

Article

SCADA Data Analysis Methods for Diagnosis of Electrical Faults to Wind Turbine Generators

Francesco Castellani , Davide Astolfi and Francesco Natili 

Department of Engineering, University of Perugia, Via G. Duranti 93, 06125 Perugia, Italy;
davide.astolfi@studenti.unipg.it (D.A.); francesco.natili@studenti.unipg.it (F.N.)

* Correspondence: francesco.castellani@unipg.it

Featured Application: Fault diagnosis of wind turbine generators.

Abstract: The electric generator is estimated to be among the top three contributors to the failure rates and downtime of wind turbines. For this reason, in the general context of increasing interest towards effective wind turbine condition monitoring techniques, fault diagnosis of electric generators is particularly important. The objective of this study is contributing to the techniques for wind turbine generator fault diagnosis through a supervisory control and data acquisition (SCADA) analysis method. The work is organized as a real-world test-case discussion, involving electric damage to the generator of a Vestas V52 wind turbine sited in southern Italy. SCADA data before and after the generator damage have been analyzed for the target wind turbine and for reference healthy wind turbines from the same site. By doing this, it has been possible to formulate a normal behavior model, based on principal component analysis and support vector regression, for the power and for the voltages and currents of the wind turbine. It is shown that the incipience of the fault can be individuated as a change in the behavior of the residuals between model estimates and measurements. This phenomenon was clearly visible approximately two weeks before the fault. Considering the fast evolution of electrical damage, this result is promising as regards the perspectives of exploiting SCADA data for individuating electric damage with an advance that can be useful for applications in wind energy practice.

Keywords: wind energy; wind turbines; renewable energy; fault diagnosis; condition monitoring; generators; SCADA data



Citation: Castellani, F.; Astolfi, D.; Natili, F. SCADA Data Analysis Methods for Diagnosis of Electrical Faults to Wind Turbine Generators. *Appl. Sci.* **2021**, *11*, 3307. <https://doi.org/10.3390/app11083307>

Academic Editor: Mohsen N. Soltani

Received: 15 March 2021

Accepted: 5 April 2021

Published: 7 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Operation and maintenance activities can reach up to 25% of the overall costs of a wind farm project [1] and this percentage can reach 35% for offshore installations [2]. This in general motivates the widespread interest in wind turbine condition monitoring [3,4], because the ability to adequately plan interventions for wind turbines is fundamental for diminishing the producible energy losses.

In [5], it is reported that for a typical 2 MW wind turbine the cost of the generator is in the order of 10% of the total component cost and that generator failure is among the most impacting factors as regards the number of producible hours lost (around 200). The electric generator is estimated to be among the top three contributors to the failure rates and the downtime of wind turbines, according to [6,7]. The above matters of fact motivate the importance of developing reliable methodologies for early detection of wind turbine generator damage and for evaluation of their health status.

Faults related to wind turbine generators are particularly difficult to diagnose in real-world applications because they evolve quickly in an uncontrolled environment with variable operation conditions. This consideration includes mechanical damage to rotating elements, for example generator bearings, and of course includes electric damage to generator components.

Furthermore, in wind energy practice, the most employed information source is constituted by supervisory control and data acquisition (SCADA) data, which have a sampling time of minutes (typically ten). The main drawback is therefore that this time scale is definitely non-optimal, in particular for the diagnosis of electrical faults. In order to acquire electrical measurements with the appropriate sampling frequency (in the order of kHz, as is done for example in [8,9]), it is at present unavoidable to intervene on site and it is inconceivable to do this on a condition basis. Consequently, the standard is that the inspections of wind turbine generators are periodic and therefore poorly related to the incipience of possible faults.

On the grounds of the above considerations, it would be appreciable to develop SCADA data analysis methodologies that could be helpful to evaluate the health status of wind turbine generators. This is exactly the objective of the present study, which is organized as a real-world test-case discussion.

The innovative aspects of the present study can be appreciated in light of a brief discussion of the literature about wind turbine generator fault diagnosis, from which it arises that most studies deal with the use of SCADA data regarding mechanical faults. As often happens when SCADA data are employed for this aim [10–14], the analysis of sub-component temperatures is useful. In [15–17], the targets are the stator winding temperature, the generator bearing temperature, and the generator slip ring temperatures. In [18], a test case of generator damage (rotor winding failure) is analyzed: the diagnosis is based on a dynamic model sensor method representing the relationship between the generator temperature, wind speed, and ambient temperature.

Two inspiring studies for the purposes of the present work are [19,20]. In [19], a series of phenomena possibly related to wind turbine generator damage are listed and these can be individuated through SCADA data analysis. In addition to anomalous heating, these are miscorrelation between generator speed and produced power, or reactive power, and anomalies in the shaft torque. In [20], it is shown that real-world generator damage can be diagnosed by analyzing the Mahalanobis distance and the correlation matrix of a set of features.

As discussed above, the present study aims at contributing to the topic of SCADA data analysis methods for wind turbine generator fault diagnosis. Thanks to the support of the Lucky Wind spa company, which provided the data sets employed in this study, it has been possible to investigate the behavior of a Vestas V52 wind turbine before and after electrical damage occurred at the generator. The most important operation variables (such as blade pitch, rotational speed, and so on) and electrical parameters have been analyzed; in particular, on the grounds of the above literature discussion, it arises that a relevant point of novelty of this study is the use of electrical parameter SCADA measurements. Actually, in this work it is shown that it is possible to construct data-driven normal behavior models describing the relation between electrical parameters and operation variables and these models are responsive in individuating incipient electrical damages. The normal behavior model is constructed through support vector regression (SVR) with a Gaussian kernel because this has been shown to be effective for tackling the typical non-linear problems arising in wind energy data practice. The features are orthogonalized and reduced in dimension through principal component analysis (PCA).

To summarize, in this work a reference healthy wind turbine and the target damaged one are analyzed in parallel, and it is shown that it is possible to distinguish the damaged wind turbine with respect to the healthy one when the fault is incipient (in the order of two weeks before the fault) and that, after the replacement of the generator, the observations are compatible with the normal behavior model.

The organization of this work is as follows: Section 2 is devoted to the description of the test cases and of the data sets, in Section 3 the methods are described; the results are collected and discussed in Section 4; finally, conclusions are drawn in Section 5.

2. Test Cases and Data Sets

The wind farm of interest features six Vestas V52 wind turbines, installed in the year 2007, and it is sited in Italy on mountainous terrain.

