Article

Computation and Statistical Analysis of Bearings' Time- and Frequency-Domain Features Enhanced Using Cepstrum Pre-Whitening: A ML- and DL-Based Classification

David Cascales-Fulgencio [1] , Eduardo Quiles-Cucarella [2],* and Emilio García-Moreno [2]

1   Escuela Técnica Superior de Ingeniería Industrial, Universitat Politècnica de València, Camino de Vera, s/n, 46022 Valencia, Spain
2   Instituto de Automática e Informática Industrial, Universitat Politècnica de València, Camino de Vera, s/n, 46022 Valencia, Spain
*   Correspondence: equiles@isa.upv.es

Abstract: Vibration signals captured with an accelerometer carry essential information about Rolling Element Bearings (REBs) faults in rotating machinery, and the envelope spectrum has proven to be a robust tool for their diagnosis at an early stage of development. In this paper, Cepstrum Pre-Whitening (CPW) has been applied to REBs' signals to enhance and extract health-state condition indicators from the preprocessed signals' envelope spectra. These features are used to train some of the state-of-the-art Machine Learning (ML) algorithms, combined with time-domain features such as basic statistics, high-order statistics and impulsive metrics. Before training, these features were ranked according to statistical techniques such as one-way ANOVA and the Kruskal–Wallis test. A Convolutional Neural Network (CNN) has been designed to implement the classification of REBs' signals from a Deep Learning (DL) point of view, receiving raw time signals' greyscale images as inputs. The different ML models have yielded validation accuracies of up to 87.6%, while the CNN yielded accuracy of up to 77.61%, for the entire dataset. In addition, the same models have yielded validation accuracies of up to 97.8%, while the CNN, 90.67%, where signals from REBs with faulty balls have been removed from the dataset, highlighting the difficulty of classifying such faults. Furthermore, from the results of the different ML algorithms compared to those of the CNN, frequency-domain features have proven to be highly relevant condition indicators combined with some time-domain features. These models can be potentially helpful in applications that require early diagnosis of REBs faults, such as wind turbines, vehicle transmissions and industrial machinery.

Keywords: condition monitoring of wind turbines; rolling element bearings; vibration analysis; envelope spectrum; cepstrum pre-whitening; time-domain features; machine learning; deep learning

1. Introduction

One of the clean energy sources that has gained the most momentum in recent years is wind energy, whose global installed capacity grew by 53% year-on-year in 2020 to 743 GW, with more than 93 GW of newly installed capacity [1]. Wind energy is generated by wind turbines that convert the wind's kinetic energy into electrical energy. The geometry of a wind turbine blade is designed to spin the rotor hub when positioned at a certain angle of attack with the wind. The torque is transmitted through a mechanical system consisting of shafts, couplings, REBs and gearboxes to an asynchronous machine specifically designed for this system: a doubly fed asynchronous machine.

The study of REBs failures from the point of view of vibration analysis is a field extensively researched by the international scientific community, and an essential subject of condition monitoring of rotating machines. Kiral & Karagülle developed a method based on finite element vibration analysis to detect single or multiple faults under the action of an imbalanced force on different components of the REB structure using time

and frequency domain features [2]. Sawalhi & Randall conducted an in-depth study on the nature of vibration signals at different stages of rolling elements' impacts with faults with applications in defect size estimation [3]. Smith & Randall thoroughly analysed the Case Western Reserve University (CWRU) signals with their benchmark method, based on pre-processing them with Discrete/Random Separation and using the envelope spectrum to identify faults [4].

In the scope of artificial intelligence [5,6], numerous papers have been published proposing different methods to train models for REBs diagnosis. In the field of ML, Dong et al. proposed a method based on SVM and the Markov model to predict the degradation process of REBs [7]. Pandya et al. extracted features based on Acoustic Emission (AE) analysis with the Hilbert–Huang Transform (HHT) and used them to train a k-NN model [8]. Piltan et al. proposed a non-linear observer-based technique called Advanced Fuzzy Sliding Mode Observer (AFSMO) to improve the average performance of fault identification with a DT model [9]. In the field of DL, Pan et al. used the second-generation wavelet transform to improve the robustness of DL-based fault diagnosis [10]. Peng et al. converted vibration signals from REBs into greyscale images and used them to extract features to serve as inputs to their proposed CNN [11]. Zhao et al. conducted a DL-based benchmark study evaluating four models: a multilayer perceptron, an auto-encoder, a CNN and an recurrent neural network [12]. Duong et al. suggested obtaining Defect Signature Wavelet Images (DSWIs) from REBs signals, which show the visualization of the discriminated pattern for different types of faults, to train a CNN [13].

Nevertheless, the known advantages offered by the cepstral domain to enhance cyclostationary components in the envelope spectrum are not sufficiently developed by the community in order to extract functional condition indicators that train ML models to diagnose REBs faults automatically. In this paper, we prove the usefulness of CPW combined with the envelope spectrum to extract frequency-domain features from the (CWRU) signals database that, combined with time-domain features, have yield accuracies over 97% for some ML models. The former frequency-domain condition indicators are the impulse responses in the envelope spectrum, whose amplitudes are the maximum of all impulse responses' amplitudes spaced at each fault's characteristic frequency for every signal and their log ratios (Ball Pass Frequency of the Inner Race (BPFI) Amplitude, Ball Pass Frequency of the Outer Race (BPFO) Amplitude, Ball Spin Frequency (BSF) Amplitude, Log (BPFI Amplitude/BPFO Amplitude), Log (BPFI Amplitude/BSF Amplitude), Log (BSF Amplitude/BPFO Amplitude)). The latter time-domain indicators are basic statistics (mean, standard deviation, Root Mean Square (RMS) and shape factor), high-order statistics (kurtosis and skewness) and impulsive metrics (peak value, impulse factor, crest factor and clearance factor). In particular, the log ratio between the maximum amplitude of the impulse responses spaced at the BPFI and the maximum amplitude of the impulse responses spaced at the BPFO in the envelope spectrum has been proven a crucial feature.

The significance of the aforementioned condition indicators has been assessed by applying two statistical methods: one-way Analysis of Variance (ANOVA) and the Kruskal–Wallis test. The ML models trained with these features are Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN) and Naïve Bayes (NB), developed by the MATLAB®-powered *Statistics and Machine Learning Toolbox 11.7™*. Therefore, this paper focuses on the extraction, analysis and demonstration of feature validity, and not on developing application-specific ML models. Their performance has been compared with that of a CNN trained with greyscale images of raw time signals, as this is a simple method that has delivered outstanding results in previous studies. Furthermore, we provide our MATLAB® functions and processed datasets for a proper assessment of the merits of our method.

This manuscript is organised as follows: Section 2.1 summarises the CWRU's experimental set-up. Section 2.2 explains the theoretical basis for frequency- and time-domain feature extraction. Section 2.3 covers the statistical analysis of these features and concludes by explaining the method for training ML models. Section 2.4 summarises the

process for obtaining greyscale images from the raw CWRU's time signals. Section 2.5 summarises the proposed CNN's architecture and defines its hyper-parameters. Section 3.1 presents the statistical analysis results and the different ML algorithms' validation accuracies. Section 3.2 presents CNN's validation accuracies. Section 4 discusses the results, comparing them with some of the works referred to above. Finally, Section 5 concludes by presenting the main advantages and disadvantages of the proposed method.

2. Materials and Methods

The following section will cover the in-depth feature extraction for ML and DL models. It also details the method by which the different ML algorithms will be trained, as well as the statistical analysis to which the time- and frequency-domain features have been subjected in order to shed light on the importance of each one for the classification problem. Furthermore, the architecture of the proposed CNN and its different hyper-parameters are explained.

2.1. CWRU's Experimental Setup Overview

The test stand layout is depicted in Figure 1, and the different REBs geometries and faults' characteristic frequencies are detailed in [14].
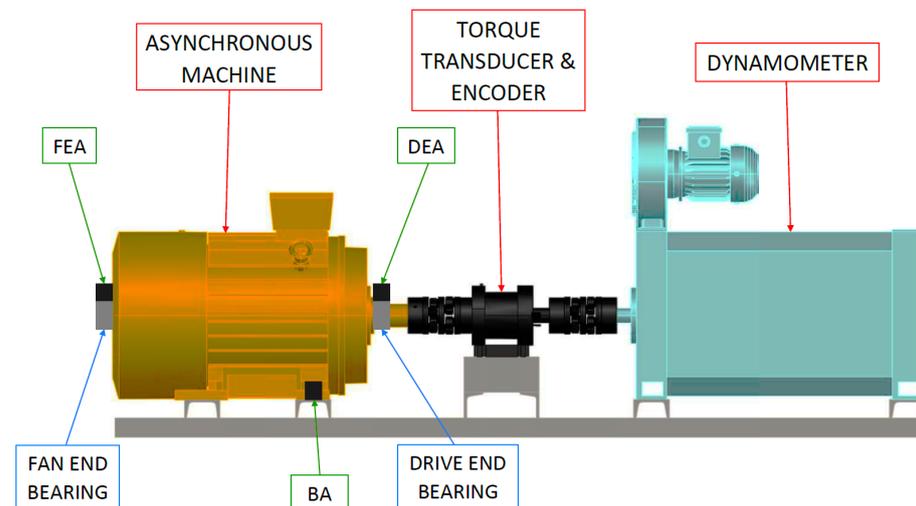

Figure 1. CWRU experimental set-up layout.

In summary, the test rig consisted of a two-hp induction motor, a torque transducer and encoder and a dynamometer. Faults were implanted into the different REBs' parts with an electro-discharge machine. Outer race faults were seeded at different positions relative to the REB load zone, since they are stationary, with the fault's position having an essential effect on the motor/REB system's vibration response. Vibration signals were captured with accelerometers. During each test, acceleration was measured at the drive-end REB housing (Drive End Accelerometer (DEA)), the fan end REB housing (Fan End Accelerometer (FEA)), and the motor support base plate (Base Accelerometer (BA)). Data were captured at 12 kHz and 48 kHz sampling rates using a 16-channel data acquisition card.

The faulty REBs vibration signals to be classified belong to the drive and fan end REBs tables [14], accompanied by healthy REBs vibration signals. These vibration signals have a commonality: they were collected at a sample rate of 12 kHz. Thus, four groups are derived from the dataset: healthy, inner race fault, outer race fault and ball fault. If vibration signals collected at 48 kHz had wanted to be included, they would have formed four different groups, taking into account the difference between sample rates. For simplification purposes, they have not been included in the dataset. As previously mentioned, each file contains FEA, DEA and BA data. Overall, there are 307 vibration

signals, 8 corresponding to healthy REBs, 76 to inner race fault REBs, 76 to ball fault REBs and 147 to outer race fault REBs, as summarised in Table 1.

Table 1. 12k-sampled REBs signals. Health conditions and class labels.

