*energies*

MDPI

*Article*

# Early Fault Diagnosis Strategy for WT Main Bearings Based on SCADA Data and One-Class SVM

**Christian Tutivén [1], Yolanda Vidal [2,3,*], Andres Insuasty [4,5], Lorena Campoverde-Vilela [1] and Wilson Achicanoy [4]**

[1] Escuela Superior Politécnica del Litoral (ESPOL), Faculty of Mechanical Engineering and Production Science (FIMCP), Mechatronics Engineering, Campus Gustavo Galindo, Km. 30.5 Vía Perimetral, Guayaquil EC090902, Ecuador; cjtutive@espol.edu.ec (C.T.); lscampov@espol.edu.ec (L.C.-V.)
[2] Control, Data and Artificial Intelligence (CoDAlab), Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Universitat Politècnica de Catalunya (UPC), Campus Diagonal-Besós (CDB), Eduard Maristany 16, 08019 Barcelona, Spain
[3] Institute of Mathematics (IMTech), Universitat Politècnica de Catalunya (UPC), Pau Gargallo 14, 08028 Barcelona, Spain
[4] Departamento de Electrónica, Universidad de Nariño, Clle 18 Cr 50 Ciudadela Universitaria Torobajo, Pasto 52001, Colombia; andresinsuasty@udenar.edu.co (A.I.); wilachic@udenar.edu.co (W.A.)
[5] Mecatrónica, Facultad de Ingenierías, Universidad ECOTEC, Km. 13.5 Vía a Samborondón, Guayaquil EC092302, Ecuador
* Correspondence: yolanda.vidal@upc.edu

**Abstract:** To reduce the levelized cost of wind energy, through the reduction in operation and maintenance costs, it is imperative that the wind turbine downtime is reduced through maintenance strategies based on condition monitoring. The standard approach toward this challenge is based on vibration monitoring, which requires the installation of specific tailored sensors that incur associated added costs. On the other hand, the life expectancy of wind parks built during the 1990s wind power boom is dwindling, and data-driven maintenance strategies issued from already accessible supervisory control and data acquisition (SCADA) data is an auspicious competitive solution because no additional sensors are required. Note that it is a major issue to provide fault diagnosis approaches built only on SCADA data, as these data were not established with the objective of being used for condition monitoring but rather for control capacities. The present study posits an early fault diagnosis strategy based exclusively on SCADA data and supports it with results on a real wind park with 18 wind turbines. The contributed methodology is an anomaly detection model based on a one-class support vector machine classifier; that is, it is a semi-supervised approach that trains a decision function that categorizes fresh data as similar or dissimilar to the training set. Therefore, only healthy (normal operation) data is required to train the model, which greatly expands the possibility of employing this methodology (because there is no need for faulty data from the past, and only normal operation SCADA data is needed). The results obtained from the real wind park show that this is a promising strategy.

**Keywords:** anomaly detection; condition-based maintenance; condition monitoring; fault diagnosis; main bearing; one-class support vector machine; predictive maintenance; SCADA data; wind turbine



**Citation:** Tutivén, C.; Vidal, Y.; Insuasty, A.; Campoverde-Vilela, L.; Achicanoy, W. Early Fault Diagnosis Strategy for WT Main Bearings Based on SCADA Data and One-Class SVM. *Energies* **2022**, *15*, 4381. https://doi.org/10.3390/en15124381

Academic Editor: Davide Astolfi

Received: 23 May 2022
Accepted: 14 June 2022
Published: 16 June 2022



## 1. Introduction

Renewable energy had a record-breaking year in 2020, in stark contrast to the decline observed in the fossil fuel sectors owing to the COVID-19 pandemic, as its installed power capacity grew by more than 260 gigawatts (GW) [1]. Renewable energy refers to the energy obtained from natural resources that are available nearly everywhere; they are inexhaustible, and they cause little to no greenhouse gas emissions.

Among the different forms of renewable energies, wind energy showed remarkable growth in 2020, with 111 GW of new installations [1]. However, taking into account that

the levelized cost of energy (LCOE) is the key tool for evaluating the essential economics of any type of power project, it is noteworthy that operating and maintenance (O&M) costs typically score 20% to 25% of the overall LCOE of modern wind parks [2]. Thus, to further extend the deployment of wind energy, reducing O&M costs has become a priority.

In recent decades, much effort has been devoted to initiating a shift from wind turbine (WT) preventive and/or corrective maintenance to maintenance based on the actual asset condition, that is through condition monitoring systems (CMS), e.g., [3–10]. However, the high cost associated with the installation of the specific tailored extra sensors needed by CMS and the communication infrastructure and protocol required to deal with the high-frequency sampling data from these sensors have postponed their widespread usage [2]. On the other hand, data-driven maintenance approaches from already accessible supervisory control and data acquisition (SCADA) data is an auspicious competitive answer [11]. The present study contributes such a strategy for early warning in the event of main bearing failures in WTs. This is an ambitious goal, as SCADA data is primarily used for control purposes and not for condition monitoring.

Usually, SCADA data logs four different statistics for each variable (sensor) and per certain duration (which is usually 10 min): minimum, maximum, average, and standard deviation. In industrial size WTs, these SCADA data contain between 100 and 200 different variables. Thus, with the properties of low sampling rate and high dimension, SCADA data is usually managed and used by the owners and/or operators for live monitoring and history querying in a remote monitoring center setting. Because data-driven models were not anticipated as a feasible solution to condition monitoring when these systems were established, most operators now have several years of unexplored SCADA data. Furthermore, they did not anticipate the importance of having maintenance annotations. As a result, maintenance logs lack a consistent structure and frequently include a great deal of noise.

There have been several successful contributions to condition monitoring with actual SCADA data in recent years (based on data from real in-production WTs). For example, in [12] the prognosis of a WT gearbox bearing is accomplished, but it proposes a supervised approach that needs historical faulty data to be tagged, which is a tedious and time-consuming task prone to errors (due to the maintenance log's non-standardized and noisy characteristics). In [13], an unsupervised approach is proposed for WT condition monitoring; however, because this study does not have access to work order data, it is not possible to fully validate whether the model is detecting the faults appropriately. In [14], an ensemble technique to detecting abnormalities and diagnosing defects is offered; however, it has only been tried on two WTs, and the alarm is triggered just days ahead of the fatal breakdown occurring, not giving sufficient time to plan the repair.

In contrast to the aforementioned references, this work proposes a main bearing early fault detection strategy based solely on WT SCADA data, which is the main contribution that it addresses. At the same time, the following six main challenges are found in the literature. (i) It uses only standard SCADA data (10-min average); thus, it can be applied to any wind turbine. (ii) It is a normal behavior model; that is, to be constructed (trained) only requires normal (healthy) data. Because it does not require any faulty data, any wind park (even those where the failure of interest has not yet occurred) can benefit from it, and it avoids the problem of highly unbalanced data sets. (iii) It is validated on real (not simulated or experimental) SCADA data and proven to be robust to seasonality and operating and environmental conditions. A significant number of references use simulated SCADA data or experimental data (from a test bench) to validate the results. Although it is understandable, as real SCADA data sets are often proprietary and are not easily available by the scientific community, it is an important drawback as relying on synthetically generated data may not generalize well to actual real-world conditions. (iv) The warning is given months in advance to the fault, thus allowing wind park operators to program the maintenance, in contrast to a non-negligible number of studies based on SCADA data that detect the fault with less than a week in advance, thus not being helpful in a real application. (v) It advances an indicator based on an exponential weighted moving average filter, depending on the

weekly number of anomalies, to reduce the number of false positive alerts in contrary to a substantial number of references that result in a significant number of false alerts, making the contribution inconvenient in the real world, as it would result in alarm fatigue for operators. (vi) The validation is performed at wind farm level encompassing 18 WTs (not only on one or two WTs), as opposed to the majority of the literature that bases its results on a relatively small amount of data, usually only one to four wind turbines. Thus, it is not clear whether the proposed strategies will generalize well to the whole wind farm.