The SCADA data sets that were used have ten minutes of sampling time and go from 1 January 2017 to 31 December 2018. At the beginning of March 2018, the target wind turbine (named Tar in this study) experienced electrical damage at the generator, in consequence of which the generator had to be replaced. In [21], a study was devoted to the analysis of the performance of the Tar wind turbine before and after the generator replacement, in comparison to the other wind turbines in the farm. The objective of that study was the assessment of the impact of the generator aging on wind turbine performance. The result was that after the generator replacement the performance of Tar slightly recovered, while the performance of the rest of the wind farm kept slightly worsening due to the effect of aging. In [21], only the main operation variables were analyzed.

The present study is instead devoted to the diagnosis of generator damage before it occurs and, for this purpose, the data set used also included the most important electrical parameters. The measurements used for this study were:

- Wind speed v (m/s);
- Power P (kW);
- Blade pitch angle β ($^{\circ}$);
- Rotor speed ω (rpm);
- Generator speed Ω (rpm);
- Gear bearing temperature T_b (K);
- Generator Phase 1 temperature T_1 (K);
- Generator Phase 2 temperature T_2 (K);
- Generator Phase 3 temperature T_3 (K);
- Current Phase 1 I_1 (A);
- Current Phase 2 I_2 (A);
- Current Phase 3 I_3 (A);
- Voltage Phase 1 V_1 (V);
- Voltage Phase 2 V_2 (V);
- Voltage Phase 3 V_3 (V);

Section 3.3 features a detailed discussion about how the data sets were arranged for the diagnosis of the damage at the Tar wind turbine and for the comparison against a healthy reference (named Ref). For completeness, Tar and Ref correspond to ITA4 and ITA3 according to the nomenclature of [21].

As regards the data pre-processing, in general it should be noted that the operation of wind turbines in industrial farms is affected by curtailments dictated by grid requirements and it is therefore necessary to filter the data appropriately. This was done as follows (similarly to [22]):

- filter using the run-time counter, requesting production for 600 s out 600;
- filter below rated power (approximately $v \leq 13$);
- the data corresponding to operation under grid curtailment are filtered out by automatic data clustering through the random forest algorithm [23].

An example of the scattered power curve before and after data pre-processing is reported in Figure 1.

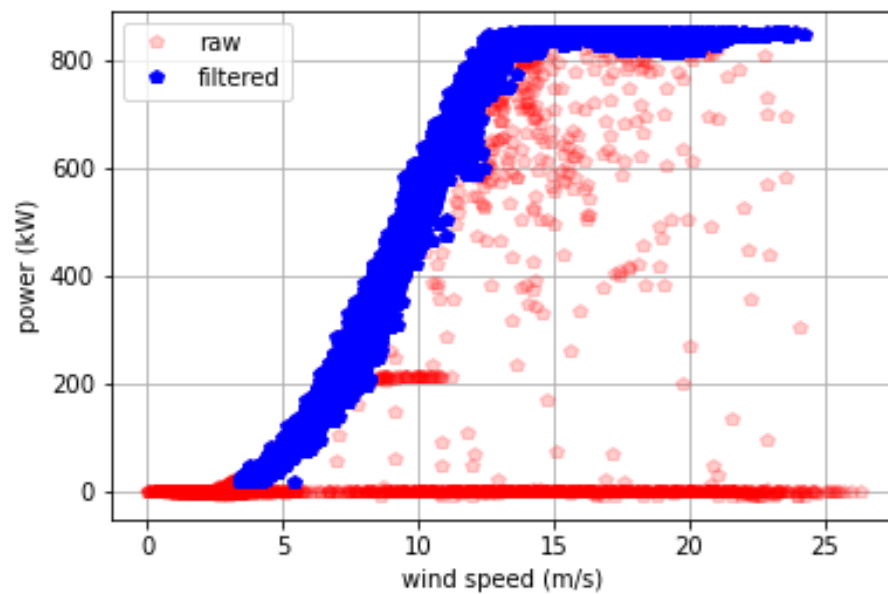


Figure 1. Effect of data clustering through the random forest algorithm: Tar wind turbine.

3. Methods

The structure of the methodology is as follows:

- SCADA data from the target damaged wind turbine (Tar) and from a reference healthy wind turbine (Ref) were pre-processed as indicated in Section 2.
- Based on the Pearson correlation coefficient between the possible input variables and the output (power, currents, and voltages), features were selected.
- The features matrix was orthogonalized and reduced in dimension through PCA transformation.
- The reduced features matrix was fed as input to a support vector regression with Gaussian kernel.
- The fault incipience was individuated through the analysis of how the mismatch between model estimates and measurements changed.

The structure of the method is summarized in Figure 2.

In Sections 3.1 and 3.2, the principles of principal component analysis and support vector regression are briefly summarized; the expert reader can skip these subsections. In Section 3.3, the use of the models for fault diagnosis is described.

3.1. Principal Component Analysis

Principal component analysis has been widely used in wind energy data analysis [24–29]. The rationale for this is that wind turbines are complex machines, which are regulated by a series of operation variables that are highly correlated among themselves but do not contain exactly the same information. Therefore, the removal of redundant information from a set of highly correlated features can be performed simply and effectively through PCA.

The principle of PCA is substituting a matrix x with a transformed matrix, whose columns are mutually orthogonal. The singular value decomposition of x is given in Equation (1):

$$x = U\Delta V^T. \quad (1)$$

The columns of U and V are orthonormal sets of vectors denoting the left and right singular vectors of x and Δ is a diagonal matrix, whose elements are the singular values of x . This allows decomposing xx^T as:

$$xx^T = V\Lambda V^T, \quad (2)$$

where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ and $\lambda_1 \geq \dots \geq \lambda_p \geq 0$.

xV_i is the i -th principal component and v_i is the i -th loading corresponding to the i -th principal value λ_i .