Health Condition	Total Dataset	Class Labels
Normal	8	0
Inner race fault	76	1
Ball fault	76	2
Outer race fault	147	3

2.2. Machine Learning Condition Indicators

From a mechanical vibration point of view, each drive train component excites the entire system at its corresponding characteristic frequency in a wind turbine. When an accelerometer is placed at a certain point in the machine, the signal obtained is a sum of the convolutions of the excitations of the different components and the transfer path of these excitations to the measuring element. The latter can be modelled mathematically as a specific impulse response function, as shown by Barszcz in [15].

REBs are components that transmit the loads coming from the supported shaft. They consist of an outer race, an inner race, the rolling elements and a cage to maintain their relative position. When a spall appears on the surface of one of these parts due to fatigue, the rolling elements repetitively impact it. These impacts excite the system in the form of repetitive impulses at the corresponding fault's characteristic frequency, a function of the shaft rotation frequency, the REB geometry, the number of rolling elements and the load angle, as shown in Table 2.

Table 2. Characteristic frequencies of REB faults

Fault Description	Characteristic Frequency	Fault Location
BPFO	$f_r \cdot \frac{N_r}{2} \cdot \left(1 - \frac{d_B \cdot \cos\theta}{d_P}\right)$	Outer race
BPFI	$f_r \cdot \frac{N_r}{2} \cdot \left(1 + \frac{d_B \cdot \cos\theta}{d_P}\right)$	Inner race
BSF	$f_r \cdot \frac{d_P}{d_B} \cdot \left(1 - \frac{d_B^2 \cdot \cos^2\theta}{d_P^2}\right)$	Rolling element

Where, f_r is Rotational frequency of the shaft [Hz]. d_B is Rolling element diameter [m]. d_P is Pitch diameter [m]. N_r is Number of rolling elements (for a single row) [-]. θ is Load angle (contact angle from radial) [rad].

By computing the fast Fourier transform [16] of a time-domain signal captured by an accelerometer, the harmonics of such excitation will be observed in a specific bandwidth, where the impulse's resonant frequency will be the fault's characteristic frequency. Usually, the fault information will be masked in a complex signal such as that of a wind turbine's drive train.

The envelope spectrum is a widely used tool for diagnosing REB faults in an early stage of development and limited size. Information about the different excitation sources is extracted from the spacing between impulse responses and not from excited frequencies. Therefore, the envelope spectrum consists of a series of impulse responses in the frequency domain spaced at the characteristic frequency of the different excitation sources, as shown in Figure 2. When obtaining the envelope spectrum of a time-domain signal for diagnosing REB faults, it is essential to demodulate the signal in the bandwidth where the fault information is present. For this purpose, a tool called the Kurtogram can be used.
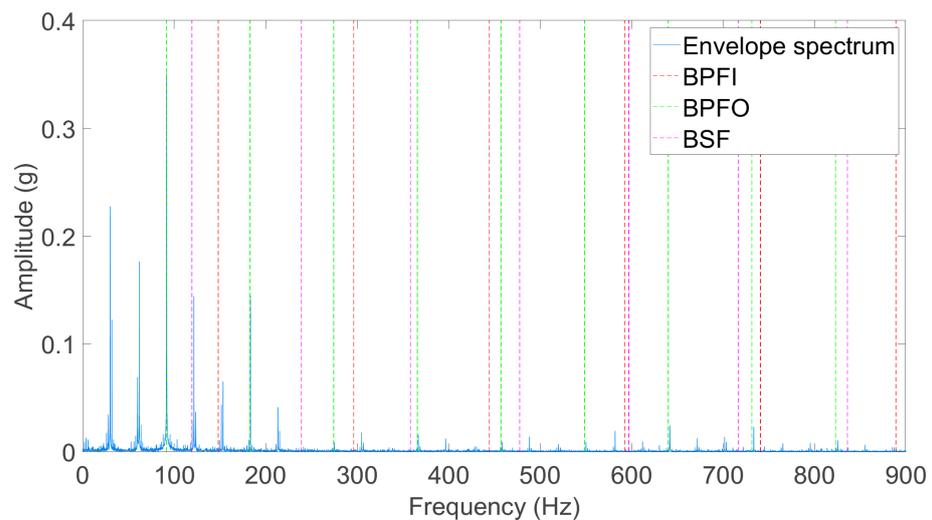
Figure 2. Envelope spectrum of a raw time signal. The signal belongs to the CWRU's bearing database, fan end REB, 0.014″ depth, centred-to-the-load outer race fault, tested at a motor speed of 1797 rpm.

As mentioned above, the rolling elements' impacts on a spall excite the mechanical system in the form of repetitive impulses at the corresponding fault's characteristic frequency. These impulses take the form of peaks in a time signal. Spectral Kurtosis (SK) is a remarkably sensitive indicator of a signal's peakedness. SK is a statistical method to detect non-Gaussian components in a signal, i.e., very impulsive components. Randall and Antoni demonstrated the usefulness of SK in detecting faults in rotating machines [17]. The practical application of SK is the Kurtogram proposed by Antoni and Randall in the same paper, a 2D colour map representing the SK at different levels and bandwidths of a signal. The Kurtogram returns the signal's most impulsive bandwidth and its centre frequency. Later, Antoni proposed an optimised version of the Kurtogram called the Fast Kurtogram [18]. This version reduces the number of variants for which the filter parameters are calculated without affecting the result's accuracy by applying the filter bank approach. The main drawback of SK is that it is susceptible to non-Gaussian random components external to the mechanical system, such as noise or interference of various sorts.

Sawalhi & Randall's CPW [19] can be applied to a raw time signal to enhance the presence of REB's fault's impulse responses in the envelope spectrum and remove both unwanted components' harmonics and sidebands. Earlier in this section, REB fault signals have been referred to as repetitive and not periodic, as they are not strictly periodic: they are second-order cyclostationary. This means that their second-order statistic, the variance, is periodic but not the impulses themselves. The repetitive impulses of REB failures are not precisely periodic due to the rolling elements' slightly random location in the cage-free space and the non-exact cage's rotational speed due to slippage. This is why the REB's fault information is not altered by applying CPW, as the cyclostationary components of a signal do not represent significant peaks in the cepstral domain. A detailed explanation of CPW can be found in the paper by Borghesani et al. [20].

In summary, when a signal is demodulated correctly, a series of impulse responses will appear in the envelope spectrum spaced at the fault's characteristic frequency. When a fault affects a specific REB part, the impacts with the fault will excite a particular frequency band in the form of repetitive impulses. Suppose this band (including the transfer path from the place of impact to the sensor) is selected for demodulation. In that case, the impulse responses in the envelope spectrum spaced at the actual fault's characteristic frequency will have a greater amplitude than those spaced at other fault's characteristic frequencies, as depicted in Figure 2. Computing CPW to a raw time signal enhances cyclostationary components' responses by filtering out periodic components' harmonics and sidebands

from the envelope spectrum. Furthermore, it sets all frequency components to the same magnitude (around $\times 10^{-4}$ for the CWRU's signals), as shown in Figure 3.
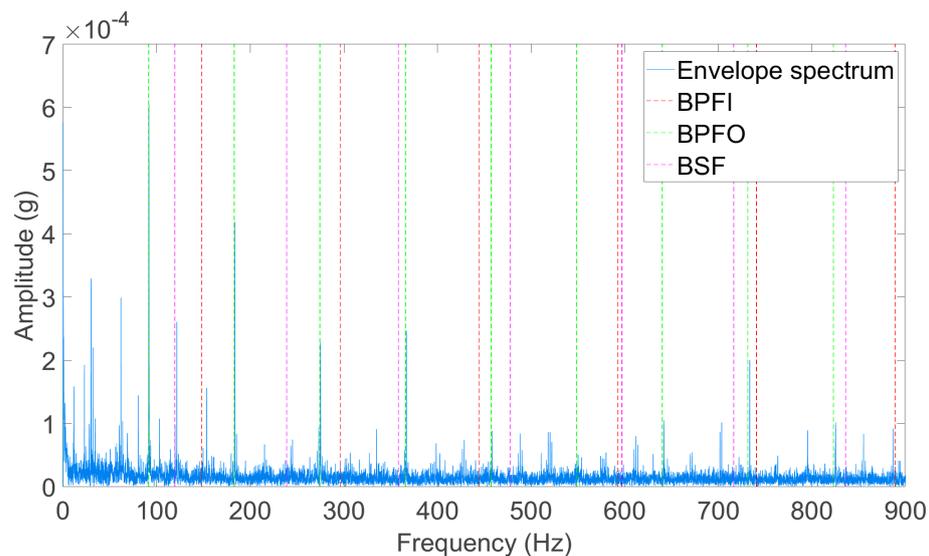

Figure 3. Envelope spectrum of the signal from Figure 2 after CPW.

The above defines a health-state condition indicator that can be calculated using MATLAB®. The idea is to set up a function [21,22] that computes and stores the impulse responses in the envelope spectrum, whose amplitudes are the maximum of all impulse responses' amplitudes spaced at each fault's characteristic frequency for every signal (BPFI Amplitude, BPFO Amplitude and BSF Amplitude). Then, by calculating the ratio of the different amplitudes, a quantified relationship between each magnitude and the others is obtained, reflecting their proportion in each case. The logarithm of these ratios is then computed to set them to the same scale. The above is known in the statistical jargon as the log ratio of two values.

$$Log\left(\frac{BPFI\ Amplitude}{BPFO\ Amplitude}\right)$$

$$Log\left(\frac{BSF\ Amplitude}{BPFO\ Amplitude}\right)$$

$$Log\left(\frac{BPFI\ Amplitude}{BSF\ Amplitude}\right)$$

The combination of time- and frequency-domain features has yielded outstanding results, such as in the research by Sánchez, R.-V. et al. [23]. Therefore, the former condition indicators have been combined with time-domain features such as basic statistics (mean, standard deviation, RMS and shape factor), high-order statistics (kurtosis and skewness) and impulsive metrics (peak value, impulse factor, crest factor and clearance factor). These have been obtained by taking advantage of the MATLAB®-powered *Predictive Maintenance Toolbox 2.2*™, ultimately leading to the MATLAB® function shown in [24,25]. Both condition indicator sets have been ranked according to one-way ANOVA [26] & the Kruskal–Wallis test [27] and used to train four widely renowned ML models (DT, SVM, k-NN and NB).

2.3. Machine Learning-Based Classification Method & Statistical Analysis

The most important step when approaching a ML classification problem is selecting the best feature combination to achieve the most accurate results possible. This is of the utmost importance when the dataset's size is limited. The most natural questions to be asked once the condition indicators have been calculated are: how are these features meaningful to the classification problem? Are any of them better than the rest? Two ranking techniques

for datasets containing more than two classes will be applied to answer these questions: one-way ANOVA [26] and the Kruskal–Wallis test [27].

Comparing the means of three or more unrelated groups within a dependent variable (feature), one-way ANOVA determines whether there are statistically significant differences between them. A non-parametric alternative to the one-way ANOVA test is the Kruskal–Wallis test. Unlike one-way ANOVA, where means are compared, the Kruskal–Wallis test contrasts the samples' distribution to determine whether they belong to the same population. Both tests make assumptions that the data of each group in every dependent variable must meet. The Kruskal–Wallis test is known to be used if one-way ANOVA's assumptions are not met for each class within every dependent variable, since they are more potent than those of the Kruskal–Wallis test.