The rest of the article is organized as follows. Sections 2 and 3 present descriptions of the wind park and the real SCADA data, respectively. Section 4 summarizes the various types of failures that the main bearing can suffer. Section 5 comprehensively delineates the fault diagnosis maintenance strategy. Section 6 showcases the results and discussion, and Section 7 lists the conclusions.

## 2. Wind Park

The wind park under study comprises 18 WTs, all with the same characteristics. The specific WT model cannot be revealed because of a non-disclosure agreement; however, some technical specifications are listed in Table 1.

**Table 1.** The wind turbines' technical parameters.

| | |
|---|---|
| Number of Blades | 3 |
| Nominal power | 2300 kW |
| Voltage | 690 V |
| Gearbox type | 3-stage planetary/helical |
| Rotor diameter | 101 m |
| Rotor speed | 6–16 rpm |
| Cut-in wind speed | 3–4 m/s |
| Rated wind speed | 12–13 m/s |
| Cut-out wind speed | 25 m/s |
| Monitoring SCADA system | WebWPS |
| Power regulation | Independent pitch |

Each WT has a three-bladed rotor and a diameter of 101 m with a sweep area of 8000 $m^2$ that generates 2300 kW. These are variable speed and pitch-controlled WTs that operate at a higher percentage of their maximal aerodynamic efficiency for a longer period of time [15]. Furthermore, variable-speed functioning helps to lower turbine loads because abrupt increases in wind energy caused by gusts can be handled by increasing rotor speed instead of by component bending [16]. The WTs' theoretical power curve is presented in Figures 1 and 2 exhibits the principal components of the WTs.

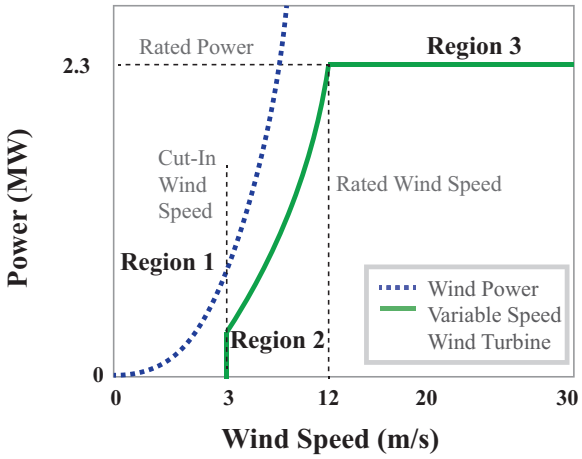


**Figure 1.** Wind turbines' operation regions.

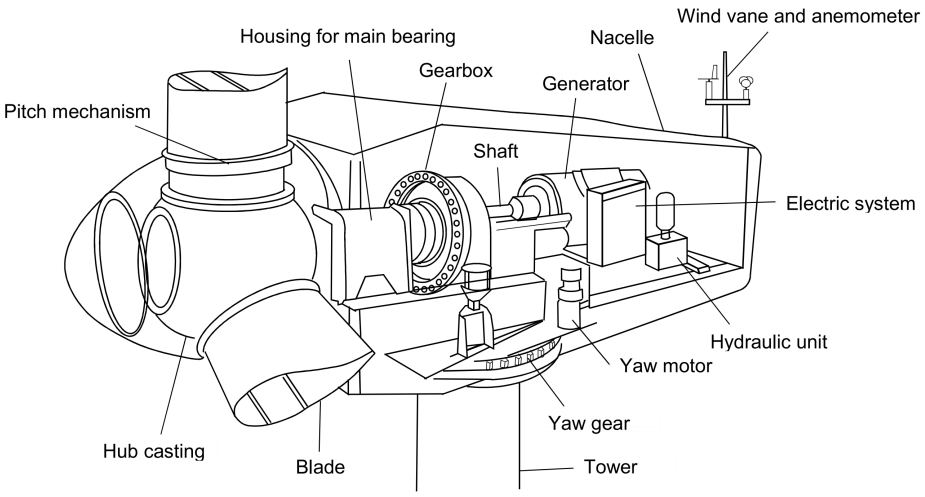


**Figure 2.** Main components of the wind turbine [17].

The WTs under consideration additionally feature a SCADA web processing server (WPS), which provides remote control as well as a range of relevant status displays and statistics. Electrical and mechanical data, operation and fault status, meteorological data, and network station data are all displayed in status displays. However, recall that, as stated in the introduction section, the SCADA system was not developed specifically for condition monitoring purposes.

## 3. Real SCADA Data Description

A collection of measurements from the wind park's SCADA systems and information from its alert log that records work orders (information about maintenance and repair operations) gathered over the same period are utilized in this article. The continuous operational data were measured between 1 January 2014 and 12 December 2019 (a period of around five years). The available SCADA data contains more than 75 different variables that can be classified into the following groups: control variables, temperature variables, environmental variables, electrical variables, and hydraulic variables (see Table 2).

**Table 2.** Data variables collected by the SCADA system.

| Variable Group | Number of Variables |
|---|---|
| Control variables | 13 |
| Component temperature variables | 35 |
| Environmental variables | 9 |
| Electrical variables | 13 |
| Hydraulic variables | 6 |

The mean, maximum, minimum, and standard deviation values of the average period of 10 min are available for each of the gathered variables. Thus, a data matrix with the following structure is obtained:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{iN} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mN} \end{pmatrix} \in \mathcal{M}_{m\times N}(\mathbb{R}), \tag{1}$$

where $m$ = 312,446 is the number of samples; $N = 304$ is the number of collected variables; and $\mathcal{M}\ (m \times N)$ is the vector space of matrices with $m$ rows and $N$ columns over $\mathbb{R}$. Finally, $x_{ij}$ indicates the acquired variable $j$ at time step $i$.

In addition to the SCADA data, an information file on maintenance and repair operations (work order logs) was accessed, as previously described. These records detail when and how the failures occurred as well as when and how the work was completed and whether the subsystem was fixed or replaced.