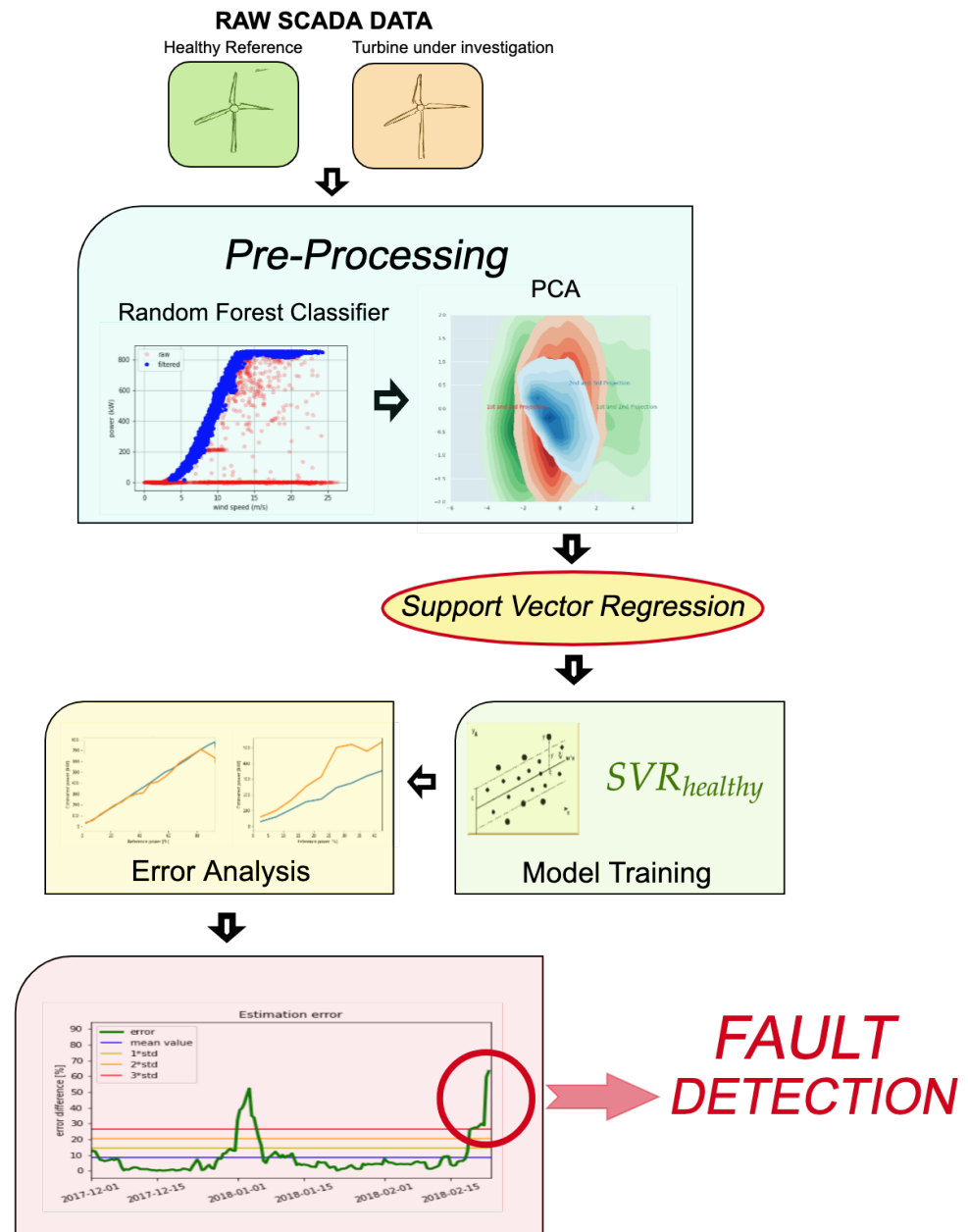


Figure 2. Schematic of the overall process for fault detection.

3.2. Support Vector Regression

Support vector regression has been vastly employed in general for renewable power generation applications [30] and in particular for several kinds of problem in wind energy, such as, for example, operation curve analysis [31,32], wind farm layout optimization [33], and forecasting [34,35].

Here follows a brief recap of the principles of support vector regression, starting from a linear model that is given in Equation (3):

$$y = x\beta + \epsilon, \quad (3)$$

where β is a vector of regression coefficients, whose best estimate has to be inferred from the input variable data matrix x and the output vector y .

Support vector regression is a methodology for estimating the β parameters, given the matrix of input variables x and the vector of output y . It is a constrained optimization problem, because the objective of this kind of regression is the minimum norm of $\beta'\beta$, subjected to the request that the residuals between the measurements y and the model estimate $f(x)$ are lower than a threshold ϵ for each n -th observation (Equation (4)):

$$|y_n - x_n\beta + b_n| \leq \epsilon. \quad (4)$$

The optimization problem can be rephrased in the Lagrange dual formulation; the function to minimize is $L(\alpha)$, given in Equation (5):

$$L(\alpha) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) x_i' x_j + \epsilon \sum_{i=1}^N (\alpha_i + \alpha_i^*) + \sum_{i=1}^N y_i (\alpha_i^* - \alpha_i), \quad (5)$$

with the constraints (Equation (6))

$$\begin{aligned} \sum_{n=1}^N (\alpha_n - \alpha_n^*) &= 0 \\ 0 &\leq \alpha_n \leq C \\ 0 &\leq \alpha_n^* \leq C, \end{aligned} \quad (6)$$

where C is the box constraint.

The solution for the β parameters in terms of the input variable matrix x and of the coefficients α_n or α_n^* is given in Equation (7):

$$\beta = \sum_{n=1}^N (\alpha_n - \alpha_n^*) x_n. \quad (7)$$

The interpretation of Equation (7) is that the algorithm selects the most meaningful observations (rows of the x matrix) by setting most of the α or α^* coefficients to zero. The observations associated to non-vanishing α or α^* are consequently called support vectors.

Once the β coefficients have been computed on a reference data set as in Equation (7), given a new input variable x' it is possible to simulate the output by multiplying x' and β (Equation (8)):

$$f(x) = \sum_{n=1}^N (\alpha_n - \alpha_n^*) x_n' x + b. \quad (8)$$

A non-linear support vector regression is obtained by replacing the products between the observation matrix with a non-linear kernel function (Equation (9)):

$$G(x_1, x_2) = \langle \varphi(x_1) \varphi(x_2) \rangle, \quad (9)$$

where φ is a transformation mapping the x observations into the feature space.

A Gaussian kernel selection is given in Equation (10):

$$G(x_i, x_j) = e^{-\kappa \|x_i - x_j\|^2}, \quad (10)$$

where κ is the kernel scale.

