cdot 10^{-3}$	$4.073 \cdot 10^{-5}$

Table 11: Quantitative performance with, and without, discrete layer

values between the correct class and the others (low MSE and low MedAE), obtaining a great confidence in the response.

5.3. Comparison with others feature extraction methods

In order to compare the proposed method with others, two comparison blocks are performed. The first one uses an unsupervised feature extraction methods without knowledge of the specific task to obtain the features from data neither raw data nor otherwise. The second one uses a supervised feature extraction and selection method together with a shallow classifier as a representative of a classical approach (see Figure 5).

To minimize comparison bias resulting from the random selection of data in train and test sets, a 5-folds cross-validation is performed for each type of method, i.e., for both the unsupervised method comparison and for the supervised method comparison.

5.3.1. Unsupervised feature extraction methods

The comparison is performed using the accuracy (acc) metric between the proposed method with 5 additional methods:

1. T-PCA-MLP(Mdlazi et al., 2007): Principal Components Analysis (PCA) technique is applied to vibration signal, to select the number of principal components the Maximum Likelihood Estimation of Minka is used obtaining 765 features. The selected features are the inputs to a MLP with the same structure used in SCAE3.
2. F-MLP(ZHOU et al., 2015): Fast Fourier Transform (FFT) is applied to vibration signal, all spectrum components are the inputs (4097 features) to a MLP with the same structure used in SCAE3.

3. F-PCA-MLP(Mdlazi et al., 2007): Fast Fourier Transform (FFT) is applied to vibration signal, then PCA technique is applied to spectrum, to select the number principal components the Maximum Likelihood Estimation of Minka is used obtaining 780 features. The selected features are the inputs to a MLP with the same structure used in SCAE3.
4. TF-MLP: The time-frequency representation proposed in this work is obtained from vibration signal, it is flatten and the resulting vector of 8704 features is the input to MLP with the same structure used in SCAE3.
5. TF-PCA-MLP: The time-frequency representation proposed in this work is obtained from vibration signal, it is flatten and the resulting vector PCA is applied, to select the number principal components the Maximum Likelihood Estimation of Minka is used obtaining 754 features. The selected features are the inputs to a MLP with the same structure used in SCAE3.
6. TF-SVM: In order to compare the results of our proposal with a method using other classifier, the time-frequency representation proposed in this work is obtained from vibration signal, it is flatten and the resulting vector of 8704 features is the input to multi-class one-vs-rest Support Vector Machine (SVM) with Gaussian Kernel.

The results are presented in the Table 12. They show that methods using traditional feature extraction techniques cannot extract informative features for the present task. This is due to, firstly by high variability in the signal vibration with redundant information, but mainly to the combination of stationary and non-stationary signals from multiples operational parameter, that hide the important features. These techniques have been applied to stationary or minimum non-stationary condition separately with good results, but they fail in more complex scenarios.

5.3.2. Classical feature extraction method

An additional comparison is performed using the classical approach to fault diagnosis based on machine learning. The vibration signals are processed to extract statistical features proposed by Pacheco et al. (2016), in time, frequency

Code	fold 1	fold 2	fold 3	fold 4	fold 5	avg. acc
SCAE3	95.02	95.05	95	94.22	95.53	94.96 ± 0.47
T-PCA-MLP	24.1	26.22	17.46	20.56	13.78	20.42 ± 5
F-MLP	10.08	10.5	10	10.71	10.42	10.34 ± 0.3
F-PCA-MLP	17.67	20.01	18.52	15.05	16.45	17.54 ± 1.91
TF-MLP	15.4	13.56	15.63	17.2	14.28	15.21 ± 1.39
TF-PCA-MLP	24.3	26.12	25.47	22.5	26.23	24.92 ± 1.56
TF-SVM	15.02	15.13	14.58	14.63	14.87	14.85 ± 0.24

Table 12: Comparison of SCAE with other unsupervised feature extraction methods. Each fold column has the accuracy with the corresponding fold. The last column has the average accuracy with the respective standard deviation.

and time-frequency domain obtaining a set of 817 condition indicators. Subsequently a process to filter the correlated indicators is applied obtaining a set of 183 non-correlated features to which a zero mean and one standard deviation normalization is applied. Later, 16 features are selected using RELIEF as the best feature selection technique together a multi-class one-vs-rest SVM with Gaussian Kernel classifier proposed by Vakharia et al. (2015) to make the classification of the severity level.