## 4. Main Bearing Failures

The main bearing is a major element of a WT. Figure 3 shows a spherical roller main bearing used in the WTs. It is obliged to meet stringent requirements in terms of load capacity, accuracy, noise level, friction and frictional heat, life, and dependability. This component does not always attain the necessary service life. Failures typically result in financial losses owing to lost production and contact component damage as well as the expense incurred from repairs. All of the bearing sections are susceptible to failure and can be harmed in a number of ways [18]. A bearing can fail prematurely for a number of reasons; however, there are 12 basic causes of bearing failure [19]:

- Excessive load
- Overheating
- False brinelling
- True brinelling
- Normal fatigue failure
- Reverse loading
- Contamination
- Lubricant failure
- Corrosion
- Misalignment
- Loose fits
- Tight fits

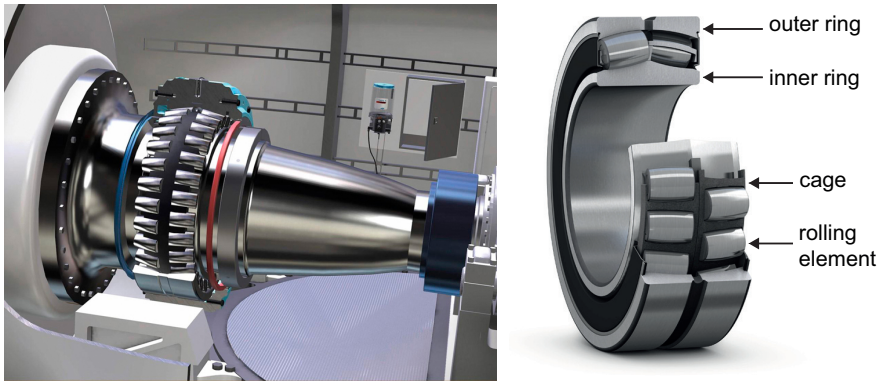


**Figure 3.** Spherical roller main bearing used in WTs. Courtesy of SKF.

Various types of bearing deterioration and failure exist, and some publications use different nomenclature for the same mode. As a result, the International Organization for Standardization (ISO) created and released a document (ISO-15243), early in 2004, that explains and categorizes the features, changes in appearance, and likely reasons for roll bearing failure in service. The failure modes are divided into six major classes (and many subgroups) in this standard: fatigue, wear, corrosion, electrical erosion, plastic deformation, and fracture and cracking. For further detailed explanation of all types of failure modes, see [16,20].

Most of these types of failures generate heat in the initial stages, which may be identified using the SCADA temperature variables relevant (closer) to the main bearing. This will be further examined in Section 5.1.2, where the feature selection step is explained.

## 5. Fault Diagnosis Strategy

This article proposes an anomaly detection approach for predicting WT's main bearing failure in advance, before the disastrous breakdown. The overall idea is to determine

whether (new) test data belongs to a certain class, determined by the training data. To cope with this problem, one-class classification methodologies are considered. By just providing normal (healthy) data to train, an algorithm creates a (representational) model of these data. If newly encountered data is different, according to some metric, from this model, it is labeled as out-of-class (faulty). In this work, the one-class SVM algorithm is utilized, see Section 5.2. It employs a collection of SCADA data to train the one-class SVM algorithm (train data set). Then, utilizing the remainder of the SCADA data, it goes on to create future predictions (test data set). In addition, it advances an indicator based on an exponential weighted moving average filter, depending on the weekly number of anomalies, to reduce the number of false positive alerts. The following steps are discussed in detail in the next subsection devoted to data preprocessing:

- Data split;
- Feature selection;
- Data cleaning;
- Data normalization.

### *5.1. Preprocessing*

When constructing a data-driven maintenance model, it is critical to ensure that the data used to train the model and the data that is inferred from the model is of high quality [21]. Data preprocessing often entails the following tasks: data cleaning, data normalization, feature extraction, feature selection, and unbalanced data handling. Furthermore, data received by real-world systems (such as data from SCADA sensors) is frequently partial, inconsistent, or inaccurate; hence, it is critical to perform data preprocessing to resolve any possible issues pertaining to the data before it is used by the model. In this study, the feature selection, data cleaning, and normalization stages are carried out. The feature selection stage is carried out with expert knowledge and is based on the relevance of the variables to the component under study. On the other hand, as it is a normality model, it is not necessary to manage unbalanced data.

#### 5.1.1. Data Split

It is critical to separate the data prior to taking any action based on it so that data leakage from the test data set into the training data set is avoided.

The data split is based on the work orders of the turbine that presented the main bearing failure, which was WT5. The work orders evidenced that the WT entered maintenance on 11 June 2018. Data from several years prior to the failure were used as training data (from 1 January 2014 to 11 December 2017). It is important to emphasize that this article is based on a one-class SVM model, which means that the model is trained with a single class of (healthy) data. Thus, in the work orders, no fault related to the main bearing was found in the period used to define the training set. Furthermore, because the data used spanned almost four years, a variety of data for each of the four seasons of the year (winter, spring, summer, and autumn) was obtained, which will aid in training a seasonal robust anomaly detection strategy. Furthermore, all regions of operation of the WT were used to train the model with the goal of obtaining a strategy able to perform in any operational region (see Figure 2). Then, to create the test data set, the data from 11 December 2017 to 12 December 2019 were used. These data will be used to make inferences from the previously trained model and to validate its efficiency. Table 3 specifies the size of the two subsets.

**Table 3.** Matrix dimensionality.

| Set | Matrix | Size |
|---|---|---|
| Training | $X_{\text{train}}$ | 208,275 × 304 |
| Test | $X_{\text{test}}$ | 104,171 × 304 |

5.1.2. Feature Selection

The data set has 304 variables, as stated in Section 3; however, when studying a specific failure, experts must skillfully select the most significant variables for the physical system to be investigated [22]. As a result, only the mean values for the main shaft temperature (the temperature variable most closely associated to the examined failure) and the mean values for the following exogenous variables were chosen for this study:

- Ambient temperature;
- Primary wind speed;
- Secondary wind speed;
- Actual wind speed;
- Anemometer measure.

The main objective is that the model must be insensitive to other faults that are not closely connected to the component of interest. To accomplish this aim, the model uses only exogenous variables and the main shaft temperature (the closest to the main bearing).

A correlation matrix analysis is performed to determine which of the exogenous variables contributes the most to the model. There are different distinct types of correlations, such as the Spearman, Kendall, and Pearson correlations, among which the present study employs the Pearson correlation to establish the linear connection between the variables [23]. Numbers close to +1 suggest a significantly positive correlation, whereas values nearer to −1 represent a significantly negative correlation. The findings of the correlation analysis are compiled in Table 4. It can be observed that, except for the ambient temperature, all other factors have a high connection. As a result, another method is required for determining the preferred variable from among the associated variables.

**Table 4.** Correlation analysis.

| | **wtc_AmbieTmp** | **wtc_PrWindSp** | **wtc_SeWindSp** | **wtc_AcWindSp** | **wtc_SecAnemo** |
|---|---|---|---|---|---|
| wtc_AmbieTmp | 1 | − 0.20 | −0.08 | −0.20 | −0.08 |
| wtc_PrWindSp | −0.20 | 1 | 0.94 | 0.99 | 0.94 |
| wtc_SeWindSp | −0.08 | 0.94 | 1 | 0.94 | 0.99 |
| wtc_AcWindSp | −0.20 | 0.99 | 0.94 | 1 | 0.94 |
| wtc_SecAnemo | −0.08 | 0.94 | 0.99 | 0.94 | 1 |

The variables with the largest variance (regarding the variable most associated with the main bearing failure, wtcMainTmp) are selected using a variance comparison feature selector. The variance for each characteristic is shown in Figure 4.

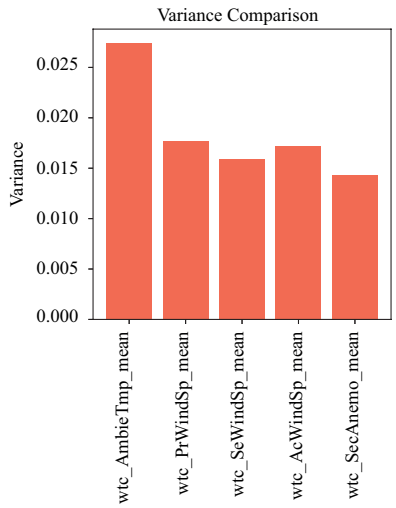


**Figure 4.** Variance threshold for ambient temperature, primary wind speed, secondary wind speed, actual wind speed, and secondary anemometer measure.

