In review: Applied Energy

Wind Turbine Power Curve Model Driven Conditional Monitoring and Fault Detection of Wind Turbines.

Emerald U. Henry[a], Olayinka S. Ohunakin[a,b], Victor U. Ezekiel[a]

[a]The Energy and Environment Research Group (TEERG), Mechanical Engineering Department, Covenant University, Ogun State, Nigeria
[b]Faculty of Engineering & the Built Environment, University of Johannesburg, South Africa

Highlights

- A new method based on Kolmogorov-Smirnov test is proposed for conditional monitoring of Wind Turbines.
- The modeled power curve is used to generate bin-wise significance levels and bin-wise ground truth distributions.
- Selection of the most accurate power curve model is achieved by Mariano-Preve test
- Fault detection and model selection are validated by hypothesis tests
- Other variables, other than wind speed, are used in developing power curve models that are most representative of actual field conditions.
- Results indicate that the proposed method can adequately monitor wind turbines and detect faults before they occur.

Abstract

This research presents a new method for conditional monitoring based on the wind turbine power curve. The Kolmogorov-Smirnov distribution test from the field of statistics is employed in the assessment of turbine data and the detection of abnormality (faults) in wind turbines. The process begins with anomaly detection and filtration of faulty SCADA data by a quantile based filtration approach. Useful data comprising wind speed, density, ambient temperature and pitch angle are utilized in the development of wind turbine power curve models that represents actualities within wind farms. The radial basis function (RBF), multi-layer Perceptron (MLP) and gradient boosting (GBR) methods utilized for model development are compared for predictive accuracy using Mariano-Preve test, the null hypothesis assumes equal predictive ability (EPA). If rejected, an algorithm compares the coefficients of correlation of the models and selects the closest to one (unity). The most accurate model is utilized for the creation of a bin-wise distribution from past data and bin-wise confidence levels from the plot of wind speed and output power. Cochran's method validates the minimum sample size that will possess a sampling distribution similar to that of the population, a fault is detected if there is a reasonable difference between the sample distribution and population distribution. The Kolmogorov-Smirnov test, having a null hypothesis of equivalent distributions, signals a fault if the null hypothesis is rejected. Two wind turbine SCADA data sets associated with two fault events are used for the assessment of our method. Our results indicate that our method effectively highlights abnormalities in power output relating to increased bearing temperature and reduced generator rpm, aiding in the detection of faults long before the occur.

Keywords: Wind Turbine, Condition monitoring, Fault detection, Power curve, Mariano-Preve test, Cochran test, Kolmogorov-Smirnov test, Nonparametric statistical test, SCADA.

1 Introduction

Wind energy has acquired much attention in recent times because asides from being one of the most potent renewable energy sources, it is available almost anywhere and lacks pollution [1]. This development could address the power problems associated with fossil fuels and lead to a more sustainable future [2]. The Global Wind Energy Council has recorded an increase in wind power installations by 93 GW in 2020 alone totaling the offshore and onshore installed capacity to 743 GW [3]. However due to the complexity of the wind turbine assembly and the ever changing operating conditions in sites where they are deployed, they experience high failure rates that are a result of failures associated with gearbox, generators and blades [4]–[6]. Because of the size of these components, their failures are very expensive and usually leads to excessive downtime [7]. It is therefore necessary to monitor and detect faults before they transit to critical failures [8]. Condition based monitoring of wind farms has been considered the most effective solution for effective maintenance of wind turbines [8]–[12].