Then Equation (5) can be rewritten as Equation (11):

$$L(\alpha) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) G(x_i, x_j) + \epsilon \sum_{i=1}^N (\alpha_i + \alpha_i^*) + \sum_{i=1}^N y_i (\alpha_i^* - \alpha_i), \quad (11)$$

and Equation (8) for predicting can be rewritten as Equation (12):

$$f(x) = \sum_{n=1}^N (\alpha_n - \alpha_n^*) G(x_n, x) + b. \quad (12)$$

In this work, the hyperparameters of the regression κ , C , ϵ were automatically optimized using Python. It was observed that, as long as the hyperparameters are set in order to achieve error metrics in the order of those reported in Section 4, their fine tuning does not sensibly affect the detection of the damage.

3.3. Normal Behavior Model

The data sets were arranged as follows for the analysis. It should be intended that a sensitivity analysis has been conducted, in order to individuate the effect of the data set size and proximity to the fault on the obtained result. This aspect is discussed in some detail in Section 4. For the moment, the general arrangement of the data set is:

- D_{train} is the data set employed for training the normal behavior model and describes the target wind turbine reasonably before the incipience of the fault.
- D_{test} is the data set for testing the model and setting the standards of the residuals between measurements and model estimates. Therefore, this data set also describes the target wind turbine reasonably before the incipience of the fault.
- D_{fault} describes the target wind turbine in proximity of the abrupt shutdown due to the fault.
- D_{after} describes the target wind turbine after the replacement of the damaged generator.

The same data set selection was done for the target wind turbine and for a reference healthy wind turbine.

The normal behavior model was subsequently constructed by selecting a target. Three options were explored in this study, namely:

- select the power P as the target of the model;
- select one phase current I as the target of the model;
- select one phase voltage V as the target of the model.

For each target y , the most appropriate input variables for the model were identified by constructing a cross-correlation map based on the Pearson correlation coefficient (Equation (13)) with each possible regressor x :

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (13)$$

where n is the number of observations in the D_{train} data set, and \bar{x} and \bar{y} are the averages of the output and of the regressor x in the D_{train} data set.

For each target above indicated, the correlation coefficients were computed with respect to each variable indicated in Section 2. It should be clarified that, when the target was voltage or current, the voltages and currents of the other phases were excluded from the possible regressors because they basically contain the same information and the resulting model would not be explanatory. The criterion for including a regressor is a threshold on the Pearson coefficient r_{xy} ; a sensitivity study was performed and a lower threshold of 0.6 was set. The selected regressors and the Pearson coefficients r_{xy} with the target are reported in Table 1.

Table 1. Selected features for the normal behavior models and Pearson coefficient r_{xy} with the target: D_{train} data set.

Target	Regressors	r_{xy}
P	$\{v \ \beta \ T_b \ \omega \ \Omega \ T_1 \ T_2 \ T_3 \ V_1 \ V_2 \ V_3\}$	$\{0.97 \ 0.61 \ 0.89 \ 0.90 \ 0.90 \ 0.85 \ 0.85 \ 0.85 \ 0.72 \ 0.67 \ 0.75\}$
I	$\{v \ \beta \ T_b \ \omega \ \Omega \ T_1 \ T_2 \ T_3 \ V_1 \ V_2 \ V_3\}$	$\{0.97 \ 0.61 \ 0.86 \ 0.90 \ 0.90 \ 0.85 \ 0.85 \ 0.85 \ 0.71 \ 0.66 \ 0.74\}$
V	$\{v \ T_b \ \omega \ \Omega \ T_1 \ T_2 \ T_3 \ I_1 \ I_2 \ I_3\}$	$\{0.67 \ 0.67 \ 0.67 \ 0.67 \ 0.63 \ 0.63 \ 0.63 \ 0.71 \ 0.71 \ 0.71\}$

After selecting the regressors, the principal component transform was performed (Equation (2)) and the first two principal components were fed as input to the support vector regression for the target. The hyperparameters of the support vector regression were then optimized based on an automatic cross validation. Subsequently the target was simulated for the data sets D_{test} , D_{fault} and D_{after} and the behavior of the residuals between model estimates and measurements was analyzed qualitatively and quantitatively, by comparing the different periods of the target wind turbine and by comparing the target and the reference wind turbines. The selected metrics are the most commonly employed and are the normalized mean absolute error (NMAE), the mean absolute percentage error (MAPE) and the normalized root-mean-square error (NRMSE). The NMAE is defined in Equation

$$NMAE = \frac{100}{\bar{y}M} \sum |y - f(x)|, \quad (14)$$

where M is the number of observations in the data set of interest, y are the observations and $f(x)$ are the model estimates, and \bar{y} is the nominal value of the target. The rationale for normalizing to the nominal value of the target is that in this study models were formulated for three different quantities (power, voltage, and current). The NRMSE is defined in Equation (15):

$$NRMSE = \frac{100}{\bar{y}} \sqrt{\frac{\sum (y - f(x) - ME)^2}{M}}, \quad (15)$$

where ME is the mean error and is defined in Equation (16):

$$ME = \frac{1}{M} \sum y - f(x). \quad (16)$$

The MAPE is defined in Equation (17):

$$MAPE = \frac{100}{M} \sum \frac{|y - f(x)|}{y}. \quad (17)$$

4. Results

A preliminary result regards the positive effect of the use of PCA for the normal behavior model. Table 2 reports an estimation of the computational time required for the models with and without PCA. The improvement in computational time is appreciable.

Table 2. Simulation time with and without the use of principal component analysis (PCA).

Target	Sim. Time PCA	Sim. Time No PCA
Power P	84 s	354 s
Current I	91 s	383 s
Voltage V	170 s	968 s

Figure 3 reports an example of a time series of simulated and measured current for the healthy Ref wind turbine. From Figure 3, it is possible to notice that the dimension reduction through the PCA does not significantly affect the behavior of the model estimates, whose error metrics for the test data set are summarized in Table 3. From Table 3, it arises that the error metrics on the test data set are acceptably low and therefore the model is

potentially capable of detecting anomalies, as will be seen herein. In order to estimate margins for the reliability of the proposed model, a 10-fold cross validation was performed with 300 model runs: for each model run, the error metrics were computed and the results reported in Table 3 should be intended as the average metrics over the model runs, to which a standard deviation was associated. From Table 3 it arises that the average error metrics can be considered particularly robust because their standard deviation is very low. Furthermore, as regards the power P , it is possible to compare with the literature: the order of the error metrics is similar to the results collected in [36] for the same test case. Therefore, it can be stated that the dimension reduction through the PCA does not remarkably affect the quality of the regression and helps in highlighting the dependence of the target on the orthogonalized input variable matrix.