Finally, three variables are chosen for inclusion in the proposed model. The first two variables have no physical link with the physical components of the WT, as it is not desirable that the model learns to detect failures other than the one being examined.

The mean main wind speed and the mean ambient temperature are the two exogenous variables. The benefit of utilizing a variable linked to wind speed is that it is also tied to the turbine's operating region (see Figure 1). On the other hand, the variable mean ambient temperature offers information regarding seasonality, as the temperature might vary with the seasons of the year. Finally, the model's third variable is the mean main shaft temperature measurement, which is the variable most closely connected to the main bearing failure. Table 5 describes the new dimensionality of the two different subsets after the feature selection process.

**Table 5.** Matrix dimensionality after feature selection.

| Set | Matrix | Matrix Dimensionality |
|---|---|---|
| Training | $X_{\text{train}}$ | 208,275 × 3 |
| Test | $X_{\text{test}}$ | 104,171 × 3 |

### 5.1.3. Data Cleaning

It is important to clean the real data before training a model because it can be inconsistent, noisy, and incomplete [24].

Missing values in a data set can be treated in several ways, one of which is by ignoring the records and another is by filling in the missing values (manually, by constant, by mean, by most probable value, etc.). In this article, the missing values are treated as indicated in [16], a single imputation by the piecewise cubic Hermite polynomial interpolation [25]. For the missing values at the start and end of the data set, the closest nonmissing value is employed to fill in the value.

As indicated in [22], outlier values are not always systematically eliminated, as this could cause information related to failure prognosis to be lost. For this reason, the current study first creates handcrafted specified ranges based on feasible values of the sensors' signals (see Table 6). Then, the values that are outside these ranges are replaced by missing values and treated in a similar way.

After the imputation stage, the mean ambient temperature is subtracted from the mean shaft temperature to avoid the problem of seasonality [26]. Table 7 describes the dimensionality of the two different subsets, training and test, after feature selection.

**Table 6.** Selected SCADA variables.

| Variable Name | Variable Description | Range | Units |
|---|---|---|---|
| wtc_MainBTmp | Mean main shaft temperature | [0, 120] | °C |
| wtc_AmbienTmp | Mean ambient temperature | [− 5, 40] | °C |
| wtc_PrWindSp | Mean primary wind speed | [0, 60] | m/s |

**Table 7.** Matrix dimensionality after the preprocessing stage.

| Set | Notation | Matrix Dimensionality |
|---|---|---|
| Training | $X_{\text{train}}$ | 208,275 × 2 |
| Test | $X_{\text{test}}$ | 10,471 × 2 |

### 5.1.4. Data Normalization

Finally, each of the variables that will be utilized in the model must be normalized. Otherwise, the model's output could be skewed towards large-scale data [21]. Furthermore, normalization aids model training by making it easier to learn a single, rather than numerous, data distribution. Hence, standardizing the data range of SCADA sensors is an

important step in data preparation. In this article, the z-score normalization is employed, which is computed as

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{2}$$

where $\mu_j$ and $\sigma_j$ are the mean and the standard deviation of column $j$, respectively, in the training set. Note that $x_{ij}$ is measurement at time step $i$ at sensor $j$, and $z_{ij}$ is its normalized value.

### *5.2. One-Class Support Vector Machine (OCSVM)*

By tackling a quadratic optimization problem, the SVM technique described by Vapnik [27] has demonstrated great performance for classification problems. This is why, as the fundamental SVM paradigm recommends, a vast majority of condition monitoring research uses both healthy and failure samples to train the model. However, in many cases where negative (failure) samples are either unavailable or impossible to collect, such as with WT systems, this technique is not practicable. OCSVM is a variant of SVM that may be used to address this problem [28], as it is a semi-supervised algorithm; that is, it is trained only on normal data, such as healthy samples in the present scenario, and it computes the surface of a minimum hyper-sphere (decision boundary) containing all normal data (healthy data). Then, after being trained, it classifies any points outside the hyper-sphere as anomalies (failure-related data). The decision boundary is computed in a high-dimensional feature space, $H$, using a suitable kernel function. In the present study, the radial basis function (RBF) kernel is used. The fundamental goal of a OCSVM is to categorize data:

$$f(z_i) = \begin{cases} 0, & \text{if } z_i \in S \\ 1, & \text{if } z_i \notin S, \end{cases} \tag{3}$$

where $S$ is a hyper-sphere of a high dimensional feature space $H$, $z_i$ is a given sample, $f(z_i) = 0$ corresponds to the sample being classified as healthy, and $f(z_i) = 1$ corresponds to being classified as an anomaly. In the current context, $z_1, z_2, \ldots z_l$ are training samples belonging to the healthy class, $Z$; and $l$ is the number of training samples. Next, given the kernel $\Phi : Z \to H$, the training is accomplished by solving the following quadratic programming problem:

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho, \tag{4}$$

subject to

$$w^T \cdot \Phi(z_i) \geq \rho - \xi_i \quad i = 1, 2, \ldots, l \quad \xi_i \geq 0, \tag{5}$$

where $\xi_i > 0$ are the slack variables, $v \in (0, 1]$ is a parameter that specifies an upper bound on the fraction of outliers (training points outside the estimated region), see [29]. Assuming $w$ and $\rho$ are the optimized parameters,

$$f(z_i) = \text{sign}((w^T \cdot \Phi(z_i)) - \rho) \tag{6}$$

will be positive for the vast majority of samples in the training set. Figure 5 summarizes the stages of the proposed methodology up to the training of the SVM model.

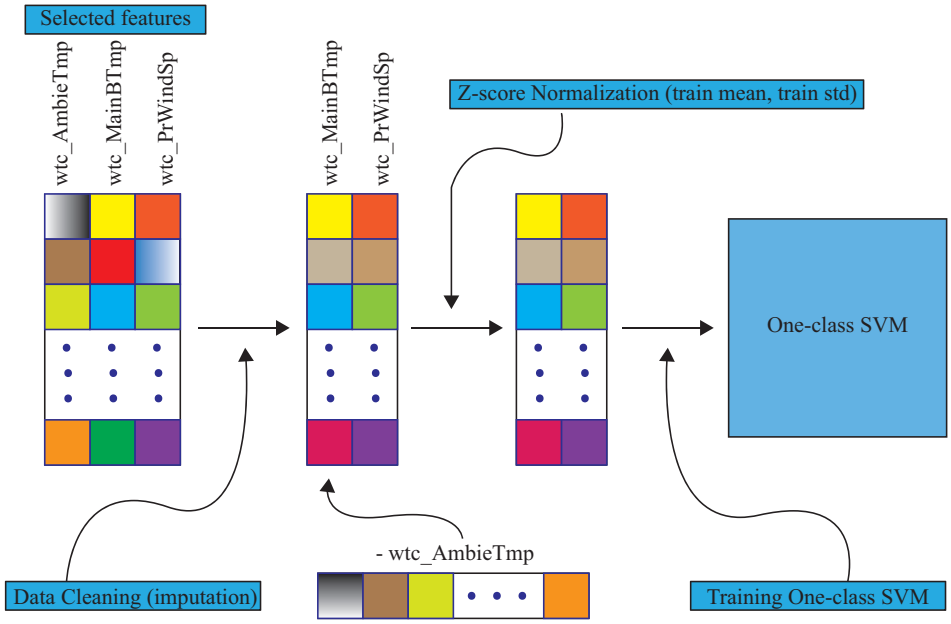


**Figure 5.** Data preprocessing and training of the one-class SVM.

### *5.3. Inference Stage*

The inferences are performed on the test data set $X_{\text{test}}$ once the model has been trained. As previously stated, the model's output must learn to distinguish healthy samples from anomalies.