One-way ANOVA tests the null hypothesis that all group means are equal (H_0 : $\mu_1 = \mu_2 = \cdots = \mu_J$) against the alternative hypothesis that at least one group mean is different from the others ($H_1 : \mu_{j'} \neq \mu_j$ for at least one j' and j) in a one-way layout, where $i = 1, \ldots, I$ is the observation number and $j = 1, \ldots, J$ is the group number. The one-way ANOVA result is the ratio of across-group variation to within-group variation F. If F is larger than the critical value of the F-distribution [28] with ($J - 1$, $J(I - 1)$) degrees of freedom and a significance level of α, the null hypothesis is rejected. Therefore, large values of F rank better if the degree of difference between the groups is considered to evaluate if the dependent variable is suitable for a classification problem.

In the Kruskal–Wallis test, the null hypothesis states that J groups from potentially different populations actually derive from a similar population, at least regarding their central tendencies or medians. As an alternative hypothesis, not all groups are derived from the same population. The Kruskal–Wallis test result is the test statistic H. If H is larger than the critical value of the chi-square distribution [29] with $J - 1$ degrees of freedom and a significance level of α, the null hypothesis is rejected. Therefore, large values of H rank better if the degree of difference between the groups is to be considered.

For the one-way ANOVA results to be reliable, the residuals of the I observations belonging to the J groups need to meet the following assumptions:

1. They have to be normally distributed in each group being compared. In practice, the dependent variable is tested to be normally distributed in each group rather than the residuals since the results are the same.
2. There is homogeneity of variances (homoscedasticity) between each group of residuals. Again, the population variances in each group are tested to be equal rather than the residuals' variances in each group.
3. Independence. The residuals (or rather the observations) need to be independent.

One-way ANOVA is known to be a robust test against the normality assumption, especially for large datasets. When the homoscedasticity condition has been violated, the Welsch-correlated ANOVA test has proved reliable. It has been stated that the most critical assumption to fail is the lack of independence between observations. For the results of the Kruskal–Wallis test to be reliable, the observations within each group need to meet the following assumptions:

1. They do not have to be normally distributed in each group. However, the observations within each group have to belong to the same continuous distribution.
2. There is homogeneity of variances (homoscedasticity) between each group of observations.
3. Independence of observations.

The use of the Kruskal–Wallis test is recommended when the populations to be compared are clearly asymmetric; it is fulfilled that they are all in the same direction and that the variance is homogeneous. To assess if the data within every group and feature are suitable for the results of one-way ANOVA or the Kruskal–Wallis test to be reliable, several hypothesis tests will be applied using the MATLAB®-powered *Statistics and Machine Learning Toolbox 11.7*™. These are the Anderson–Darling test [30] to test the

normality assumption, the Levene test [31] to test the homoscedasticity condition and the Kolmogorov–Smirnov test [32] to test the equality of continuous distributions condition.

Once the features are ranked, the procedure to investigate which features will deliver the highest accuracy for every ML algorithm will be as follows: for every case, each ML algorithm will be computed with a features vector that will be of dimension 1, firstly, increasing its dimensionality until it reaches its maximum, following the order of importance established by each statistical test. Therefore, two tables will be produced, each with the features vector's dimensionality being increased differently. This way, the best results can be achieved with the fewest features possible, understanding which features are valid and which are not for every algorithm.

2.4. Deep Learning Condition Indicators Overview

In addition, a CNN has been built to classify REBs signals from a DL point of view. CNNs are a type of feed-forward neural network initially designed for image processing. Therefore, the CWRU's healthy and faulty REBs signals must be transformed into images, with every pixel within an image being a single feature. CNNs are known to deliver outstanding results when the training dataset size is considerably large [33]. This paper will test its performance using a 14.736 image database generated from the CWRU's raw time signals using the code shown in [34,35].

To transform a 1D time-domain signal into a 2D image to serve as an input to the CNN, N data are split off from a signal $\mathbf{x}$. Each datum is aligned sequentially, shaping m rows of n points. A matrix is then built by placing one of the m rows onto the following. To create square images of acceptable size while preserving as much information as possible about the defect, a 120 k data signal section has been transformed into 48 50×50 pixels images. The REBs signals sampled at 12 kHz have approximately 120 k data, given that each reading took about 10 seconds. Suppose the test stand's shaft rotational speed is 1730 rpm. In that case, a single rolling element will impact approximately 29 times against the fault every second, giving a total of 290 impacts of each rolling element with the fault during these ten seconds (this pattern is not valid if the defect is in the rolling element). In mathematical terms, the transformation described above looks like this:

$$\mathbf{I}_{mxn} = \begin{pmatrix} x(t) & \dots & x(t+n-1) \\ \vdots & \ddots & \vdots \\ x(t+(m-1)n) & \dots & x(t+mn-1) \end{pmatrix} \tag{1}$$

where $\mathbf{I}$ denotes the signal image and $x(t)$ the vibration datum of time t. In this particular case, $t_0 = 1$ and $m = n = 50$.
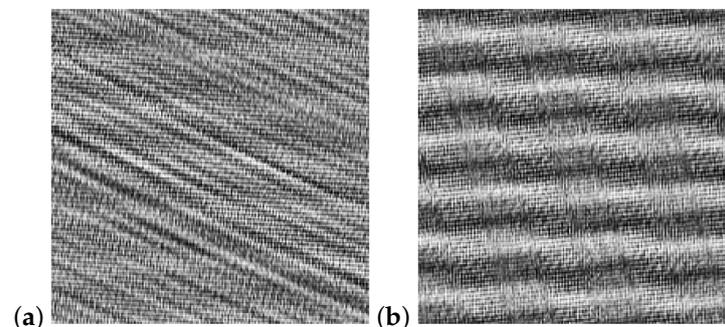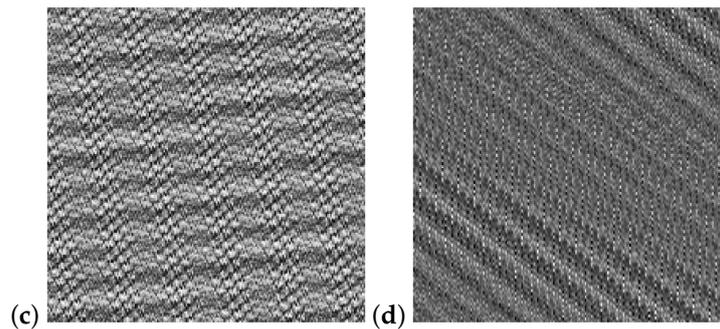
Some signal images are depicted in Figure 4:

(a)   (b)

Figure 4. *Cont.*

Figure 4. Greyscale images of drive-end REB signals. Faults located in the rolling elements (**a**,**b**), inner race (**c**) and outer race orthogonal to the load zone (**d**).

2.5. CNN Architecture and Hyper-Parameters Overview

The architecture of the proposed CNN model, shown in [36,37], which comprises one input layer responsible for receiving external data, two hidden layers accountable for filtering the inputs, a fully connected layer responsible for the classification and an output layer. Each hidden layer comprises a convolutional layer, a batch normalisation layer, an activation layer and a max-pooling layer. The convolutional layer convolves the local input regions with filter kernels and then generates the output features by computing the activation function. The *ReLU* function is used as the activation layer to sort out the issue of vanishing gradients. The batch normalisation layer, placed between the convolutional layer and the activation layer, helps to reduce the CNN's sensitivity to network initialisation. Finally, a max-pooling layer is placed between the first activation layer and the second convolutional layer. This is known to improve the CNN's accuracy. The fully connected layer comprises as many neurons as there are labels—four, in this case. The *softmax* function is used as an activation layer for the latter.

Every ML algorithm aims to minimise a loss/cost function or maximise a likelihood function, depending on the existing model, to find its optimal values and achieve the most accurate prediction. In a CNN, every unit within a convolutional layer is a slight regression model trained by learning the filter kernel's parameters. The loss function, in this case, is defined by the task. For image classification, the cross-entropy loss function is used to compute the difference between the *softmax* output probability distribution and the label probability distribution. The stochastic gradient descent with momentum is used to minimise the cross-entropy loss function. It is an upgraded version of the stochastic gradient descent that accelerates gradient vectors in the right direction. The hyper-parameters for the proposed CNN are shown in Table 3. Furthermore, the training options for the SGDM are set to 0.01 for the learning rate, 15 epochs, a mini-batch size of 35 and a frequency of network validation in 50 iterations. These parameters have been tuned by trial and error.

Table 3. Hyper-parameters of the proposed CNN.

	Kernel		Stride		Padding			
	Height	Width	Vertical Step Size	Horizontal Step Size	Top	Bottom	Left	Right
Convolutional Layers	5	5	1	1	aut.	aut.	aut.	aut.
Pooling layers	2	2	2	2	0	0	0	0

3. Results

The following section presents the outcomes of applying one-way ANOVA and the Kruskal–Wallis test to the time- and frequency-domain features and the validation results of the different ML algorithms and the CNN.

3.1. Machine Learning-Based Classification

The metric used to assess the different algorithms under the various cases is accuracy, defined as the proportion of correctly classified examples divided by the total number of

examples within the dataset. As mentioned in the previous section, to learn which features work better for every ML algorithm under each case, the procedure will be to compute each algorithm with a features vector of dimension 1, increasing its dimensionality until it reaches its maximum for every case, following the order of importance established by each statistical test. Prior to ranking any feature, let us observe various scatter plots of both datasets. The aim is to find significant differences between the features' distributions for each group.