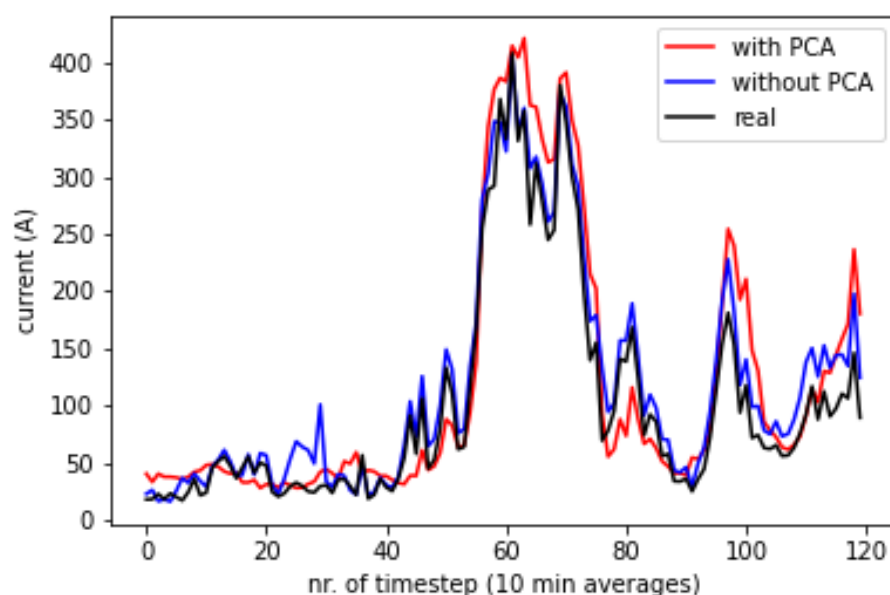


Figure 3. Comparison of the simulation of the current for the healthy Ref wind turbine, with and without PCA.

Table 3. Average error metrics and standard deviations for the 10-fold cross validation on the D_{test} data set: PCA model.

Target	NMAE (%)	NRMSE (%)	MAPE (%)
Power P	2.0 ± 0.1	3.1 ± 0.1	5.5 ± 1.2
Voltage V	0.63 ± 0.01	0.80 ± 0.01	0.04 ± 0.02
Current I	1.9 ± 0.1	2.9 ± 0.2	6.3 ± 1.0

In Figure 4, qualitative results are reported that are meaningful for the identification of the incoming fault and which will be elaborated further on. The simulated power is reported as a function of the measured power for two data sets approaching the date of the fault. On the left, the plots for the reference wind turbine Ref are reported, while the plots for the target faulty wind turbine Tar are on the right. It arises that, approximately two weeks before the fault, it is possible to distinguish anomalous behavior at the Tar wind turbine: the simulated power is significantly higher than the measured value. This means that the input features are such that the extracted power should be higher, according to the normal behavior model. In Figure 5, a similar kind of plot is reported for the simulation of the current, approaching the fault (the same data set as Figure 4c,d); from this Figure, it arises that the fault onset can be individuated, also modeling directly the electric parameters of the generator.

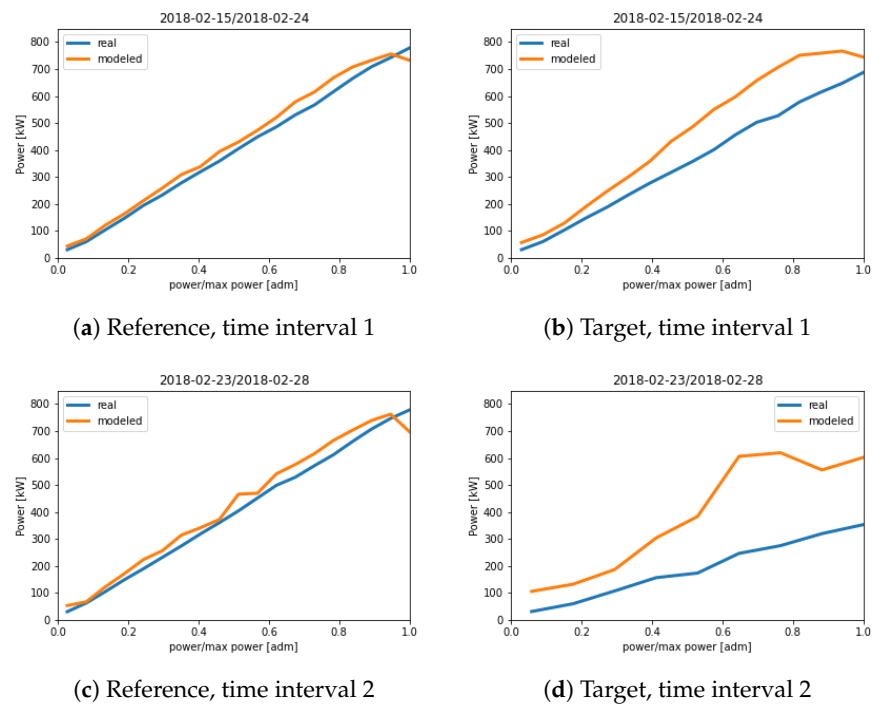


Figure 4. Measured and simulated power for the reference (Ref) and the target (Tar) wind turbines, for several examples of time windows approaching the fault (D_{fault}).

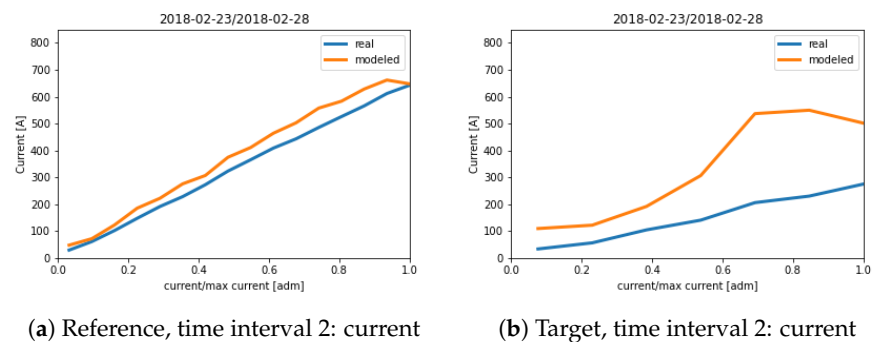


Figure 5. Measured and simulated current for the reference (Ref) and the target (Tar) wind turbines: data set D_{fault} .