### *5.4. Fault Prognosis Indicator (FPI)*

Generally, a threshold is set for FPIs. An alert is generated when the value of anomalies (outliers) discovered by the model exceeds the set threshold. However, because 10 min samples were utilized in this study, an overwhelming amount of false alarms might arise, therefore rendering the approach worthless. This section explains the three steps involved in computing a fault prognosis indicator (FPI) that is proposed to avoid this issue. The one-class SVM technique is used to aggregate (count) outliers weekly (see Section 5.4.1), following which the exponential weighted moving average (EWMA) filter is applied to the weekly count of outliers, and lastly, a threshold is set.

#### 5.4.1. Weekly Grouping

In the present study, the SCADA samples are obtained every ten minutes; thus, 1008 samples are acquired per week. The weekly grouping entails keeping track of the number of samples, out of the total of 1008, that are categorized as anomalies by the one-class SVM algorithm per week. Stated formally, given the $i$-th week, the number of samples classified as anomalies in that week are counted and noted as $C_i$.

#### 5.4.2. EWMA Filter

The moving average (MA) method is a well-known procedure used to smooth historical data [30]. It takes a time series and defines a fixed subset size. Taking the average of the initial fixed subset of the sequence of numbers yields the first element of the moving average. The subset is then updated as it moves ahead, with the initial number in the series being omitted and the next value in the subset being included. By calculating the MA, the impacts of random short-term fluctuations over a specified period of time are mitigated (in the study case, false positives). The MA has various extensions, each one with its respective characteristics, but their underlying purpose remains the same. The easiest type is the

simple moving average (SMA), which averages the previous $n$ data points in time series data as follows:

$$\mathrm{SMA}_M = \frac{P_M + P_{M-1} + \cdots + P_{M-(n-1)}}{n}, \tag{7}$$

where $P_M$ is the time series value at time $M$, and $n$ is the number of samples used in the calculation.

Another typical MA extension is the weighted moving average (WMA), which improves the behavior of the SMA by giving more relevance (weight) to recent data:

$$\mathrm{WMA}_M = \frac{nP_M + (n-1)P_{M-1} + \cdots + P_{(M-n+1)}}{n + (n-1) + \cdots + 1}. \tag{8}$$

Furthermore, the exponential moving average (EMA) assigns a weight factor to each sample depending on its seniority. The EMA can be calculated recursively as follows:

$$\mathrm{EMA}_1 = P_1, \tag{9}$$

$$\text{for } t > 1, \mathrm{EMA}_t = \alpha \cdot P_t + (1-\alpha) \cdot \mathrm{EMA}_{t-1}, \tag{10}$$

where $P_t$ is the value at time $t$, $\mathrm{EMA}_t$ is the EMA value at $t$, and $\alpha$ is the decrease weight degree (factor between 0 and 1) calculated as

$$\alpha = \frac{2}{n+1}. \tag{11}$$

Finally, an approach that integrates the computation of weight factor for WMA and EMA, called weighted exponential moving average (EWMA), is used in this study [31]. An EWMA reacts more significantly to recent sample changes than a WMA, which applies equal weight to all observations in the period. The EWMA has one parameter, $\alpha$, to be defined. This parameter is related to the importance of the current point in the EWMA computation. The greater the value of $alpha$, the better the EWMA follows the initial time series. The EWMA's formula used in this study is applied to the weekly grouping time series, $C_i$, as

$$\mathrm{EWMA}_t = \alpha C_t + (1-\alpha)\mathrm{EWMA}_{t-1}, \tag{12}$$

where $\mathrm{EWMA}_0$ is the mean of historical data, and $\alpha$ is the weight decided by the user. The parameter $\alpha$ can be defined in terms of spans, $s$, commonly understood as an $n$-day EW moving average,

$$\alpha = \frac{2}{s+1}. \tag{13}$$

For the present study, $s = 4$, which means that it considers 4-week groups (around a month). Actually, this selection is influenced by the findings of McKinnon et al. [32]. Their research, on the influence of time history on WT failures using SCADA data, tests three distinct moving windows: daily, weekly, and monthly. In comparison to the others, the weekly moving window has the best performance in identifying failures. On the one hand, a daily window contains too much noise, leading to a large percentage of false alarms. On the other hand, a monthly window removes much information and does not allow any specification of when an anomaly occurred. Finally, note that the EWMA is a recursive function.

Finally, the FPI activates an alert when the EWMA of the number of anomalies is higher than a prescribed threshold. To define this threshold, first the training data set is passed through the model. Then, the weekly counting and the EWMA filter are applied. The mean ($\mu$) and the standard deviation ($\sigma$) of the EWMA over the training set are calculated. Recall that, for an approximately normally distributed data set, the values within one standard deviation of the mean account for about 68% of the set; whereas within two standard deviations account for about 95%; and within three standard deviations account for about 99.7%. These percentages are rounded theoretical probabilities intended only

to approximate the empirical data derived from a normal population. Thus, the so-called three-sigma rule (or $3\sigma$ rule) expresses a conventional heuristic that nearly all values are taken to lie within three standard deviations of the mean, and thus it is empirically useful to treat 99.7% probability as near certainty. Thus, in this work, the threshold indicates that values above three standard deviation of the mean should be considered as anomalies.

Therefore, the threshold is declared as

$$\text{threshold} = \mu + 3\sigma. \tag{14}$$

The final steps of the methodology (from the already trained SVM model) are described in the flowchart given in Figure 6. The diagram explains how the $X_{\text{train}}$ inferences are grouped by weeks and then filtered to calculate a threshold. Then, the same process is carried out with the $X_{\text{test}}$ inferences, not to calculate a new threshold but to use the already defined threshold (with $X_{\text{train}}$) to trigger an alarm when the outputs trespass it.