Figures 5 and 6 depict 3D scatter plots from the time- and frequency-domain datasets. As drawn from these figures, the frequency-domain features seem to be the ones that take the most segregated values for each group, especially the logarithmic ratios between the maximum amplitudes of the envelope spectrum around the pulses spaced at the faults' characteristic frequency. The time-domain features seem to take less segregated values between the groups. Therefore, lower accuracy is expected to be obtained from these features. Moreover, data belonging to the faulty balls class (2) are more spread than the other groups' data in the cases where a significant difference between the latter can be observed.
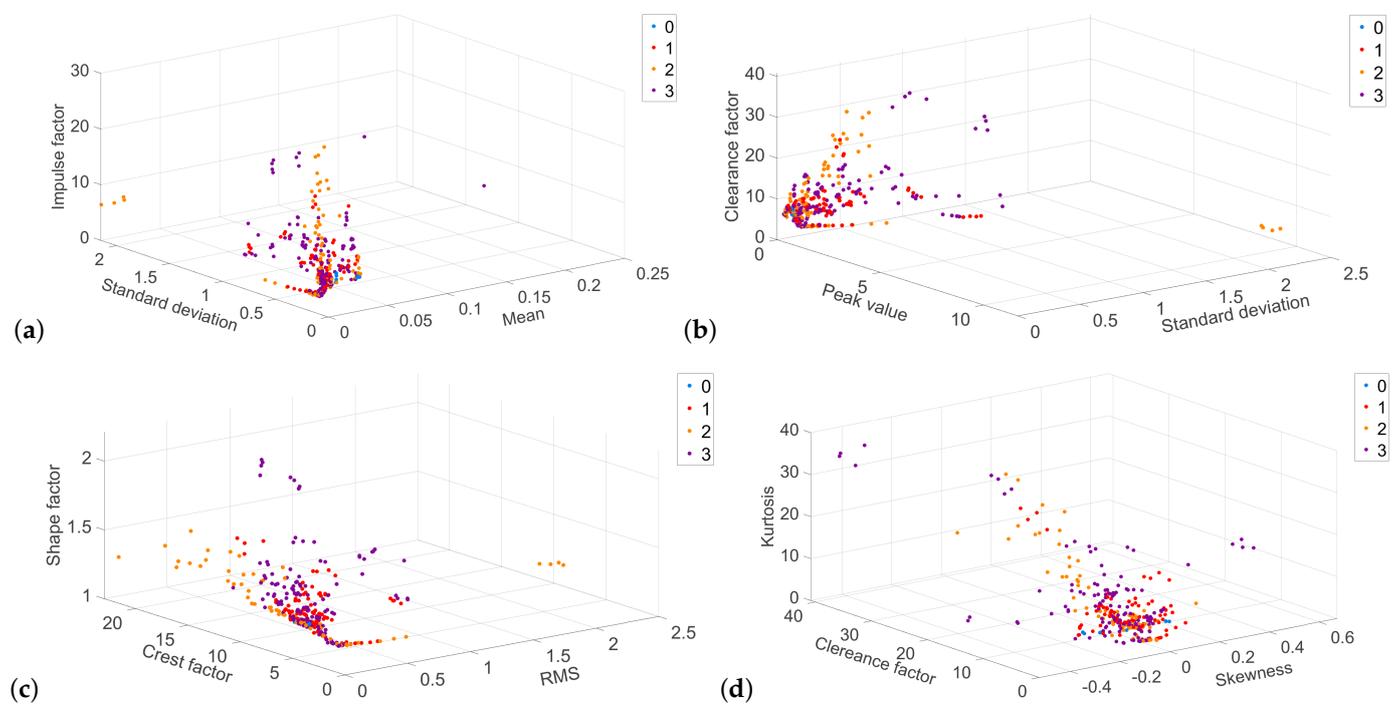
Figure 5. Scatter plots of time−domain features. (**a**) Mean, Standard deviation & Impulse factor. (**b**) Standard deviation, Peak value & Clearance factor. (**c**) RMS, Crest factor & Shape factor. (**d**) Skewness, Clearance factor & Kurtosis.

Considering all of the above, two case studies will be defined.

1. Case (A) Classification using the whole dataset.
2. Case (B) Classification removing the second class from the dataset.

The features' ranking results are shown in Figure 7 and collected in Tables 4–7.
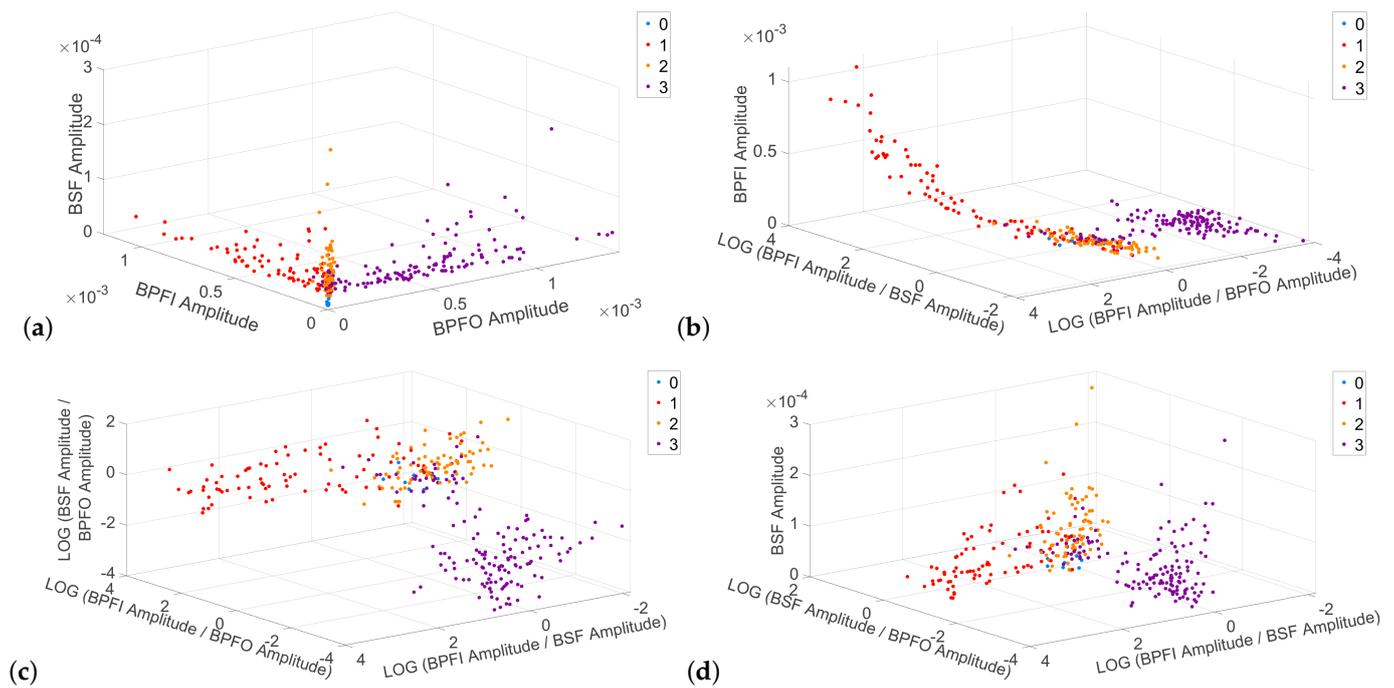
Figure 6. Scatter plots of frequency−domain features. (**a**) BPFO Amplitude, BPFI Amplitude & BSF Amplitude. (**b**) Log(BPFI Amplitude/BPFO Amplitude), Log(BPFI Amplitude/BSF Amplitude) & BPFI Amplitude. (**c**) Log(BPFI Amplitude/BSF Amplitude), Log(BPFI Amplitude/BPFO Amplitude) & Log(BSF Amplitude/BPFO Amplitude). (**d**) Log(BPFI Amplitude/BSF Amplitude), Log(BSF Amplitude/BPFO Amplitude) & BSF Amplitude.

Table 4. Features sorted by importance according to one-way ANOVA. Case A.

Features	F
LOG_BPFI_Amplitude_BPFO_Amplitude	258.3292
LOG_BSF_Amplitude_BPFO_Amplitude	122.8081
LOG_BPFI_Amplitude_BSF_Amplitude	116.2330
BPFI_Amplitude	92.2322
BPFO_Amplitude	77.7082
BSF_Amplitude	9.5288
Crest factor	6.0064
Shape factor	5.8229
Impulse factor	3.4917
Clearance factor	2.7894
Kurtosis	1.9256
Mean	1.4055
Peak value	1.3931
Standard deviation	0.9941
RMS	0.9637
Skewness	0.6254

Table 5. Features sorted by importance according to one-way ANOVA. Case B.

Features	F
LOG_BPFI_Amplitude_BPFO_Amplitude	289.5471
LOG_BPFI_Amplitude_BSF_Amplitude	129.0033
BPFI_Amplitude	95.1479
LOG_BSF_Amplitude_BPFO_Amplitude	90.1346
BPFO_Amplitude	61.6727
BSF_Amplitude	6.0888
Shape factor	4.0389
Clearance factor	2.9775
Kurtosis	2.8668
Standard deviation	2.7570
Impulse factor	2.6555
RMS	2.6403
Peak value	2.5531
Crest factor	1.8718
Mean	1.2457
Skewness	0.7408

Table 6. Features sorted by importance according to the Kruskal–Wallis test. Case A.

Features	H
LOG_BPFI_Amplitude_BPFO_Amplitude	217.3216
BPFI_Amplitude	167.1512
LOG_BSF_Amplitude_BPFO_Amplitude	155.3533
BPFO_Amplitude	138.4140
LOG_BPFI_Amplitude_BSF_Amplitude	120.2450
BSF_Amplitude	43.3603
Standard deviation	26.0577
RMS	25.2882
Shape factor	17.9515
Peak value	15.7401
Mean	12.3161
Kurtosis	11.9904
Clearance factor	5.3638
Impulse factor	5.1157
Crest factor	4.8157
Skewness	3.0968

Table 7. Features sorted by importance according to the Kruskal–Wallis test. Case B.