In Figure 6, the same kind of plot as in Figure 4 is reported for a sample data set after the replacement of the Tar generator. From Figure 6, it arises that the reference and the target are hardly distinguishable, which is different to what happened in proximity of the fault.

The situation depicted in Figures 4 and 5 can be interpreted quantitatively by computing the Pearson correlation coefficients r_{xy} between the measurements of the targets and of the input variables in the data set D_{after} for the target wind turbine. The results are reported in Table 4, from which it arises that the correlation coefficients change drastically with respect to the D_{train} data set (Table 1). In this sense, it is confirmed that the onset of generator damage can also be individuated as a change in the correlation between relevant operation variables (as suggested in [19]). In Table 5, the Pearson correlation coefficients r_{xy} are reported for the data set D_{after} and, compatibly, it results that the obtained values return to being of the order of those reported in Table 1 for the target wind turbine in healthy conditions.

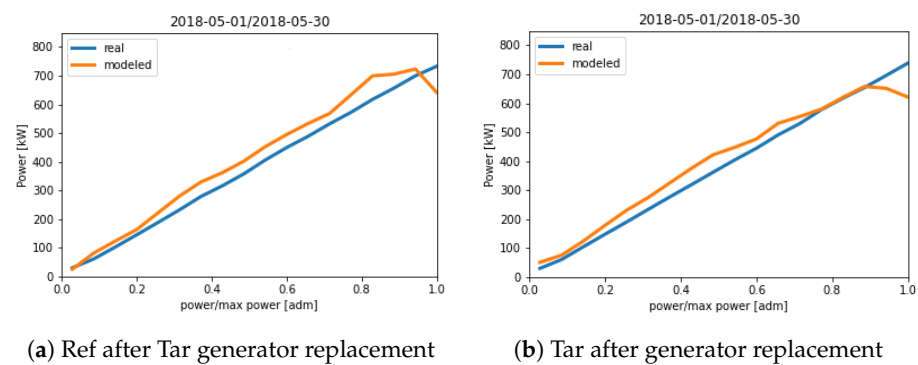


Figure 6. Comparison between Ref and Tar after the replacement of the Tar generator: data set D_{after} .

Table 4. Correlation coefficients r_{xy} for the D_{fault} data set.

Target	Regressors	r_{xy}
P	$\{v \beta T_b \omega \Omega T_1 T_2 T_3 V_1 V_2 V_3\}$	$\{0.98 -0.22 0.92 0.99 0.99 0.38 0.37 0.36 -0.14 -0.17 -0.11\}$
I	$\{v \beta T_b \omega \Omega T_1 T_2 T_3 V_1 V_2 V_3\}$	$\{0.98 -0.22 0.92 0.99 0.99 0.38 0.38 0.36 -0.15 -0.17 -0.12\}$
V	$\{v T_b \omega \Omega T_1 T_2 T_3 I_1 I_2 I_3\}$	$\{-0.15 -0.05 -0.12 -0.12 -0.03 -0.02 -0.04 -0.15 -0.15 -0.15\}$

Table 5. Correlation coefficients r_{xy} for the target turbine after the replacement of the generator.

Target	Regressors	r_{xy}
P	$\{v \beta T_b \omega \Omega T_1 T_2 T_3 V_1 V_2 V_3\}$	$\{0.97 0.59 0.86 0.88 0.88 0.81 0.81 0.80 0.68 0.65 0.73\}$
I	$\{v \beta T_b \omega \Omega T_1 T_2 T_3 V_1 V_2 V_3\}$	$\{0.97 0.58 0.87 0.89 0.89 0.81 0.81 0.80 0.67 0.64 0.72\}$
V	$\{v T_b \omega \Omega T_1 T_2 T_3 I_1 I_2 I_3\}$	$\{0.67 0.56 0.50 0.50 0.62 0.62 0.61 0.67 0.67 0.67\}$

In Figure 7, the behavior of Figure 4 is elaborated through the indication of meaningful historical thresholds for the detection of the incoming fault. In Figure 7, the difference in NMAE between the Tar and the Ref wind turbines is plotted for the D_{fault} data set. The statistical indicators for the quantity of Figure 7 have been computed using the D_{test} data set: these are the average (blue line) and the standard deviation (yellow line). Therefore, for each measurement of Figure 7 in the D_{fault} data set, it is possible to estimate by how many standard deviations the difference in NMAE between the Tar and the Ref deviates from the historical normal trend. This allows observation of the evolution of the incoming fault and two reasonable thresholds can be defined: a pre-alert one (two standard deviations) and an alert one (three standard deviations). From Figure 7, it arises that two relevant peaks above the alert thresholds have occurred before the stopping of the wind turbine. It was crosschecked against the alarm log book for the wind turbines that both peaks are concomitant with the onset of alarms individuating current anomalies reaching the converter: in light of the present work, this phenomenon can be explained as being due to anomalous electrical behavior of the generator, eventually resulting in converter current anomalies. It should be pointed out that the present method is more informative with respect to the analysis of the alarm log book: alarms are impulsive events, while the approach of this study can be employed for online continuous monitoring. Defining appropriately, as is done in this work, a set of thresholds, it is possible to individuate possible faults with more advance with respect to the mere elaboration of the alarm logs. A further observation regards the fact that in the D_{fault} , the quantity reported in Figure 7 deviates more than one standard deviation with respect to the historical only in anticipation of an alarm event. This supports that the number of false positives indicated by this method can be reasonably low and this aspect will be analyzed further when other test case studies are available to the authors. Another interesting aspect of Figure 7 is that the latter peak anticipates the damage and two observations arise: this peak is higher with respect to the

former alarm event and the alarm threshold is passed approximately two weeks before the stopping of the wind turbine. From these matters of fact, the usefulness of the proposed method is supported because it is responsive and can also anticipate error logs that are not associated with short-term stopping of the wind turbine. This is particularly important in the perspective of using this kind of method for condition assessment.