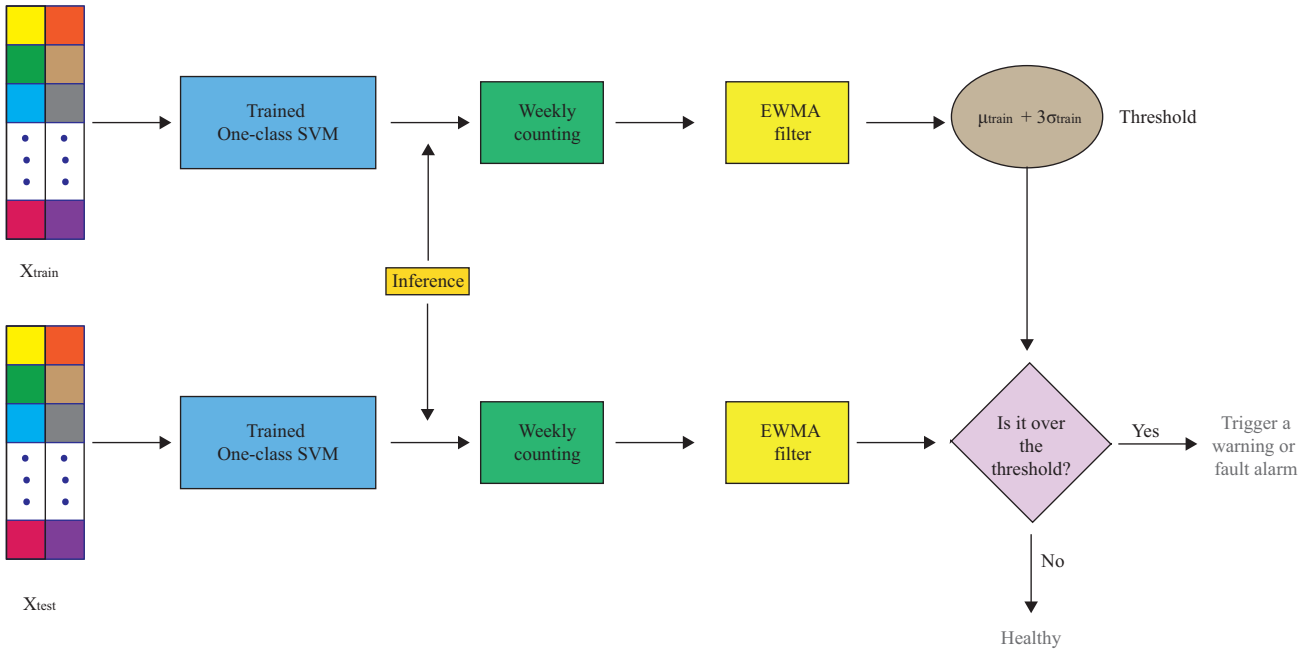


**Figure 6.** Final steps of the proposed methodology (from the already trained SVM model).

## 6. Results and Discussion

This section details and analyzes the results of the proposed approach over the entire wind park.

The inferences generated by the one-class SVM model from the test data of two adjacent turbines impacted by similar wind speeds are shown in Figure 7. Recall that thanks to the work orders, it is known that WT5 had the failure of interest to this study and a main bearing repair scheduled for 11 June 2018, whereas WT6 had no working orders related to this fault type. It can be observed, in Figure 7, that WT5 had many samples (obtained every 10 min) labeled as anomalies prior to the repair date. On the other hand, WT6 had a relatively small number of samples classified as anomalies.

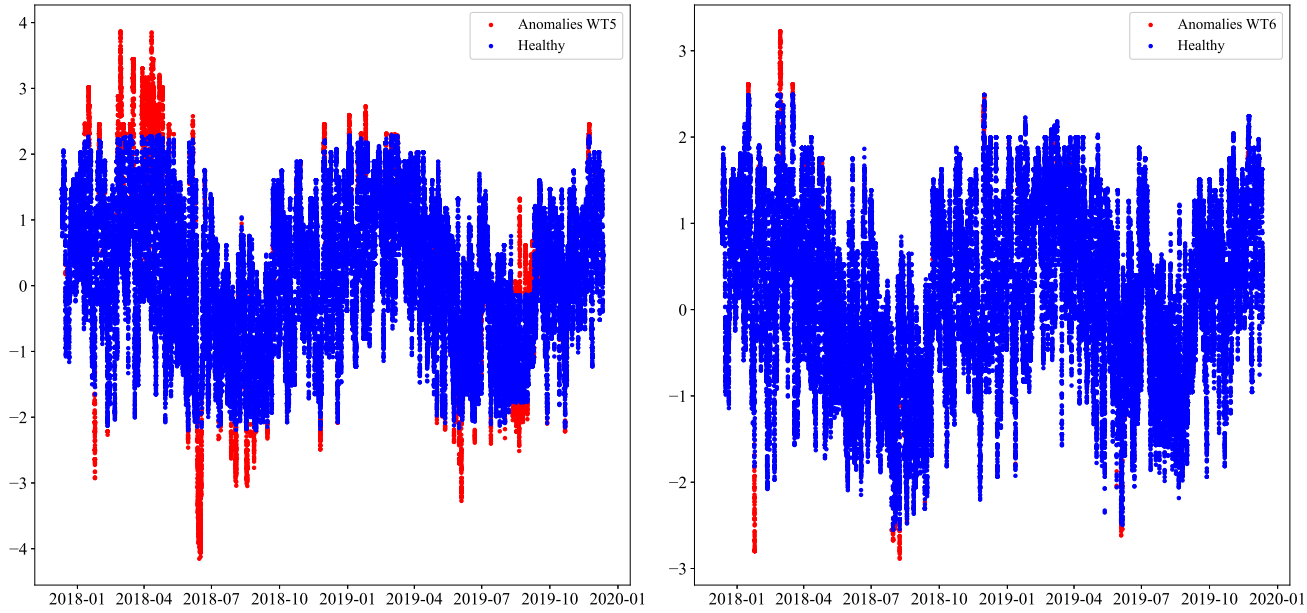


**Figure 7.** Mean main shaft temperature samples classified as healthy (blue) or anomaly (red) by the one-class SVM algorithm for WT5 (**left**) and WT6 (**right**).

The next step involved counting the samples designated as anomalies on a weekly basis. The acquired result for WT5 and WT6 is shown in Figure 8. It can be observed that WT5 exhibited many anomalies in successive weeks before the component repair date. WT6, on the other hand, had far fewer weekly anomalies. Although this weekly grouping is a strong indication of a pre-failure in the main bearing, there were still a high number of false positives that had to be addressed. The EWMA was applied to the weekly counting to alleviate this problem.

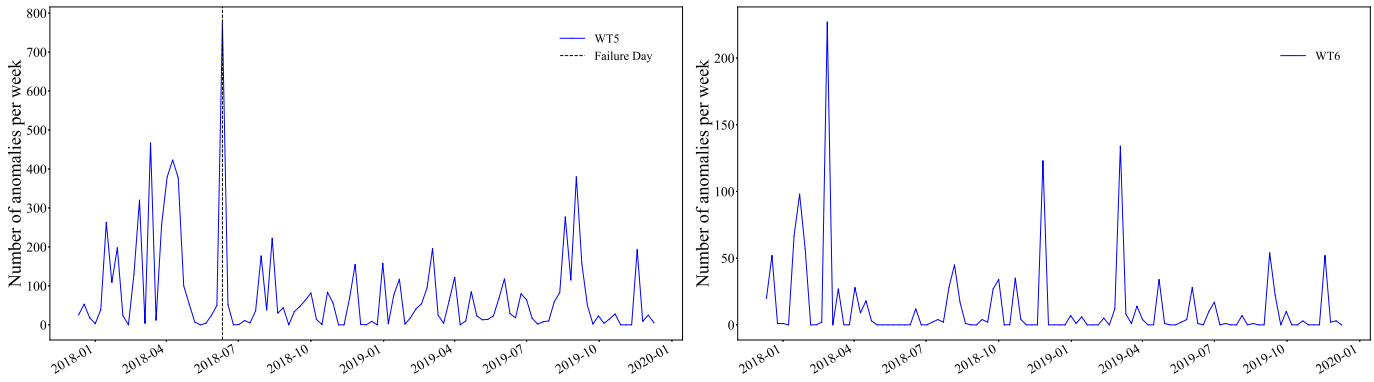


**Figure 8.** Number of anomalies per week for WT5 (**left**) and WT6 (**right**).

Figure 9 shows the results obtained with the proposed FPI for the whole wind park composed of 18 WTs over the test data set. On 15 October 2018, two days before a blade failure, the model set off an alarm at WT4. The method identified it, even though it was not a failure of interest that could be forecast well in advance. At WT5, which presented the failure of interest (main bearing) on 11 June 2018, the model raised an alarm for the first time in March, followed by three times in April, and once on the day of failure. Thus, the proposed strategy detects the fault months before it happens. Additionally, note that, for WT5, there is a clear trend of residuals increasing and then decreasing. When bearing failure initiates (or develops), there is usually a brief heat release rendered as temperature increasing. As it is stated in [20], almost all bearing failure modes (excessive current, fatigue fracture, thermal cracking, etc.) are driven by unforeseen heat release. After that, the