The available methods employed for WT conditional monitoring can be grouped into signal-based, physical model and data-driven approaches. Signal-based methods require the installation sensors for information capture leading to significant increase in O&M costs. Their methods include; acoustic emission [13], [14], infrared [15], vibration analysis [16]–[18] etc. physical model based methods often require expertise knowledge of various systems and their interactions in order to develop an accurate physical model, it is also very difficult to obtain accurate physical models because of the system to system interactions [19]. Hence this approach rather focuses on developing component-specific physical models, such as physical model representations for gears and bearings [20]. Data driven approaches develop methods that rely on the data collated by the turbine's supervisory control and data acquisition (SCADA) system [11], [21]–[32]. This is a cost effective approach because no additional cost is incurred in installation of sensors or other devices [11]. Additionally, due to the diversity of parametric and nonparametric techniques available in the literature, it is considered the most effective solution to WT conditional monitoring. SCADA data contains a number of parameters that are captured at mostly 10 minute intervals, each of which have to be meticulously and continuously analyzed with the aim of detecting faults early. It is often a daunting endeavor for the most skillful analyst leading to poor assessment because of the mental strain caused by the volume of data available for analysis. Hence researchers have attempted developing various approaches to aid in the classification and detection of abnormalities from turbine SCADA, a plethora of research employ ML methods for classification and regression in an attempt at distinguishing between faulty and proper readings, these methods include k-nearest neighbor [33], support vector machines [34], decision trees [35] and neural networks [8], [11], [36]–[38]. However, while they have been known to yield comparative performance they are complicated, require long training times and enormous computation cost; these have led researchers to investigate more efficient solutions from the fields of statistics and econometrics [39], [40]. Despite these limitations, in recent times, AI and ML based solutions have been employed for conditional monitoring by attempting to develop models that extract information that separates the normal operating state from faulty state of a turbine. However, because not only are multiple parameters needed to confirm a fault, each turbine has a different fault representation meaning that the wind farm analysist will have to manage lots of models alongside various SCADA parameters. Additionally, the representations for normal functioning and faulty operations change overtime due to obvious reasons like turbine ageing, recalibration of sensors and change of critical parts rendering previously developed models inefficient. At this point, newer models should be developed for such turbine however, wind farm analysists usually don't have this skill set. In unique cases where the analysist possesses such skills, there exist no metric to identify when a model is due for replacement. In general, the performance of AI and ML based approaches for CM decreases in efficiency over time rendering them practically less useful [31]. Another approach that has been attempted in the literature is the use of parametric models for CM [41]–[43], deep learning based models have been argued to be parametric to some degree [25]. They problem with parametric approaches is that they generally assume that

the data follows a normal distribution. Tests like fishers exact, student's t and ANOVA usually assume homogeneous variance leading to very accurate analysis on normally distributed data. The cointegration method [22], [42], [43], CUSUM based approach [41] and chow test [27] are examples of parametric implementations for CM. However, in cases where the data doesn't follow a normal distribution, as is the case with most SCADA parameters analyzed during CM, they end up with misleading results. Most recently, researchers have attempted developing nonparametric techniques in order to accommodate for both normal and non-normal parameter distributions that a given SCADA parameter may follow. Nonparametric approaches only require the data to follow a continuous distribution.

This present study utilizes a nonparametric method in an attempt to close the gaps discussed thus far. The wind turbine power curve alongside Kolmogorov-Smirnov's test, a nonparametric statistical approach are utilized for the development of a new CM technique. Before providing discourse on the nonparametric test used in this paper, this author intends to briefly highlight the nature of wind turbine power curve models found in the literature. A plethora of algorithms and methods have been applied to the problem of developing a power curve that is a true representation of the actual conditions experienced in real world, they have been discrete models [44], stochastic models [45], parametric [46]–[48] and nonparametric models [49]–[51]. These models are mainly focused on the problems of power prediction and forecasting with little considerations to monitoring and troubleshooting or predictive control and optimization. Attempts have been made to include the concept of performance monitoring into papers focused on power prediction without the definition of a clear technique for performance monitoring based on the developed power curve model. Bear in mind that performance monitoring is closely related to condition monitoring but not identical to condition monitoring, while performance monitoring indicates underperformance of a wind farm it isn't aimed at fault detection. In essence, an assessment of the predicted power and the actual power output of a certain turbine over a time period could suffice as performance monitoring because any reasonable difference between the actual power output and the predicted power output will indicate if the turbine or wind farm is performing acceptably or underperforming but won't provide enough information on the imminence of a fault. This is because various field conditions and turbine variables (e.g. turbine age) could be responsible for the noticed underperformance. On the other hand, condition monitoring for fault detection requires a more specialized approach. To the best of this author's knowledge, no research within the literature provides a method for condition monitoring and fault detection based on the developed power curve. Additionally, the technique developed in this paper employs a bin-wise approach for SCADA analysis by attempting to discretize the continuous power variable into wind speed intervals and identify the probability distribution of each interval (bin), while simultaneously extracting significance levels (α). A few research in the literature have employed the concept of binning in the development of wind turbine power curves (WTPC). In Llombart *et al.*[44] modifications were made to the IEC 61400-12 bins method that developed a single line power curve by least squares method and binning. A comparison between the binning approach and support vector regression for estimating the rotor speed based power curve of a wind turbine is carried out in [52] with the aim of comparing efficiencies between the two approaches. No research in the literature, to the best of the author's knowledge, utilizes power values that fall within wind speed intervals (bins) for the formation of probability distributions and the extraction of confidence levels (α) in any form and for any other application.