Features	H
LOG_BPFI_Amplitude_BPFO_Amplitude	152.5596
BPFI_Amplitude	145.4471
LOG_BPFI_Amplitude_BSF_Amplitude	100.1278
BPFO_Amplitude	97.8711
LOG_BSF_Amplitude_BPFO_Amplitude	91.5079
BSF_Amplitude	24.0144
Standard deviation	18.0231
RMS	17.5023
Peak value	15.3398
Mean	11.6661
Kurtosis	10.9086
Shape factor	10.1311
Clearance factor	5.9409
Impulse factor	5.4390
Crest factor	3.8549
Skewness	2.7573
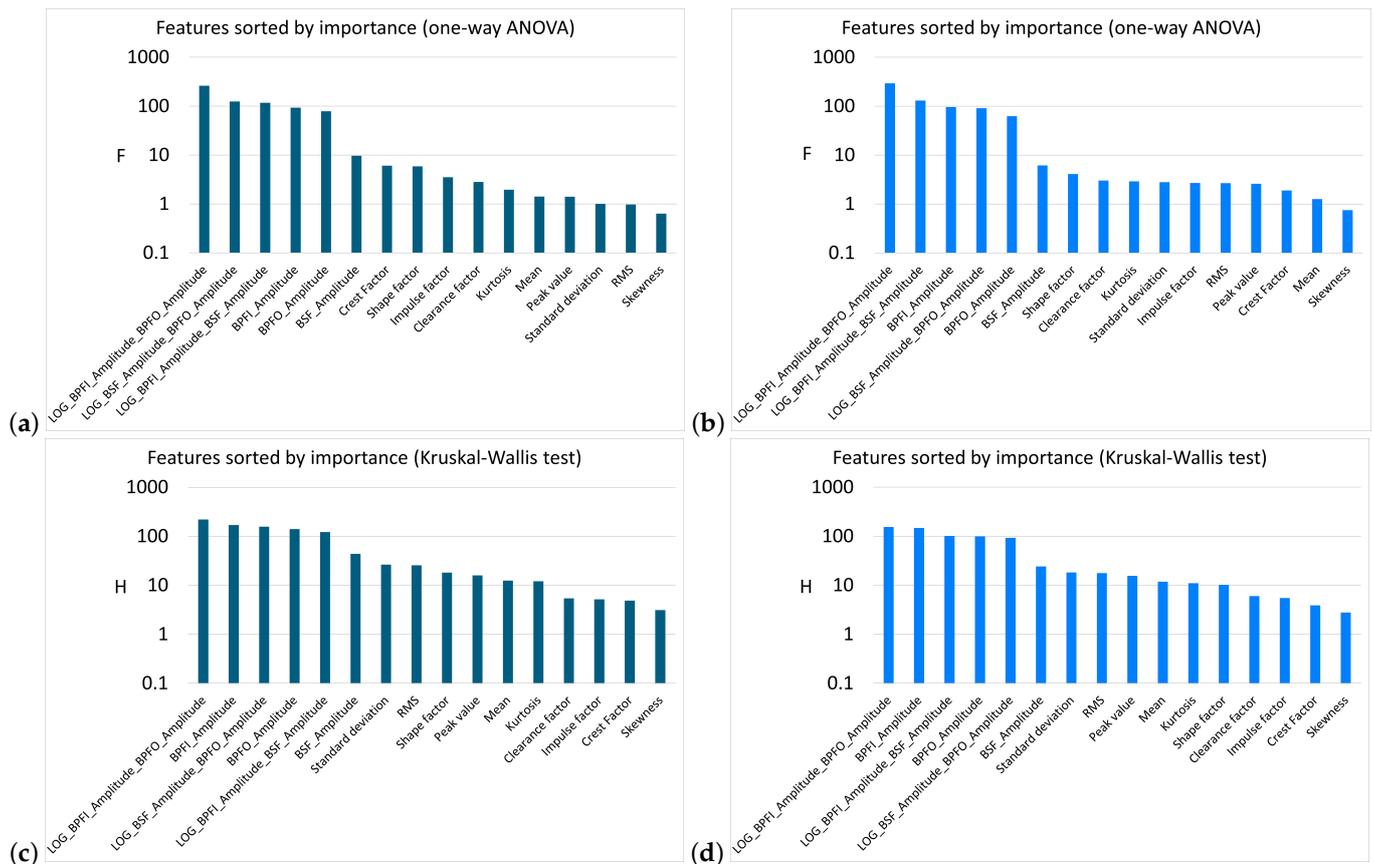
Figure 7. Bar graphs of cases A & B for both one-way ANOVA and the Kruskal-Wallis test. (**a**) Dataset of case A ranked according to the one-way ANOVA test. (**b**) Dataset of case B ranked according to the one-way ANOVA test. (**c**) Dataset of case A ranked according to the Kruskal-Wallis test. (**c**) Dataset of case B ranked according to the Kruskal-Wallis test.

As shown in Figure 7, for every case, the frequency-domain features are ranked better than the time-domain features, especially with regard to the logarithmic ratios between the maximum amplitudes of the envelope spectrum around the pulses spaced at the faults' characteristic frequencies. This matches the conclusions drawn from the scatter plots in Figures 5 and 6. The hypothesis tests results are summarised in Table A1, Appendix A. Every feature within the dataset violates the normality and equality of continuous distributions requirements. Only one feature (mean) fulfils the homoscedasticity condition. The independence of observations condition is known to be fulfilled for the whole dataset, since the features have been calculated from the raw time signals, which were independently obtained as described in Section 2.1, from different REBs placed in different parts of the setup, in addition to the accelerometers.

The classification results are depicted in Figures 8–11 and collected in Tables 8–11. Using the MATLAB®-powered *Statistics and Machine Learning Toolbox 11.7*™, models have been trained by applying five-fold cross-validation, using 80% of the dataset for training and 20% for validation during each iteration. Five-fold cross-validation also protects against overfitting by estimating accuracy on each fold.

Table 8. ML algorithms performance. Features vector's dimensionality increased following one-way ANOVA order. Case A.

Features Vector's Dimensionality	DT	SVM	k-NN	Naïve Bayes
1	0.788	0.805	0.785	0.805
2	0.808	0.824	0.827	0.821
3	0.765	0.818	0.801	0.808
4	0.824	0.840	0.821	0.834
5	0.821	0.831	0.818	0.831
6	**0.83**	0.837	0.840	0.837
7	0.821	0.824	0.824	0.844
8	0.808	0.840	0.827	0.840
9	0.821	0.808	0.821	0.840
10	0.821	0.841	0.847	0.840
11	0.811	0.831	0.824	**0.847**
12	0.827	0.834	0.824	0.847
13	0.814	0.824	0.827	0.837
14	0.831	0.831	0.850	0.847
15	0.827	**0.847**	0.866	0.837
16	0.821	0.821	**0.876**	0.840

The best result obtained for each model by applying the proposed method has been marked in bold.

Table 9. ML algorithms performance. Features vector's dimensionality increased following one-way ANOVA order. Case B.

Features Vector's Dimensionality	DT	SVM	k-NN	Naïve Bayes
1	0.896	0.909	0.922	0.905
2	0.900	0.909	0.900	0.887
3	0.939	0.944	0.926	0.931
4	0.935	0.939	0.913	0.935
5	0.944	0.939	0.920	0.944
6	0.948	0.948	0.957	**0.948**
7	0.952	0.952	0.957	0.935
8	**0.957**	0.952	0.957	0.939
9	0.944	0.948	0.952	0.935
10	0.948	0.965	0.961	0.944
11	0.948	0.965	**0.965**	0.935
12	0.952	**0.974**	0.961	0.935
13	0.939	0.970	0.965	0.948
14	0.948	0.961	0.965	0.944
15	0.944	0.961	0.961	0.939
16	0.948	0.948	0.961	0.939

The best result obtained for each model by applying the proposed method has been marked in bold.

Table 10. ML algorithms performance. Features vector's dimensionality increased following the Kruskal–Wallis test's order. Case A.

Features Vectors' Dimensionality	DT	SVM	k-NN	Naïve Bayes
1	0.788	0.805	0.785	0.805
2	0.808	0.818	0.785	0.814
3	0.818	0.831	0.818	0.831
4	0.824	0.834	0.821	0.831
5	0.834	0.840	0.814	0.834
6	0.834	0.837	0.840	0.837
7	0.821	0.837	0.821	0.837
8	0.831	0.837	0.824	0.834
9	0.798	0.811	0.827	0.834
10	**0.850**	0.827	0.834	0.840
11	0.824	0.834	0.824	0.840
12	0.795	0.818	0.831	0.834
13	0.837	0.834	0.870	**0.847**
14	0.824	0.837	0.837	0.834
15	0.827	**0.847**	0.866	0.837
16	0.821	0.821	**0.876**	0.840

The best result obtained for each model by applying the proposed method has been marked in bold.

Table 11. ML algorithms performance. Features vectors' dimensionality increased following the Kruskal–Wallis test's order. Case B.

Features Vector's Dimensionality	DT	SVM	k-NN	Naïve Bayes
1	0.896	0.909	0.922	0.905
2	0.948	0.948	0.939	0.935
3	0.939	0.944	0.926	0.931
4	0.948	0.944	0.926	0.948
5	0.944	0.939	0.920	0.944
6	0.948	0.948	0.957	0.948
7	**0.957**	0.970	**0.974**	0.939
8	0.948	0.974	0.974	**0.957**
9	0.952	0.974	0.974	0.952
10	0.952	0.970	0.970	0.944
11	0.944	0.970	0.974	0.952
12	0.952	0.965	0.965	0.935
13	0.952	**0.978**	0.970	0.948
14	0.952	0.974	0.970	0.939
15	0.944	0.961	0.961	0.939
16	0.948	0.948	0.961	0.939

The best result obtained for each model by applying the proposed method has been marked in bold.