temperature goes back to normal, i.e., crack is not growing. It is crucial that despite the residuals going back under the threshold, the triggered alarm must be kept active. Thus, the methodology's approach is to predict this typical heat release in advance (to raise an alert when there is a potential abnormal state) before the bearing is entirely damaged. At WT17, a failure occurred in the gearbox on 29 May 2019, and it was repaired on 7 June of the same year according to the work order. This was not the failure of interest, but it was also predicted by the strategy on different occasions: three times in April and four times in May. The activation alerts were presented several months in advance because the failure in this element can be transmitted to the main bearing as they are closely connected. Therefore, the strategy is capable of detecting failures at components close to the main bearing. It is noteworthy that the remaining WTs in the wind park were accurately predicted as healthy, with no false alarms. This is because the weekly counting of anomalies with the EWMA filter effectively avoided false alarms, which is one of the main difficulties encountered by this type of strategy.

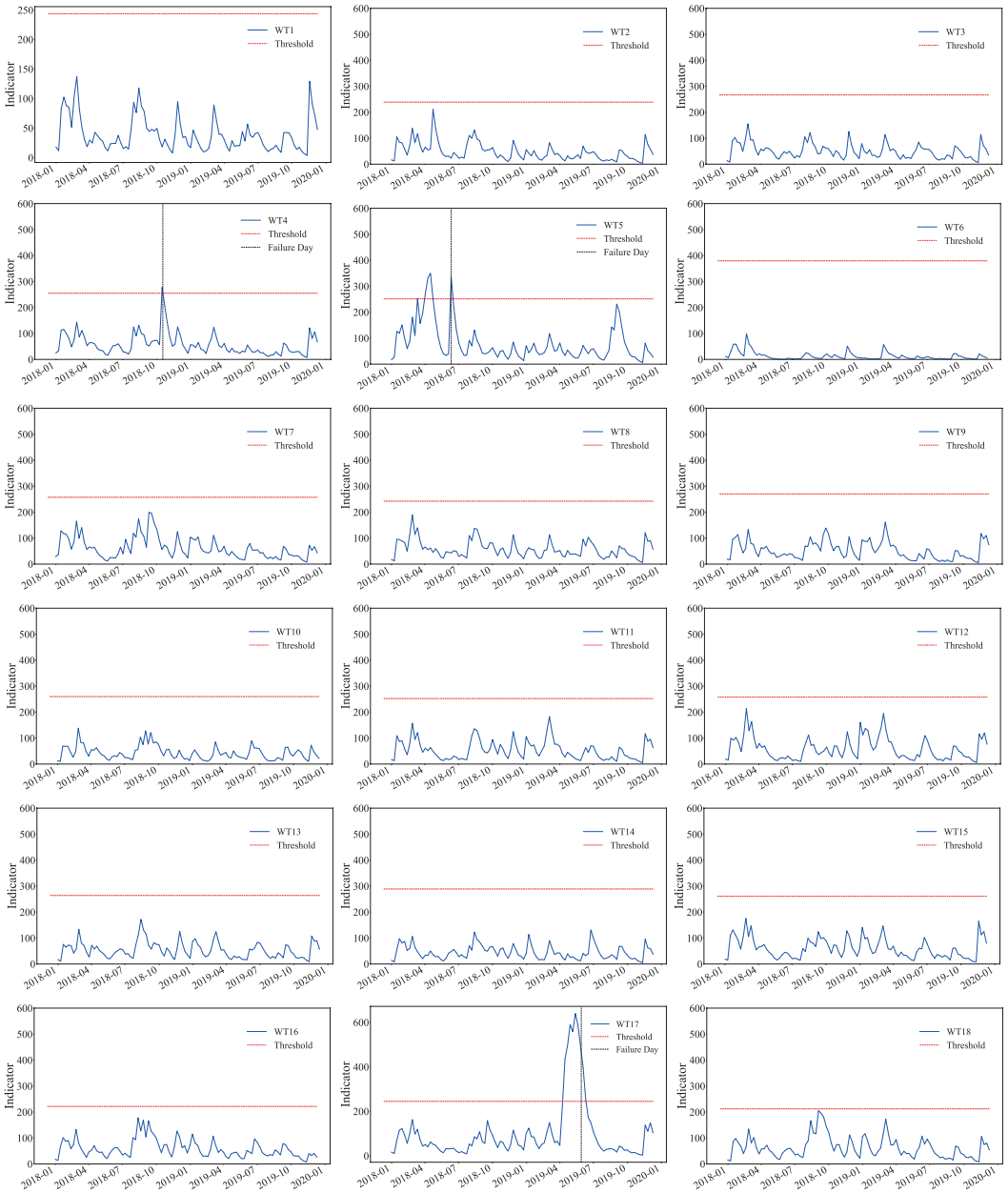


**Figure 9.** Fault prognosis indicator (FPI) results over 18 in-production wind turbines. The red horizontal line is the threshold value. The black vertical line indicates the main bearing fault.

## 7. Conclusions

A methodology for early fault diagnosis of WT main bearings was presented in this study. In particular, it proposed an anomaly detection model based only on wind speed, ambient temperature, and main bearing temperature (given by the SCADA data). The key advantages of the proposed strategy are listed below:

- The model is insensitive to other faults that are not closely connected to the component of interest, because it only uses exogenous variables and the main shaft temperature (the temperature variable most closely associated to the examined failure);
- The methodology is robust to seasonality and operating and environmental conditions;
- It does not require historical faulty data to construct the model;
- There is no need to install extra sensors, as only SCADA data (already available in all industrial-sized WTs) is used;
- The methodology can be employed even when historical faulty data is unavailable;
- The results, obtained in an in-production wind park with 18 WTs, show that there are very few false alarms, and for the fault of interest, an alarm is triggered several months in advance.

The wind park data employed in this work include gearbox faults. Hence, future work will tackle this system. The gearbox comprises several components (bearings, pinions, and gears) and has a wide variety of fault scenarios. Furthermore, in WTs, it is a system that frequently fails prematurely, and its maintenance is costly. Thus, a predictive maintenance strategy capable of coping with the various faulty scenarios for this system would help to reduce costs associated with wind park operation and maintenance.