Kolmogorov-Smirnov test is a nonparametric test from the field of statistics that measures the goodness of fit. It compares the cumulative distribution functions of two data samples or one sample and a population in order to assess whether they were drawn from the same distribution. The chi-squared test is an alternative however, it is most sensitive at the center of the distribution and least sensitive at the edges. Additionally, as the sample size decreases, the chi-squared test becomes inapplicable. On the other hand, Kolmogorov-Smirnov's test retains its efficiency on small samples as well as large ones, it also doesn't suffer from reduced efficiency around the edges of a distribution making it a better alternative for comparing data samples. The Kolmogorov-Smirnov's test has been applied in a diversity of fields for comparing sample with sample and sample with distribution. In

the works of Zhang et al. [53] Kolmogorov-Smirnov test is utilized for fast and robust sensing of spectrums in radio systems by computing the empirical cumulative distribution functions (ECDF) of some decision statistics and comparing it with the ECDF of the noise signal. The K-S test was applied for denoising MR images in Baselice et al. [54] by evaluating and comparing the CDF of different pixels with the aim of measuring similarities. The K-S test has also been applied for drift detection in machine learning [55], explanation of unreliable machine learning survival models [56], identifying the distribution of earthquake data to predict magnitude [57] and for detecting changes in maps of gamma spectra in radioactivity [58]. Surprisingly, the Kolmogorov-Smirnov's test has never been applied for condition monitoring of wind turbines or any other aspect of wind energy.

1.1 Contribution and outline

This study is aimed at developing a new method for conditional monitoring of wind turbines by utilizing the wind turbine power curve and three well developed test methods. Fault detection is validated by the Kolmogorov-Smirnov's test. The method comprises a sequence of steps. First, SCADA data obtained from the previous operations of a wind turbine is processed for anomaly detection and removal. A quantile based algorithm sets user defined quantiles that differentiates between normal and faulty data. Afterwards, useful SCADA parameters are utilized in developing power curve models that accurately represent actual field conditions within a wind farm. Superior predictive ability (SPA) of one of the compared models is asserted using the Mariano-Preve test of equal predictive ability (EPA) and by comparing their coefficients of determination. The most accurate model is utilized for the creation of bin-wise frequency distribution to serve as the ground truth data sample or population and bin-wise confidence levels to serve as the decision factor for Kolmogorov-Smirnov's test. Cochran [59] developed a method of identifying the minimum sample size with the capacity to retain distribution information from the population, this minimum sample size represents the minimum number of acquired SCADA temporal instances required to assert a fault. This method detects faults directly from the output power of a wind turbine in relation to an increase in bearing temperature and a reduction of generator speed. Two case studies of one year SCADA data from two onshore wind turbines are used to validate the developed method. These SCADA data are associated with two fault or abnormal events and in one of the cases we analyze the detected fault from the output power by indicating the simultaneous increase in bearing temperature and reduction in generator speed. To the best of the author's knowledge, none of the methods and tests utilized in this study have ever been investigated in the literature for condition monitoring of wind turbine or any other aspect of wind energy. This research is motivated by its efficacy, over the long term, when compared to methods based on ML algorithms, it is aimed at improving the current state of condition based maintenance measures that would in turn advance energy efficiency via the reduction of the expenses incurred during corrective maintenance, ultimately reducing the total cost of energy.

The rest of this paper is outlined as follows. Section 2 begins with a flowchart aimed at graphically explaining the methods utilized in this study. Furthermore, detailed description of the methods and algorithms are provided to familiarize the reader with the concepts used for the development of this condition monitoring technique. Additionally, a step-by-step procedure is provided for the provision of a sequential flow and breakdown of the developed method. Section 3 presents results on the application of all used algorithms and tests and a description of what these represents. Section 4 concludes the paper.