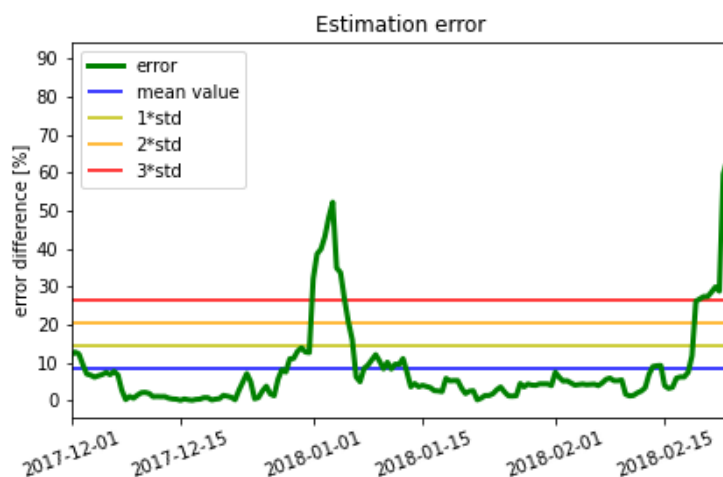


Figure 7. The time history of the difference between Tar—Ref normalized mean absolute error (NMAE) with different statistical thresholds.

5. Conclusions

The present study has been devoted to use of SCADA data for the diagnosis of electrical damage at wind turbine generators. As discussed in Section 1, this objective is particularly challenging because the sampling time of SCADA data is not the most appropriate for interpreting the dynamics of electrical phenomena of machines subjected to non-stationary conditions, as wind turbines are. Nevertheless, the widespread use of SCADA data and the potential applications in wind energy practice motivate a continuously growing scientific interest.

The main result of this study is that appropriate SCADA data analysis methods are helpful in diagnosing electrical damage to wind turbine generators. This has been accomplished through the analysis of a real-world test case, which is the breakdown and the consequent replacement of a generator at a Vestas V52 wind turbine sited in southern Italy.

The proposed methodology is based on the construction of normal behavior models for the power and for the voltages and currents of the wind turbine. The relevant features were selected based on their Pearson correlation coefficient with the target and the PCA was employed to reduce the dimension of the problem and to deal with the collinearity of the regressors. The non-linear relation between the features and the target was taken into account using support vector regression with a Gaussian kernel.

Through the analysis of how the mismatch between model estimates and measurements evolved approaching the time of the fault, it was possible to individuate the damage with an advance in the order of two weeks. Considering the nature of the damage and the type of employed data, this result is very promising as regards the use of SCADA data for monitoring the health status of wind turbine generators.

A valuable further direction of the present work would be the use of time-resolved SCADA data with a sampling time of the order of the second [37]. Despite the fact that in wind energy practice they are more complex to manage from the OPC-DA servers of wind turbines, these kinds of data could likely provide a deeper insight into the dynamics of wind turbines and this could further improve the diagnostic capability of data-driven methods.

In general, a realistic application for the type of analysis presented in this work is to constitute a first advice for the assessment of generator conditions, based on which it could

be possible to plan devoted inspections using methodologies similar to those in [8]. From this point of view, the present study represents a contribution to the methodologies for condition-based maintenance of wind turbine generators.

Author Contributions: Conceptualization, D.A., F.C., F.N.; data curation, D.A., F.C., F.N.; formal analysis, D.A., F.C., F.N.; investigation, D.A., F.C., F.N.; methodology, D.A., F.C., F.N.; software, D.A., F.C., F.N.; supervision, F.C., validation, D.A., F.C., F.N.; writing—original draft, D.A.; writing—review editing, F.C., F.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors thank the company Lucky Wind Spa for the technical support and for providing the data sets employed for the study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DA	Data access
IEC	International Electrotechnical Commission
MAPE	Mean absolute percentage error
ME	Mean error
NMAE	Normalized mean absolute error
NRMSE	Normalized root-mean-square error
OPC	Open Platform Communications
PCA	Principal component analysis
SCADA	Supervisory control and data acquisition
SVR	Support vector regression

References

1. Sinha, Y.; Steel, J.A. A progressive study into offshore wind farm maintenance optimisation using risk based failure analysis. *Renew. Sustain. Energy Rev.* **2015**, *42*, 735–742. [\[CrossRef\]](#)
2. Fischer, K.; Besnard, F.; Bertling, L. Reliability-centered maintenance for wind turbines based on statistical analysis and practical experience. *IEEE Trans. Energy Convers.* **2011**, *27*, 184–195. [\[CrossRef\]](#)
3. Tchakoua, P.; Wamkeue, R.; Ouhrouche, M.; Slaoui-Hasnaoui, F.; Tameghe, T.A.; Ekemb, G. Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges. *Energies* **2014**, *7*, 2595–2630. [\[CrossRef\]](#)
4. Hossain, M.L.; Abu-Siada, A.; Mueen, S. Methods for advanced wind turbine condition monitoring and early diagnosis: A literature review. *Energies* **2018**, *11*, 1309. [\[CrossRef\]](#)
5. Pérez, J.M.P.; Márquez, F.P.G.; Tobias, A.; Papaelias, M. Wind turbine reliability analysis. *Renew. Sustain. Energy Rev.* **2013**, *23*, 463–472. [\[CrossRef\]](#)
6. Carroll, J.; McDonald, A.; McMillan, D. Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. *Wind Energy* **2016**, *19*, 1107–1119.
7. Artigao, E.; Martín-Martínez, S.; Honrubia-Escribano, A.; Gómez-Lázaro, E. Wind turbine reliability: A comprehensive review towards effective condition monitoring development. *Appl. Energy* **2018**, *228*, 1569–1583. [\[CrossRef\]](#)
8. Artigao, E.; Honrubia-Escribano, A.; Gómez-Lázaro, E. In-service wind turbine DFIG diagnosis using current signature analysis. *IEEE Trans. Ind. Electron.* **2019**, *67*, 2262–2271. [\[CrossRef\]](#)
9. Artigao, E.; Sapena-Bano, A.; Honrubia-Escribano, A.; Martínez-Roman, J.; Puche-Panadero, R.; Gómez-Lázaro, E. Long-term operational data analysis of an in-service wind turbine DFIG. *IEEE Access* **2019**, *7*, 17896–17906. [\[CrossRef\]](#)
10. Maldonado-Correa, J.; Martín-Martínez, S.; Artigao, E.; Gómez-Lázaro, E. Using SCADA Data for Wind Turbine Condition Monitoring: A Systematic Literature Review. *Energies* **2020**, *13*, 3132. [\[CrossRef\]](#)
11. Zaher, A.; McArthur, S.; Infield, D.; Patel, Y. Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy Int. J. Prog. Appl. Wind. Power Convers. Technol.* **2009**, *12*, 574–593. [\[CrossRef\]](#)
12. Corley, B.; Carroll, J.; McDonald, A. Fault detection of wind turbine gearbox using thermal network modelling and SCADA data. *J. Phys. Conf. Ser.* **2020**, *1618*, 022042. [\[CrossRef\]](#)
13. Zeng, X.; Yang, M.; Bo, Y. Gearbox oil temperature anomaly detection for wind turbine based on sparse Bayesian probability estimation. *Int. J. Electr. Power Energy Syst.* **2020**, *123*, 106233. [\[CrossRef\]](#)
14. Guo, P.; Fu, J.; Yang, X. Condition Monitoring and Fault Diagnosis of Wind Turbines Gearbox Bearing Temperature Based on Kolmogorov-Smirnov Test and Convolutional Neural Network Model. *Energies* **2018**, *11*, 2248. [\[CrossRef\]](#)