Figures 10 and 11 clearly show that signals from REBs with faulty balls are the most difficult to classify using the ML models. The algorithms computed with case A dataset give, in all scenarios, worse results than those calculated with case B dataset. Delving deeper into the development of each model, in Figures 8 and 9, it is not always true that a features vector's higher dimensionality leads to better results. Two models exemplify that: DT and NB, which in all cases have required fewer features to achieve their best results than SVM or k-NN. Regardless, all models needed fewer features to achieve their best results in case B rather than in case A. This is because features calculated on the REBs' signals with faulty balls do not help find correlations that ultimately lead to a correct classification. The best-performing models are k-NN for case A [38,39] and SVM for case B [40,41].
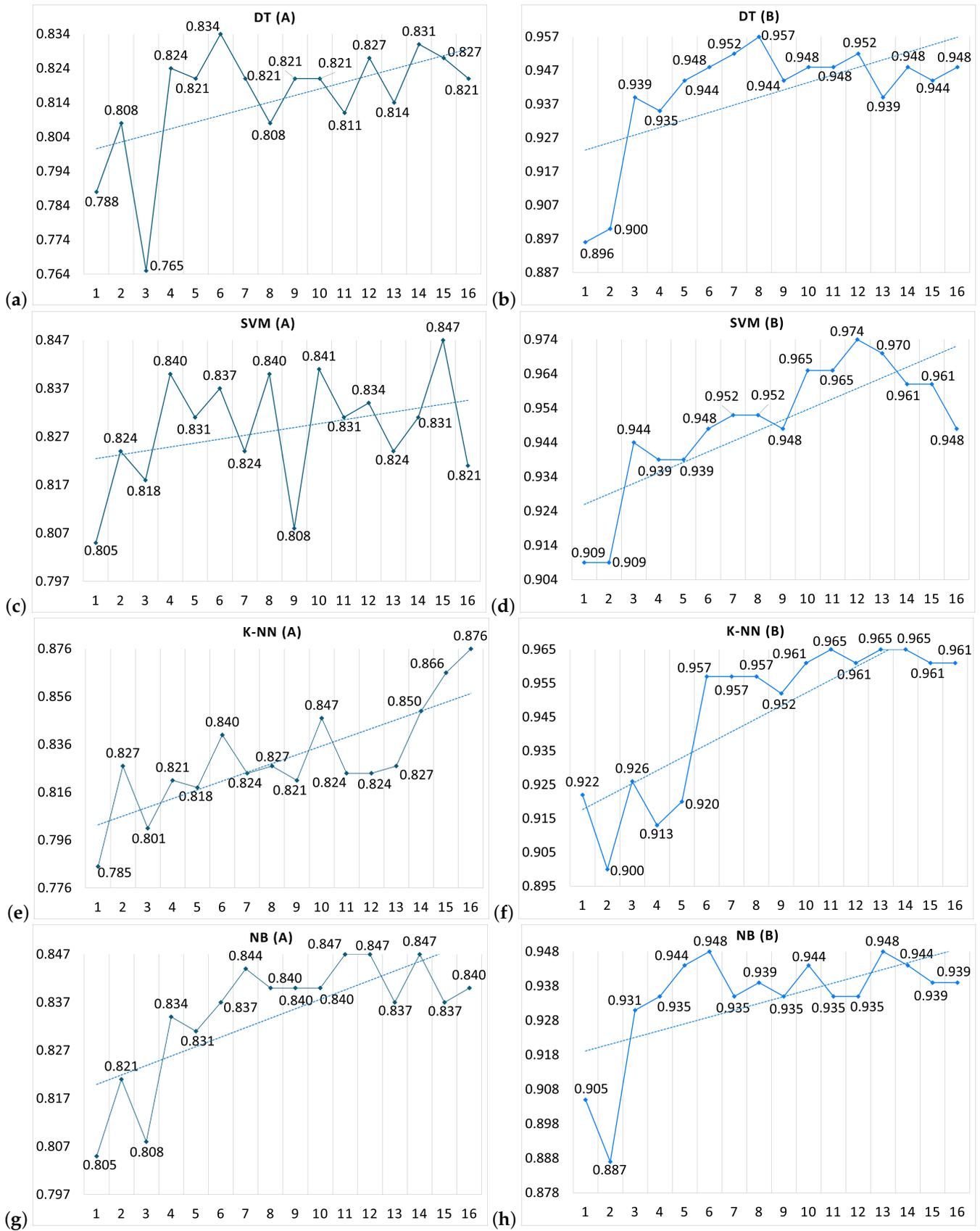
Figure 8. ML algorithms performance. Features vectors' dimensionality increased following one-way ANOVA order. Plots (**a**,**c**,**e**,**g**) for case A and plots (**b**,**d**,**f**,**h**) for case B.
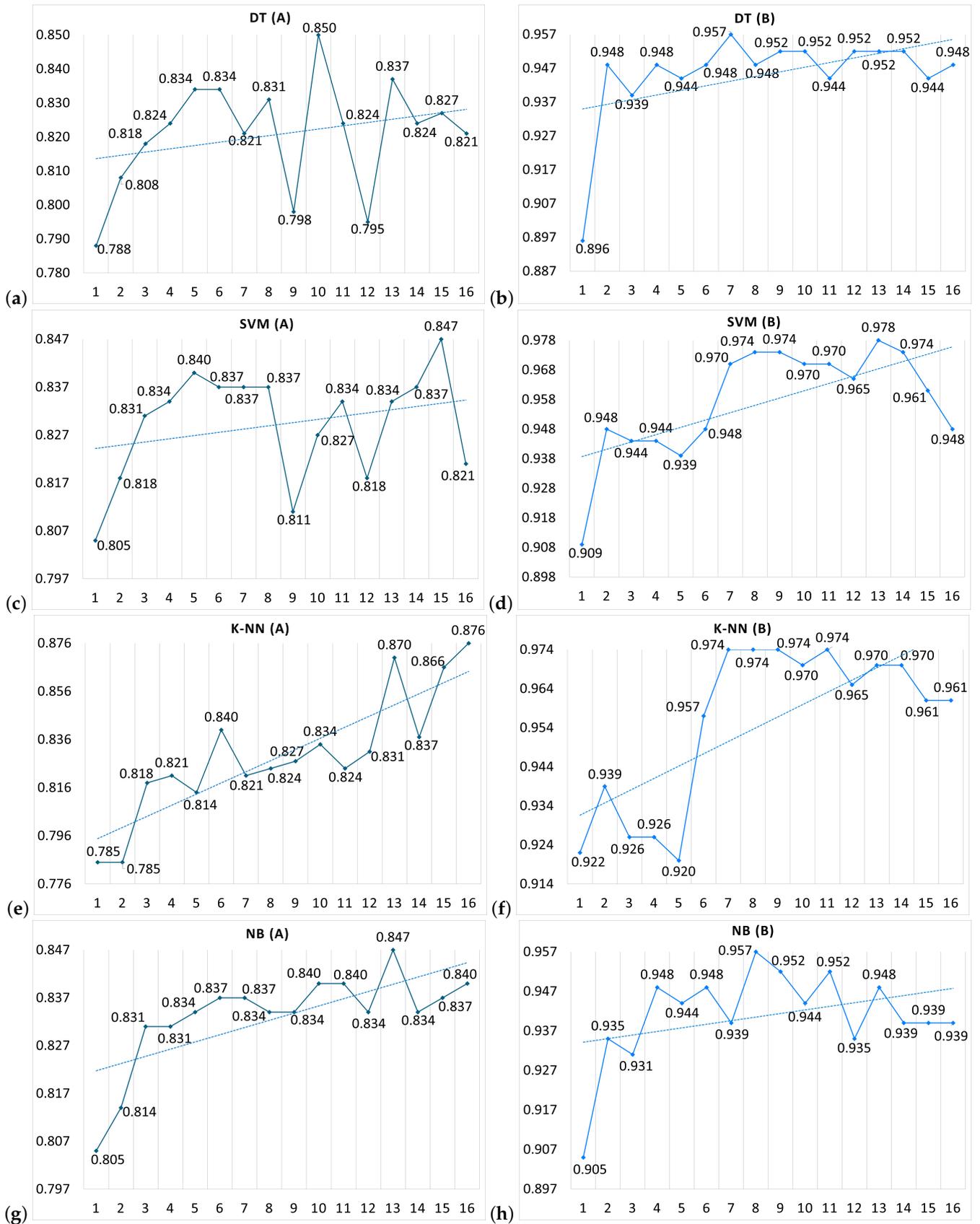
Figure 9. ML algorithms performance. Features vectors' dimensionality increased following the Kruskal–Wallis test's order. Plots (**a,c,e,g**) for case A and plots (**b,d,f,h**) for case B.
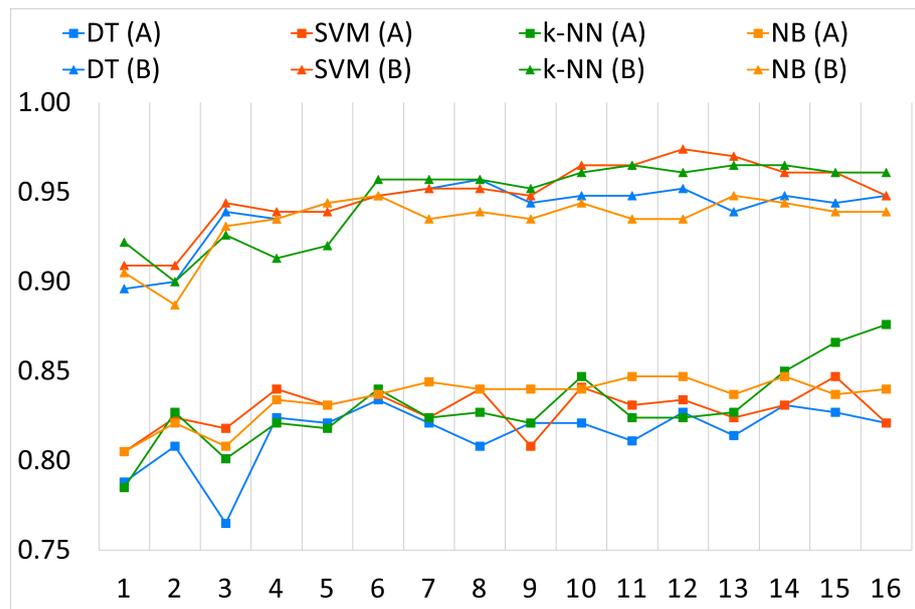
Figure 10. All ML algorithms' performance. Features vectors' dimensionality increased following one-way ANOVA order.
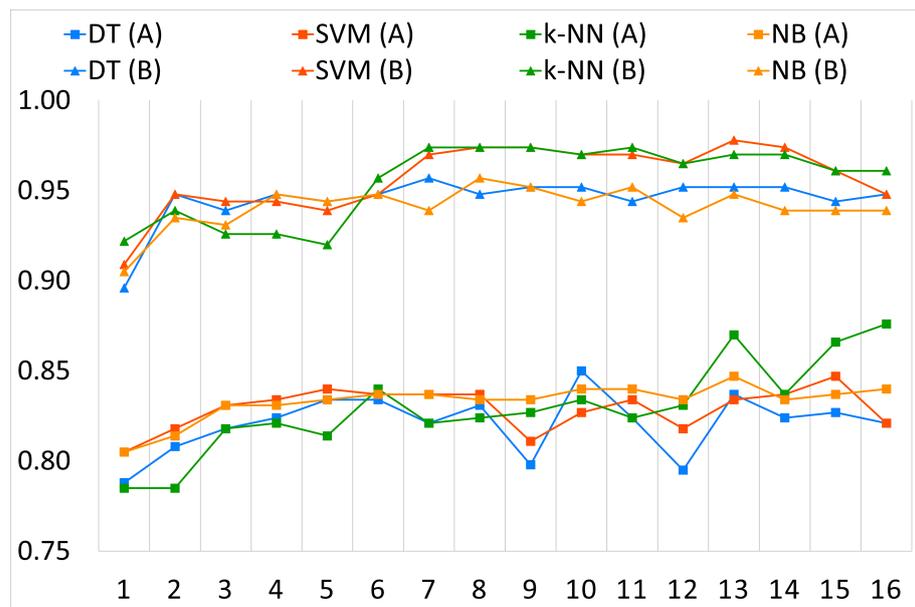
Figure 11. All ML algorithms performance. Features vectors' dimensionality increased following the Kruskal–Wallis test's order.

3.2. Deep Learning-Based Classification

The previous section stated that REBs' signals with faulty balls are expected to be the most difficult for the ML algorithms to classify. The question is whether the CNN will also have the same difficulties finding correlations between REBs' signals with faulty balls. The two case studies defined in Section 3.1 will be applied to answer this question. The proposed CNN architecture and the chosen hyper-parameters have been defined in Section 2.5. For the ML case, the CNN has been trained with 80% of the images in the dataset and validated with the remaining 20% of them in each case. Both subsets have been randomly split. The classification results are shown in Table 12.

Table 12. CNN results for cases A and B.

	Case A	Case B
Validation accuracy	0.7761	0.9067
Elapsed time	5 min 14 s	5 min 44 s
Epochs	15	15
Maximum iterations	5040	3765
Iterations per epoch	336	253
Frequency	50 iterations	50 iterations
Learning rate	0.01	0.01

Worse results have been obtained from the CNN. Although a database of 14.736 images has been generated, taking full advantage of the length of the CWRU's signals, this has not been sufficient to improve the results obtained from some ML models. The computational time to train the proposed CNN has been much higher than what was required for the ML models, which was a few seconds. To improve the CNN results, more images could be obtained from more signals, generating a more extensive dataset.

4. Discussion

This paper has trained various ML algorithms and a CNN to classify healthy and faulty REBs. Each model's best results are summarised in Figure 12. The different models have been trained using the CWRU bearing data center signals. These models are prepared to receive features of never-before-seen REBs' signals to diagnose their state of health, with an accuracy corresponding to the validation result obtained for each model. Considering the results obtained, the following can be stated:

1. Although the frequency-domain features are ranked better, according to the applied statistical methods, than the time-domain features, it can be observed in the accuracy development of some models that some frequency-domain features contribute to worse signal classification against some time-domain features.
2. The logarithmic ratio between the maximum amplitude of the impulse responses spaced at the BPFI and the maximum amplitude of the impulse responses spaced at the BPFO is the feature containing the most relevant information about REBs' health status. That follows from the high accuracy obtained by computing the algorithms with this feature alone. According to the statistical methods, it is also the best-positioned feature in the rankings.
3. The combination of time- and frequency-domain features yields better results than by being computed separately. This can be deduced since the best results combine both types of features. To design a robust ML model that generalises correctly, the most important thing to find is the features' combination providing the most relevant information about the REBs' health status.