**Author Contributions:** Conceptualization, C.T. and Y.V.; Data curation, C.T., A.I. and L.C.-V.; Formal analysis, C.T., A.I., L.C.-V., Y.V. and W.A.; Funding Acquisition, Y.V.; Investigation, C.T., A.I., L.C.-V., Y.V. and W.A.; Methodology, C.T., A.I., L.C.-V. and Y.V.; Project administration, Y.V.; Resources, Y.V.; Software, A.I. and L.C.-V.; Supervision, Y.V. and C.T.; Validation, C.T.; Writing—original draft, C.T., A.I. and L.C.-V.; Writing—review and editing, Y.V. and C.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been partially funded by the Spanish Agencia Estatal de Investigación (AEI)—Ministerio de Economía, Industria y Competitividad (MINECO), and the Fondo Europeo de Desarrollo Regional (FEDER) through the research project DPI2017-82930-C2-1-R; and by the Generalitat de Catalunya through the research project 2017 SGR 388.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this work are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Whiteman, A.; Akande, D.; Elhassan, N.; Escamilla, G.; Lebedys, A.; Arkhipova, I. *Renewable Capacity Statistics 2021;* Technical Report; International Renewable Energy Agency (IRENA): Abu Dhabi, United Arab Emirates, 2021.
2. Costa, Á.M.; Orosa, J.A.; Vergara, D.; Fernández-Arias, P. New Tendencies in Wind Energy Operation and Maintenance. *Appl. Sci.* **2021**, *11*, 1386. [CrossRef]
3. Hameed, Z.; Ahn, S.H.; Cho, Y.M. Practical aspects of a condition monitoring system for a wind turbine with emphasis on its design, system architecture, testing and installation. *Renew. Energy* **2010**, *35*, 879–894. [CrossRef]
4. Tchakoua, P.; Wamkeue, R.; Tameghe, T.A.; Ekemb, G. A review of concepts and methods for wind turbines condition monitoring. In Proceedings of the 2013 World Congress on Computer and Information Technology (WCCIT), Sousse, Tunisia, 22–24 June 2013; pp. 1–9.
5. Luo, N.; Vidal, Y.; Acho, L. *Wind Turbine Control and Monitoring*; Springer: Berlin/Heidelberg, Germany, 2014.

6. Asgarpour, M.; Sørensen, J.D. Bayesian based diagnostic model for condition based maintenance of offshore wind farms. *Energies* **2018**, *11*, 300. [CrossRef]
7. Scheu, M.N.; Tremps, L.; Smolka, U.; Kolios, A.; Brennan, F. A systematic Failure Mode Effects and Criticality Analysis for offshore wind turbine systems towards integrated condition based maintenance strategies. *Ocean. Eng.* **2019**, *176*, 118–133. [CrossRef]
8. Puruncajas, B.; Vidal, Y.; Tutivén, C. Vibration-Response-Only Structural Health Monitoring for Offshore Wind Turbine Jacket Foundations via Convolutional Neural Networks. *Sensors* **2020**, *20*, 3429. [CrossRef] [PubMed]
9. Baboli, P.T.; Raeiszadeh, A.; Babazadeh, D.; Meiners, J. Two-Stage Condition-based Maintenance Model of Wind Turbine: From Diagnosis to Prognosis. In Proceedings of the 2020 IEEE International Smart Cities Conference (ISC2), Virtual, 28 September–1 October 2020; pp. 1–6.
10. Sandoval, D.; Leturiondo, U.; Vidal, Y.; Pozo, F. Entropy indicators: An approach for low-speed bearing diagnosis. *Sensors* **2021**, *21*, 849. [CrossRef] [PubMed]
11. Beretta, M.; Cárdenas, J.J.; Koch, C.; Cusidó, J. Wind Fleet Generator Fault Detection via SCADA Alarms and Autoencoders. *Appl. Sci.* **2020**, *10*, 8649. [CrossRef]
12. Elasha, F.; Shanbr, S.; Li, X.; Mba, D. Prognosis of a wind turbine gearbox bearing using supervised machine learning. *Sensors* **2019**, *19*, 3092. [CrossRef] [PubMed]
13. McKinnon, C.; Carroll, J.; McDonald, A.; Koukoura, S.; Infield, D.; Soraghan, C. Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data. *Energies* **2020**, *13*, 5152. [CrossRef]
14. Jin, X.; Xu, Z.; Qiao, W. Condition monitoring of wind turbine generators using SCADA data analysis. *IEEE Trans. Sustain. Energy* **2020**, *12*, 202–210. [CrossRef]
15. Pao, L.Y.; Johnson, K.E. Control of wind turbines. *IEEE Control. Syst. Mag.* **2011**, *31*, 44–62.
16. Encalada-Dávila, Á.; Puruncajas, B.; Tutivén, C.; Vidal, Y. Wind Turbine Main Bearing Fault Prognosis Based Solely on SCADA Data. *Sensors* **2021**, *21*, 2228. [CrossRef] [PubMed]
17. Jiang, Z.; Hu, W.; Dong, W.; Gao, Z.; Ren, Z. Structural reliability analysis of wind turbines: A review. *Energies* **2017**, *10*, 2099. [CrossRef]
18. Hamadache, M.; Lee, D. Wind turbine main bearing fault detection via shaft speed signal analysis under constant load. In Proceedings of the 2016 16th International Conference on Control, Automation and Systems (ICCAS), Gyeongju, Korea, 16–19 October 2016; pp. 1579–1584.
19. Bearings, B.P. *Bearing Failure: Causes and Cures*; Technical Report; The Barden Corporation: Danbury, CT, USA, 2008.
20. Bearing Damage and Failure Analysis. 2017. Available online: https://www.skf.com/binaries/pub12/Images/0901d1968064c148-Bearing-failures---14219_2-EN_tcm_12-297619.pdf (accessed on 8 July 2021).
21. Kang, M.; Tian, J. Machine Learning: Data Pre-processing. In *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*; IEEE: Piscataway, NJ, USA, 2018; pp. 111–130.
22. Martí-Puig, P.; Blanco-M, A.; Cárdenas, J.J.; Cusidó, J.; Solé-Casals, J. Effects of the pre-processing algorithms in fault diagnosis of wind turbines. *Environ. Model. Softw.* **2018**, *110*, 119–128. [CrossRef]
23. Kumar, S.; Chong, I. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2907. [CrossRef] [PubMed]
24. Hossen, M.S. Data preprocess. In *Machine Learning and Big Data: Concepts, Algorithms, Tools and Applications*; Wiley: Hoboken, NJ, USA , 2020; pp. 71–103.
25. Lu, S.; Wang, Y.; Wu, Y. Novel high-precision simulation technology for high-dynamics signal simulators based on piecewise Hermite cubic interpolation. *IEEE Trans. Aerosp. Electron. Syst.* **2018**, *54*, 2304–2317. [CrossRef]
26. Zhang, Z. Automatic fault prediction of wind turbine main bearing based on SCADA data and artificial neural network. *Open J. Appl. Sci.* **2018**, *8*, 211–225. [CrossRef]
27. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
28. Chen, Y.; Zhou, X.S.; Huang, T.S. One-class SVM for learning in image retrieval. In Proceedings of the 2001 International Conference on Image Processing (Cat. No. 01CH37205), Thessaloniki, Greece, 7–10 October 2001; Volume 1, pp. 34–37.
29. Schölkopf, B.; Platt, J.C.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural Comput.* **2001**, *13*, 1443–1471. [CrossRef]
30. Hansun, S. A new approach of moving average method in time series analysis. In Proceedings of the 2013 Conference on New Media Studies (CoNMedia), Tangerang, Indonesia, 27–28 November 2013; pp. 1–4.
31. Hunter, J.S. The exponentially weighted moving average. *J. Qual. Technol.* **1986**, *18*, 203–210. [CrossRef]
32. McKinnon, C.; Turnbull, A.; Koukoura, S.; Carroll, J.; McDonald, A. Effect of time history on normal behaviour modelling using SCADA data to predict wind turbine failures. *Energies* **2020**, *13*, 4745. [CrossRef]