15. Guo, P.; Infield, D.; Yang, X. Wind turbine generator condition-monitoring using temperature trend analysis. *IEEE Trans. Sustain. Energy* **2011**, *3*, 124–133. [\[CrossRef\]](#)
16. Kusiak, A.; Verma, A. Analyzing bearing faults in wind turbines: A data-mining approach. *Renew. Energy* **2012**, *48*, 110–116. [\[CrossRef\]](#)
17. Astolfi, D.; Castellani, F.; Natili, F. Wind turbine generator slip ring damage detection through temperature data analysis. *Diagnostyka* **2019**, *20*. [\[CrossRef\]](#)
18. Zhang, S.; Lang, Z.Q. SCADA-data-based wind turbine fault detection: A dynamic model sensor method. *Control. Eng. Pract.* **2020**, *102*, 104546. [\[CrossRef\]](#)
19. Zhao, Y.; Li, D.; Dong, A.; Kang, D.; Lv, Q.; Shang, L. Fault prediction and diagnosis of wind turbine generators using SCADA data. *Energies* **2017**, *10*, 1210. [\[CrossRef\]](#)
20. Jin, X.; Xu, Z.; Qiao, W. Condition monitoring of wind turbine generators using SCADA data analysis. *IEEE Trans. Sustain. Energy* **2020**, *12*, 202–210. [\[CrossRef\]](#)
21. Astolfi, D.; Byrne, R.; Castellani, F. Estimation of the Performance Aging of the Vestas V52 Wind Turbine through Comparative Test Case Analysis. *Energies* **2021**, *14*, 915. [\[CrossRef\]](#)
22. Astolfi, D.; Castellani, F.; Lombardi, A.; Terzi, L. Multivariate SCADA Data Analysis Methods for Real-World Wind Turbine Power Curve Monitoring. *Energies* **2021**, *14*, 1105. [\[CrossRef\]](#)
23. Zhang, D.; Qian, L.; Mao, B.; Huang, C.; Huang, B.; Si, Y. A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access* **2018**, *6*, 21020–21031. [\[CrossRef\]](#)
24. Pozo, F.; Vidal, Y. Wind turbine fault detection through principal component analysis and statistical hypothesis testing. *Energies* **2016**, *9*, 3. [\[CrossRef\]](#)
25. Pozo, F.; Vidal, Y.; Salgado, Ó. Wind turbine condition monitoring strategy through multiway PCA and multivariate inference. *Energies* **2018**, *11*, 749. [\[CrossRef\]](#)
26. Wang, Y.; Ma, X.; Qian, P. Wind turbine fault detection and identification through PCA-based optimal variable selection. *IEEE Trans. Sustain. Energy* **2018**, *9*, 1627–1635. [\[CrossRef\]](#)
27. Rezamand, M.; Kordestani, M.; Cariveau, R.; Ting, D.S.K.; Saif, M. A new hybrid fault detection method for wind turbine blades using recursive PCA and wavelet-based PDF. *IEEE Sens. J.* **2019**, *20*, 2023–2033. [\[CrossRef\]](#)
28. Wang, Y.; Ma, X.; Joyce, M.J. Reducing sensor complexity for monitoring wind turbine performance using principal component analysis. *Renew. Energy* **2016**, *97*, 444–456. [\[CrossRef\]](#)
29. Castellani, F.; Garibaldi, L.; Daga, A.P.; Astolfi, D.; Natili, F. Diagnosis of faulty wind turbine bearings using tower vibration measurements. *Energies* **2020**, *13*, 1474. [\[CrossRef\]](#)
30. Sharifzadeh, M.; Sikinioti-Lock, A.; Shah, N. Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian Process Regression. *Renew. Sustain. Energy Rev.* **2019**, *108*, 513–538. [\[CrossRef\]](#)
31. Pandit, R.K.; Infield, D. Comparative assessments of binned and support vector regression-based blade pitch curve of a wind turbine for the purpose of condition monitoring. *Int. J. Energy Environ. Eng.* **2019**, *10*, 181–188. [\[CrossRef\]](#)
32. Byrne, R.; Astolfi, D.; Castellani, F.; Hewitt, N.J. A Study of Wind Turbine Performance Decline with Age through Operation Data Analysis. *Energies* **2020**, *13*, 2086. [\[CrossRef\]](#)
33. Ju, X.; Liu, F.; Wang, L.; Lee, W.J. Wind farm layout optimization based on support vector regression guided genetic algorithm with consideration of participation among landowners. *Energy Convers. Manag.* **2019**, *196*, 1267–1281. [\[CrossRef\]](#)
34. Santamaría-Bonfil, G.; Reyes-Ballesteros, A.; Gershenson, C. Wind speed forecasting for wind farms: A method based on support vector regression. *Renew. Energy* **2016**, *85*, 790–809. [\[CrossRef\]](#)
35. Hu, Q.; Zhang, S.; Yu, M.; Xie, Z. Short-term wind speed or power forecasting with heteroscedastic support vector regression. *IEEE Trans. Sustain. Energy* **2015**, *7*, 241–249. [\[CrossRef\]](#)
36. Astolfi, D.; Castellani, F.; Natili, F. Wind Turbine Multivariate Power Modeling Techniques for Control and Monitoring Purposes. *J. Dyn. Syst. Meas. Control* **2021**, *143*. [\[CrossRef\]](#)
37. Gonzalez, E.; Stephen, B.; Infield, D.; Melero, J.J. Using high-frequency SCADA data for wind turbine performance monitoring: A sensitivity study. *Renew. Energy* **2019**, *131*, 841–853. [\[CrossRef\]](#)