The results are shown in the Table 13. Its has evidence of a better performance in the estimation of severity assessment using the designed features by experts in complement with a process of feature selection and classification. However in comparison with the proposed method, Rel+SVM has a poor accuracy of 6.69% minus. Also it is possible to note that there is a little variation of only 0.47% in the results with our method in front of 2.56% with Rel+SVM, which shown that performance of our proposal is not dependent of the data used to train the models.

Code	fold 1	fold 2	fold 3	fold 4	fold 5	avg. acc
SCAE3	95.02	95.05	95	94.22	95.53	94.96 ± 0.47
Rel+SVM	88	89.33	84	90.67	89.33	88.27 ± 2.56

Table 13: Comparison of SCAE with RELIEF+SVM. Each fold column has the accuracy with the corresponding fold. The last column has the average accuracy with the respective standard deviation.

6. Conclusions

A new method of automatic feature extraction has been proposed to fault severity assessment. It consists in a pattern extraction process following an unsupervised approach from a representation in time-frequency domain of time-series signals, that adjusts the convolutional layer to minimize the reconstruction error of inputs. The feature maps obtained from a layer is the input for the next layer, in order to further refine them. As a result of the successive application of this procedure a hierarchical extraction of features and pattern identification are obtained. With this approach, the parameter space is explored starting from a good initial configuration that minimizes the possibility of becoming trapped in a local minimum in a later supervised learning stage.

The case study is the identification of fault severity in helical gearboxes from a vibration signal, where the design and extraction of condition parameters is a non-trivial task when performed by the conventional classical methods. Moreover, the results obtained by the latter methods are hardly applicable to other real-world systems as the features containing the most representative information are highly dependent on the specific mechanism.

The experimental results for this case study show relevant aspects for fault severity assessment:

1. Time-frequency representation of vibration signals retains the information of the original time-series with respect to level of damage, and shows the relevant information more clearly.

2. The use of hierarchical architectures with convolutional layers allows extracting successfully features and detecting patterns from time-varying inputs.
3. The use of Stacked Convolutional Autoencoder approach improves the accuracy of level severity assessment model with respect to Convolutional Neural Network without a pre-training stage.

Acknowledgments

The authors want to thank to R&D projects TIN2012-37434 and TIN2013-41086-P supported by Ministerio de Economía y Competitividad of Gobierno de España and co-financed by the European FEDER funds by support of this research work. The work was sponsored by the GIDTEC project No. 002-002-2016-03-03 supported by Universidad Politécnica Salesiana sede Cuenca, and the Prometeo Project of the Secretariat for Higher Education, Science, Technology and Innovation (SENESCYT) of the Republic of Ecuador. The experimental work was developed at the GIDTEC research group lab of the Universidad Politécnica Salesiana sede Cuenca, Ecuador.

References

Cabrera, D., Sancho, F., Sánchez, R.-V., Zurita, G., Cerrada, M., Li, C., & Vásquez, R. (2015). Fault diagnosis of spur gearbox based on random forest and wavelet packet decomposition. *Frontiers of Mechanical Engineering*, (pp. 1–10). URL: http://dx.doi.org/10.1007/s11465-015-0348-8. doi:10.1007/s11465-015-0348-8.

Cerrada, M., Sánchez, R. V., Cabrera, D., Zurita, G., & Li, C. (2015). Multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal. *Sensors*, *15*, 23903. URL: http://www.mdpi.com/1424-8220/15/9/23903. doi:10.3390/s150923903.

Cerrada, M., Zurita, G., Cabrera, D., Sánchez, R.-V., Artés, M., & Li, C. (2016). Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mechanical Systems and Signal Processing*, *70-71*, 87–103. URL: http://dx.doi.org/10.1016/j.ymssp.2015.08.030. doi:10.1016/j.ymssp.2015.08.030.

Chen, F., Tang, B., & Chen, R. (2013). A novel fault diagnosis model for gearbox based on wavelet support vector machine with immune genetic algorithm. *Measurement*, *46*, 220232. URL: http://dx.doi.org/10.1016/j.measurement.2012.06.009. doi:10.1016/j.measurement.2012.06.009.

Chinniah, Y. (2015). Analysis and prevention of serious and fatal accidents related to moving parts of machinery. *Safety Science*, *75*, 163173. URL: http://dx.doi.org/10.1016/j.ssci.2015.02.004. doi:10.1016/j.ssci.2015.02.004.