Table 13 compares the results of some relevant papers cited in the introduction with those obtained in this paper. The Root Mean Square Error (RMSE) of the model proposed by Dong et al. suggests an almost perfect REB degradation prediction compared to the data obtained in their experiment.

The results of ML models trained with features based on AE analysis with the HHT, proposed by Pandya et al., demonstrate the usefulness of these physical features. The observer-based non-linear models of Piltan et al. have helped improve the accuracy of the DT model. However, these authors divide the CWRU dataset into numerous groups, considering fault sizes and different load regimes. The latter work overlooks the fact that the load is practically meaningless in this case, as there is no mechanism in the CWRU experiment that converts the torque into a radial load supported by the REBs. Furthermore, the usefulness of these non-physical models with more heterogeneous groups is unclear. Pan et al. obtained excellent results in their novel neural network enhanced with the second-generation wavelet transform. Peng et al. show that the simple method of converting raw

time signals into greyscale images to train a CNN provides outstanding results. Once again, the CWRU dataset is subdivided into numerous groups in this work, considering different fault sizes and load regimes. Finally, the model proposed by Duong et al. demonstrates the usefulness of DSWI obtained from faulty REBs' signals to train CNNs.

Figure 12. Summary of results for the ML and DL models.

Table 13. Results comparison.

Ref.	Method		RMSE	Ref.	Method		Accuracy
[7]	SVM model		0.0521		CNN (Greyscale Images Training Set)		0.987
	SVM Markov-model		0.0091	[11]			
Ref.	**Method**		**Accuracy**		CNN (Greyscale Images Validation Set)		0.997
	Bayesian		80.55–91.67%				
	NB		71.67–91.11%		CNN (Greyscale Images Test Set)		0.989
	MLP		87.22–96.11%				
[8]	k-NN		87.78–92.77%	**Ref.**	**Method**		**Accuracy**
	Weighted k-NN		91.67–96.11%			Vibration Image Method	51.42%
	J48		89.44–90.55%		KNN + PCA	Wavelet Spectrogram	48.99%
	Random Forest		90.56–95.55%			DSWI	61.13%
Ref.	**Method**		**Accuracy**			Vibration Image Method	80.97%
		SMO	80–100%		MCSVM + PCA	Wavelet Spectrogram	81.91%
	Constant torque load	ASMO	81–100%			DSWI	87.76%
		AFSMO	84–100%	[13]		Vibration Image Method	46.77%
[9]		SMO	75–100%		LeNet-5	Wavelet Spectrogram	89.11%
	Variable torque load	ASMO	81–100%			DSWI	95.97%
		AFSMO	75–100%			Vibration Image Method	68.54%
Ref.	**Method**		**Accuracy**		AlexNet	Wavelet Spectrogram	90.72%
	FS1 + SVM		75.00% & 73.57%			DSWI	97.98%
	FS2 + SVM		92.50% & 78.81%			Vibration Image Method	83.06%
[10]					DCNN	Wavelet Spectrogram	93.15%
	LiftingNet		99.63% & 93.19%			DSWI	98.79%

This work evaluated the applicability of physical methods, such as the CPW-enhanced envelope spectrum, for the ML-based classification problem. Feature extraction from this method has yielded excellent results demonstrating its usefulness in combination with time-domain features. These results have been superior to those of the CNN trained with greyscale images of raw time signals under the same conditions. The results obtained in this paper are to be understood as a complement to the study of relevant features for diagnosing REBs based on ML.

5. Conclusions

In general terms, excellent accuracies have been obtained from the ML models. The statistical tests' results and the different algorithms' accuracy have shown that the frequency-domain features, particularly the logarithmic ratios between the maximum amplitudes of the envelope spectrum around the pulses spaced at the faults' characteristic frequency, have been the most important ones. These, in addition to some of the most critical time-domain features, have been able to yield the highest accuracies in this paper. Among the models, the k-NN and SVM classifiers were the best performers. The accuracy of these algorithms is given in Tables 8–11. It can also be seen from Figures 10 and 11 that signals corresponding to REBs with faulty balls have been the most difficult for the algorithms to classify. The CNN's accuracy can be improved by increasing the dataset's size, as mentioned in Section 4. These models can be potentially helpful for automatically diagnosing the health status of REBs.

The main drawback of the proposed method is that the database on which the study has been conducted is imbalanced, as seen in Table 1. This means that the classification result may be biased, as the classifiers are more sensitive to detecting the majority class and less sensitive to the minority class. A common technique to counteract the imbalanced data problem is to oversample the dataset. Nevertheless, after applying cross-validation, it is known to lead to over-optimistic results, for which conventional model evaluation metrics such as accuracy are not good indicators of the algorithms' performance. In the industry, faults in REBs are typically not evenly distributed, with malfunctions such as those affecting inner and outer races occurring more frequently than others. Therefore, dealing with imbalanced databases is a common challenge. The best practice to address this issue is to apply cross-validation by stratifying each k-fold to capture the unbalanced distribution of groups on each target feature. In this paper, as discussed, conventional five-fold cross-validation has been implemented. This is an essential point of improvement for future research.

Author Contributions: Conceptualization, E.Q.-C.; methodology, E.Q.-C., D.C.-F. and E.G.-M.; software, D.C.-F.; validation, E.Q.-C., D.C.-F. and E.G.-M.; formal analysis, D.C.-F., E.Q.-C. and E.G.-M.; investigation, D.C.-F., E.Q.-C. and E.G.-M.; resources, D.C.-F., E.Q.-C. and E.G.-M.; data curation, D.C.-F.; writing—original draft preparation, D.C.-F.; writing—review and editing, E.Q.-C.; visualisation, E.Q.-C.; supervision, E.Q.-C. and E.G.-M.; project administration, E.Q.-C. and E.G.-M. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in FigShare and have been cited accordingly.

Abbreviations

The following abbreviations are used in this manuscript:

AE   Acoustic Emission.
AFSMO   Advanced Fuzzy Sliding Mode Observer.
ANOVA   ANalysis Of VAriance.
BA   Base Accelerometer.
BPFI   Ball Pass Frequency of the Outer Race.
BPFO   Ball Pass Frequency of the Inner Race.
BSF   Ball Spin Frequency.
CNN   Convolutional Neural Network.
CPW   Cepstrum Pre-Whitening.
CWRU   Case Western Reserve University.
DEA   Drive End Accelerometer.
DL   Deep Learning.
DSWI   Defect Signature Wavelet Image.
DT   Decision Tree.
FEA   Fan End Accelerometer.
HHT   Hilbert–Huang Transform.
k-NN   k-Nearest Neighbors.
ML   Machine Learning.
NB   Naïve Bayes.
REB   Rolling Element Bearing.
RMS   Root Mean Square.
RMSE   Root Mean Square Error.
SK   Spectral Kurtosis.
SVM   Support Vector Machine.

Appendix A

Table A1. Hypothesis tests applied to every group within each feature.

Features	Normality Condition [1]	Homoscedasticity Condition [2]		Equality of Continuous Distributions Condition [3]		
		W	*p*-Value	1	2	3
Clearance_Factor_0	0			1	1	1
Clearance_Factor_1	1			N/A	1	1
Clearance_Factor_2	1	10.7727	0	=	N/A	0
Clearance_Factor_3	1			=	=	N/A
Crest_Factor_0	0			1	1	0
Crest_Factor_1	1			N/A	1	0
Crest_Factor_2	1	27.6535	0	=	N/A	1
Crest_Factor_3	1			=	=	N/A
Impulse_Factor_0	0			1	1	1
Impulse_Factor_1	1			N/A	1	1
Impulse_Factor_2	1	15.1834	0	=	N/A	0
Impulse_Factor_3	1			=	=	N/A
Kurtosis_0	0			1	1	1
Kurtosis_1	1			N/A	1	1
Kurtosis_2	1	5.4175	0.0012	=	N/A	1
Kurtosis_3	1			=	=	N/A
Mean_0	1			1	1	1
Mean_1	1			N/A	0	0
Mean_2	1	1.3019	0.2739	=	N/A	0
Mean_3	1			=	=	N/A
Peak_Value_0	0			1	1	1
Peak_Value_1	1			N/A	0	0
Peak_Value_2	1	4.555	0.0039	=	N/A	0
Peak_Value_3	1			=	=	N/A
RMS_0	0			1	1	1
RMS_1	1			N/A	1	0
RMS_2	1	4.3323	0.0052	=	N/A	0
RMS_3	1			=	=	N/A

Table A1. *Cont.*

Features	Normality Condition [1]	Homoscedasticity Condition [2]		Equality of Continuous Distributions Condition [3]		
		W	*p*-Value	1	2	3
Shape_Factor_0	0			1	1	1
Shape_Factor_1	1	10.6547	0	N/A	1	0
Shape_Factor_2	1			=	N/A	1
Shape_Factor_3	1			=	=	N/A
Skewness_0	0			0	0	0
Skewness_1	1	4.3422	0.0051	N/A	1	0
Skewness_2	1			=	N/A	0
Skewness_3	1			=	=	N/A
Std_0	0			1	1	1
Std_1	1	4.3611	0.005	N/A	1	0
Std_2	1			=	N/A	0
Std_3	1			=	=	N/A
BPFO_Amplitude_0	0			1	1	1
BPFO_Amplitude_1	0	107.138	0	N/A	0	1
BPFO_Amplitude_2	1			=	N/A	1
BPFO_Amplitude_3	1			=	=	N/A
BPFI_Amplitude_0	1			1	1	1
BPFI_Amplitude_1	1	148.66	0	N/A	1	1
BPFI_Amplitude_2	1			=	N/A	1
BPFI_Amplitude_3	1			=	=	N/A
BSF_Amplitude_0	1			1	1	1
BSF_Amplitude_1	1	5.1763	0.0017	N/A	1	0
BSF_Amplitude_2	1			=	N/A	1
BSF_Amplitude_3	1			=	=	N/A
LOG_BPFI_Amplitude_BPFO_Amplitude_0	0			1	0	1
LOG_BPFI_Amplitude_BPFO_Amplitude_1	1	31.1992	0	N/A	1	1
LOG_BPFI_Amplitude_BPFO_Amplitude_2	0			=	N/A	1
LOG_BPFI_Amplitude_BPFO_Amplitude_3	1			=	=	N/A
LOG_BSF_Amplitude_BPFO_Amplitude_0	0			0	1	1
LOG_BSF_Amplitude_BPFO_Amplitude_1	1	34.6674	0	N/A	1	1
LOG_BSF_Amplitude_BPFO_Amplitude_2	0			=	N/A	1
LOG_BSF_Amplitude_BPFO_Amplitude_3	1			=	=	N/A
LOG_BPFI_Amplitude_BSF_Amplitude_0	0			1	1	0
LOG_BPFI_Amplitude_BSF_Amplitude_1	1	31.1532	0	N/A	1	1
LOG_BPFI_Amplitude_BSF_Amplitude_2	0			=	N/A	1
LOG_BPFI_Amplitude_BSF_Amplitude_3	1			=	=	N/A

[1] The Anderson–Darling test returns a decision for the null hypothesis that data in vector $\mathbf{x}$ are from a population with a normal distribution. The alternative hypothesis is that $\mathbf{x}$ is not from a population with a normal distribution. The result is 1 if the test rejects the null hypothesis at the 5% significance level, or 0 otherwise. [2] The Levene test returns a decision for the null hypothesis that columns of data matrix $\mathbf{X}$ have the same variance. The alternative hypothesis is that not all columns of data have the same variance. The Levene test rejects the null hypothesis that the variances are equal if $W >$ the upper critical value of the F distribution with $J - 1$ and $J(I - 1)$ degrees of freedom at a significance level of α. The p-value represents the probability of observing a test statistic as extreme as or more extreme than the observed value under the null hypothesis. [3] The Kolmogorov–Smirnov test returns a decision for the null hypothesis that data in vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are from the same continuous distribution. The alternative hypothesis is that $\mathbf{x}_1$ and $\mathbf{x}_2$ are from different continuous distributions. The result is 1 if the test rejects the null hypothesis at the 5% significance level, or 0 otherwise.

References

1. Global Wind Energy Council (GWEC). Global Wind Report 2021. Available online: https://gwec.net/global-wind-report-2021/#:~:text=Today%2C%20there%20is%20now%20743,carbon%20emissions%20of%20South%20America (accessed on 14 September 2022).
2. Kiral, Z.; Karagülle, H. Vibration analysis of rolling element bearings with various defects under the action of an unbalanced force. *Mech. Syst. Signal Process.* **2006**, *20*, 1967–1991. [CrossRef]
3. Sawalhi, N.; Randall, R.B. Vibration response of spalled rolling element bearings: Observations, simulations and signal processing techniques to track the spall size. *Mech. Syst. Signal Process.* **2011**, *25*, 846–870. [CrossRef]
4. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64–65*, 100–131. [CrossRef]
5. Heidari, A.; Navimipour, N.J.; Unal, M. Applications of ML/DL in the management of smart cities and societies based on new trends in information technologies: A systematic literature review. *Sustain. Cities Soc.* **2022**, *85*, 104089. [CrossRef]
6. Heidari, A.; Jabraeil Jamali, M.A.; Navimipour, N.J.; Akbarpour, S. Deep Q-Learning Technique for Offloading Offline/Online Computation in Blockchain-Enabled Green IoT-Edge Scenarios. *Appl. Sci.* **2022**, *12*, 8232. [CrossRef]
7. Dong, S.; Yin, S.; Tang, B.; Chen, L.; Luo, T. Bearing Degradation Process Prediction Based on the Support Vector Machine and Markov Model. *Shock Vib.* **2014**, 1–15. [CrossRef]
8. Pandya, D.H.; Upadhyay, S.H.; Harsha, S.P. Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. *Expert Syst. Appl.* **2013**, *40*, 4137–4145. [CrossRef]
9. Piltan, F.; Prosvirin, A.E.; Jeong, I.; Im, K.; Kim, J.-M. Rolling-Element Bearing Fault Diagnosis Using Advanced Machine Learning-Based Observer. *Appl. Sci.* **2019**, *9*, 5404. [CrossRef]

10. Pan, J.; Zi, Y.; Chen, J.; Zhou, Z.; Wang, B. LiftingNet: A Novel Deep Learning Network With Layerwise Feature Learning from Noisy Mechanical Data for Fault Classification. *IEEE Trans. Ind. Electron.* **2018**, *65*, 4973–4982. [CrossRef]

11. Peng, X.; Zhang, B.; Gao, D. Research on fault diagnosis method of rolling bearing based on 2DCNN. In Proceedings of the 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, 22 August 2020; pp. 693–697. [CrossRef]

12. Zhao, Z.; Li, T.; Wu, J.; Sun, C.; Wang, S.; Yan, R.; Chen, X. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA Trans.* **2020**, *107*, 224–255. [CrossRef] [PubMed]

13. Duong, B.P.; Kim, J.Y.; Jeong, I.; Im, K.; Kim, C.H.; Kim, J.M. A Deep-Learning-Based Bearing Fault Diagnosis Using Defect Signature Wavelet Image visualisation. *Appl. Sci.* **2020**, *10*, 8800. [CrossRef]

14. Case Western Reserve University (CWRU). Bearing Data Center. Available online: https://engineering.case.edu/bearingdatacenter/welcome (accessed on 18 September 2022).

15. Barszcz, T. *Vibration-Based Condition Monitoring of Wind Turbines*, 1st ed.; Springer: Cham, Switzerland, 2019; pp. 8–23. [CrossRef]

16. Randall, R.B. *Vibration-Based Condition Monitoring: Industrial, Automotive and Aerospace Applications*, 1st ed.; John Wiley & Sons: Hoboken, NJ, USA, 2010; pp. 66–95. [CrossRef]

17. Antoni, J.; Randall, R.B. The spectral kurtosis: Application to the vibratory surveillance and diagnostics of rotating machines. *Mech. Syst. Signal Process.* **2004**, *20*, 308–331. [CrossRef]

18. Antoni, J. Fast computation of the kurtogram for the detection of transient faults. *Mech. Syst. Signal Process.* **2006**, *21*, 108–124. [CrossRef]

19. Sawalhi, N.; Randall, R.B. Signal pre-whitening using cepstrum editing (liftering) to enhance fault detection in rolling element bearings. In Proceedings of the 24 International Congress on Condition Monitoring and Diagnostic Engineering Management (COMADEM2011), Stavanger, Norway, 30 May 2011; pp. 330–336.

20. Borghesani, P.; Pennacchi, P.; Randall, R.B.; Sawalhi, N.; Ricci, R. Application of cepstrum pre-whitening for the diagnosis of bearing faults under variable speed conditions. *Mech. Syst. Signal Process.* **2012**, *36*, 370–384. [CrossRef]

21. Cascales Fulgencio, D. *Frequency_Domain_Features*, version 4; FigShare; 2022. Available online: https://figshare.com/articles/software/Frequency_Domain_Features_m/21150967/4 (accessed on 17 September 2022).

22. Cascales Fulgencio, D. *all_12k_sets_table_I*, version 1; FigShare; 2022. Available online: https://figshare.com/articles/dataset/all_12k_sets_table_I_mat/21151042/1 (accessed on 17 September 2022).

23. Sánchez, R.-V.; Lucero, P.; Macancela, J.-C.; Rubio Alonso, H.; Cerrada, M.; Cabrera, D.; Castejón, C. Evaluation of Time and Frequency Condition Indicators from Vibration Signals for Crack Detection in Railway Axles. *Appl. Sci.* **2020**, *10*, 4367. [CrossRef]

24. Cascales Fulgencio, D. *Time_Domain_Features*, version 4; FigShare; 2022. Available online: https://doi.org/10.6084/m9.figshare.21150976.v2. (accessed on 17 September 2022).

25. Cascales Fulgencio, D. *all_12k_sets_table_II*, version 1; FigShare. 2022. Available online: https://figshare.com/articles/dataset/all_12k_sets_table_II_mat/21151033/1 (accessed on 17 September 2022).

26. NIST/SEMATECH. *e-Handbook of Statistical Methods*. One-Way ANOVA. 2012. Available online: https://www.itl.nist.gov/div898/handbook/ppc/section2/ppc231.htm (accessed on 17 September 2022).

27. NIST/SEMATECH. *e-Handbook of Statistical Methods*. Kruskal-Wallis Test. 2012. Available online: https://www.itl.nist.gov/div898/handbook/prc/section4/prc41.htm (accessed on 17 September 2022).

28. Statistics Online Computational Resource (SOCR). F-Distribution Tables. 2002. Available online: http://www.socr.ucla.edu/Applets.dir/F_Table.html (accessed on 19 September 2022).

29. Statistics Online Computational Resource (SOCR). Chi-Square Distribution Tables. 2002. Available online: http://www.socr.ucla.edu/Applets.dir/ChiSquareTable.html (accessed on 19 September 2022).

30. NIST/SEMATECH. *e-Handbook of Statistical Methods*. Anderson-Darling Test. 2012. Available online: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm (accessed on 19 September 2022).

31. NIST/SEMATECH. *e-Handbook of Statistical Methods*. Levene Test for Equality of Variances. 2012. Available online: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm (accessed on 19 September 2022).

32. NIST/SEMATECH. *e-Handbook of Statistical Methods*. Kolmogorov-Smirnov Goodness-of-Fit Test. 2012. Available online: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm (accessed on 19 September 2022).

33. Zhang, W.; Peng, G.; Li, C. Bearings fault diagnosis based on convolutional neural networks with 2-D representation of vibration signals as input. *MATEC Web Conf.* **2017**, *95*, 13001. [CrossRef]

34. Cascales Fulgencio, D. *Images_Generator*, version 4; FigShare 2022. Available online: https://figshare.com/articles/software/Images_Generator_m/21151021/4 (accessed on accessed on 19 September 2022)

35. Cascales Fulgencio, D. *all_12k_sets_table_III*, version 1; FigShare. 2022. Available online: https://figshare.com/articles/dataset/all_12k_sets_table_III_mat/21151036/1 (accessed on accessed on 19 September 2022)

36. Cascales Fulgencio, D. *TWODCNN*, version 4; FigShare; 2022. Available online: https://figshare.com/articles/software/TWODCNN_m/21151018/4 (accessed on accessed on 19 September 2022)

37. Cascales Fulgencio, D. *all_images_datastore*, version 1; FigShare; 2022. Available online: https://figshare.com/articles/dataset/all_images_datastore_mat/21151045/1 (accessed on accessed on 19 September 2022).

38. Cascales Fulgencio, D. *trainClassifier_KNN_A*, version 4; FigShare; 2022. Available online: https://figshare.com/articles/software/trainClassifier_KNN_A_m/21150970/4 (accessed on accessed on 19 September 2022).

39. Cascales Fulgencio, D. *all_12k_sets_KNN_A_features*, version 1; FigShare; 2022. Available online: https://figshare.com/articles/dataset/all_12k_sets_KNN_A_features_mat/21151030/1 (accessed on accessed on 19 September 2022).

40. Cascales Fulgencio, D. *trainClassifier_SVM_B*, version 4; FigShare; 2022. Available online: https://figshare.com/articles/software/trainClassifier_SVM_B_m/21150973/4 (accessed on accessed on 19 September 2022).

41. Cascales Fulgencio, D. *all_12k_sets_SVM_B_features*, version 1; FigShare; 2022. Available online: https://figshare.com/articles/dataset/all_12k_sets_SVM_B_features_mat/21151039/1 (accessed on accessed on 19 September 2022).