Fan, X., & Zuo, M. J. (2006). Gearbox fault detection using Hilbert and wavelet packet transform. *Mechanical Systems and Signal Processing*, *20*, 966982. URL: http://dx.doi.org/10.1016/j.ymssp.2005.08.032. doi:10.1016/j.ymssp.2005.08.032.

Feng, Z., & Zuo, M. J. (2013). Fault diagnosis of planetary gearboxes via torsional vibration signal analysis. *Mechanical Systems and Signal Processing*, *36*, 401–421. URL: http://dx.doi.org/10.1016/j.ymssp.2012.11.004. doi:10.1016/j.ymssp.2012.11.004.

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*, 14831510. URL: http://dx.doi.org/10.1016/j.ymssp.2005.09.012. doi:10.1016/j.ymssp.2005.09.012.

Kateris, D., Moshou, D., Pantazi, X.-E., Gravalos, I., Sawalhi, N., & Loutridis, S. (2014). A machine learning approach for the condition monitoring of rotating machinery. *Journal of Mechanical Science and Technology*, *28*, 61–

71. URL: http://dx.doi.org/10.1007/s12206-013-1102-y. doi:10.1007/s12206-013-1102-y.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, *86*, 2278–2324. URL: http://dx.doi.org/10.1109/5.726791. doi:10.1109/5.726791.

LeCun, Y., & Cortes, C. (2010). MNIST handwritten digit database, . URL: http://yann.lecun.com/exdb/mnist/.

Li, C., & Liang, M. (2012). Time-frequency signal analysis for gearbox fault diagnosis using a generalized synchrosqueezing transform. *Mechanical Systems and Signal Processing*, *26*, 205–217. URL: http://dx.doi.org/10.1016/j.ymssp.2011.07.001. doi:10.1016/j.ymssp.2011.07.001.

Li, C., Sanchez, R.-V., Zurita, G., Cerrada, M., Cabrera, D., & Vásquez, R. E. (2015). Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis. *Neurocomputing*, *168*, 119–127. URL: http://dx.doi.org/10.1016/j.neucom.2015.06.008. doi:10.1016/j.neucom.2015.06.008.

Li, Z., Yan, X., Tian, Z., Yuan, C., Peng, Z., & Li, L. (2013). Blind vibration component separation and nonlinear feature extraction applied to the nonstationary vibration signals for the gearbox multi-fault diagnosis. *Measurement*, *46*, 259–271. URL: http://dx.doi.org/10.1016/j.measurement.2012.06.013. doi:10.1016/j.measurement.2012.06.013.

Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I* ICANN'11 (pp. 52–59). Berlin, Heidelberg: Springer-Verlag. URL: http://dl.acm.org/citation.cfm?id=2029556.2029563.

Mdlazi, L., Marwala, T., Stander, C. J., Scheffer, C., & Heyns, P. S. (2007). Principal component analysis and automatic relevance determination in dam-

age identification. *CoRR*, *abs/0705.1672*. URL: http://dblp.uni-trier.de/db/journals/corr/corr0705.html#abs-0705-1672.

Nie, M., & Wang, L. (2013). Review of condition monitoring and fault diagnosis technologies for wind turbine gearbox. *Procedia CIRP*, *11*, 287290. URL: http://dx.doi.org/10.1016/j.procir.2013.07.018. doi:10.1016/j.procir.2013.07.018.

Pacheco, F., Oliveira, J. V. d., Sánchez, R.-V., 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*, . URL: http://dx.doi.org/10.1016/j.neucom.2016.02.028. doi:10.1016/j.neucom.2016.02.028.

Vakharia, V., Gupta, V. K., & Kankar, P. K. (2015). A comparison of feature ranking techniques for fault diagnosis of ball bearing. *Soft Computing*, *20*, 1601–1619. URL: http://dx.doi.org/10.1007/s00500-015-1608-6. doi:10.1007/s00500-015-1608-6.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In W. W. Cohen, A. McCallum, & S. T. Roweis (Eds.), *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)* (pp. 1096–1103). ACM.

ZHOU, J., QIN, Y., KOU, L., YUWONO, M., & SU, S. (2015). Fault detection of rolling bearing based on FFT and classification. *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, *9*, JAMDSM0056–JAMDSM0056. URL: http://dx.doi.org/10.1299/jamdsm.2015jamdsm0056. doi:10.1299/jamdsm.2015jamdsm0056.

Highlights

- An automatic feature extraction method is proposed.
- Manual feature extraction procedures can be replaced with deep learning approach.
- Valid features can be extracted from a time-frequency representation of time-series.
- A pre-trained layer by layer model is better than a model without pre-training.