2 Methodology and Algorithms

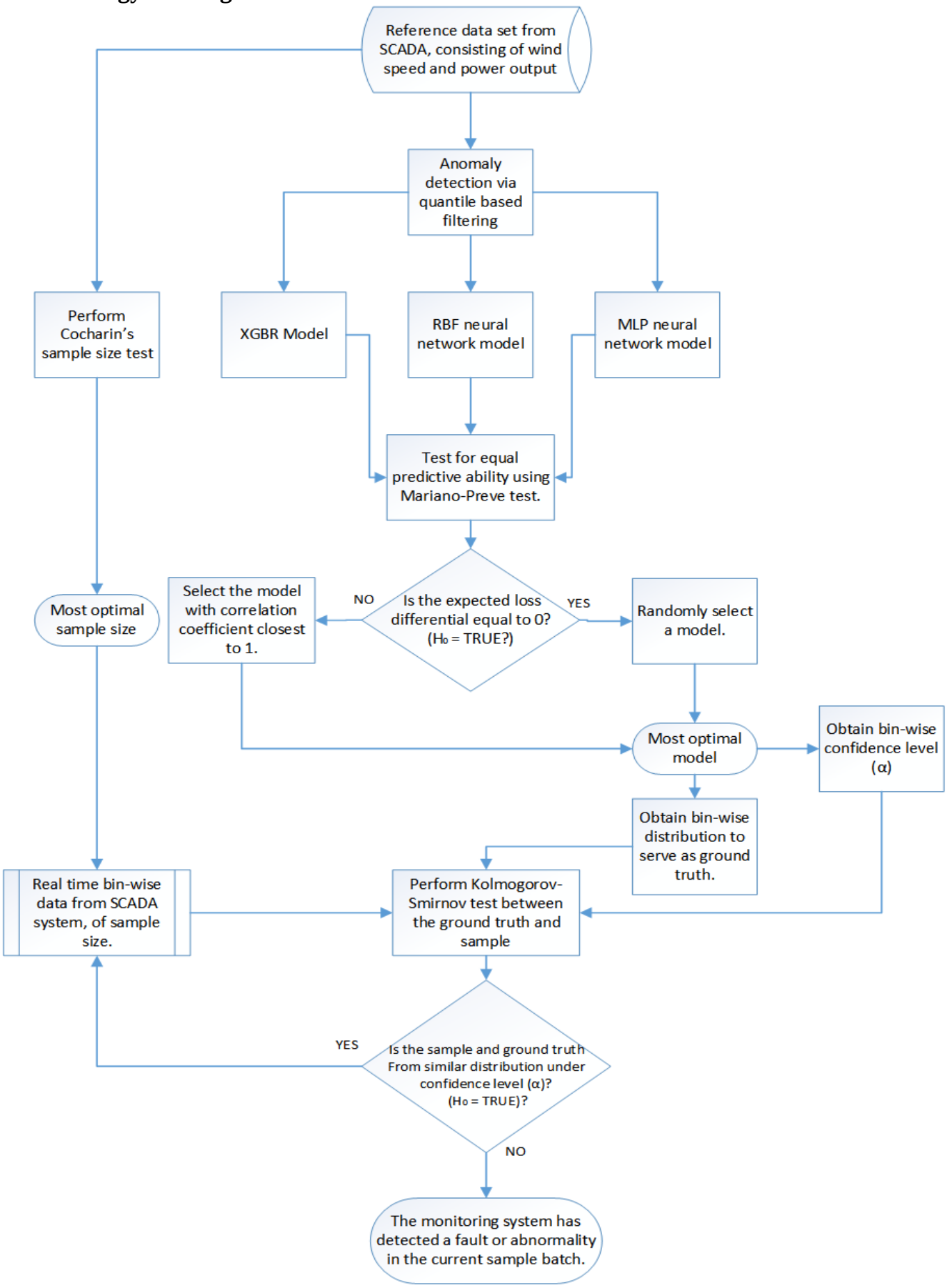


Figure 1: Computation flowchart for the detection of fault in condition monitoring of wind turbines from SCADA data and the developed power curve

2.1 Quantile Filtering

Quantiles usually define a particular part of a data in relation to other parts of the same data within a distribution. The simplest representation is a dividing plane that serves as a limiting condition to an assertion about the nature of the data. Consider the Figure (2) below, in this case we are considering a normally distributed data with no skewness (i.e. the LHS is an identical replica of the RHS) the distribution of q-quantile plots for all values $a \epsilon S$; the probability that *x* falls within quantile *q* is given by $P[X < x] \leq k/q$ (where *x* is a *k*-th *q*-quantile for a variable X), and the probability that *x* falls without the quantile *q* is given by $P[X < x] \geq 1 - k/q$ considering also that x is the *k*-th *q*-quantile for a variable *X*. the normal distribution is represented mathematically in Equation (1)

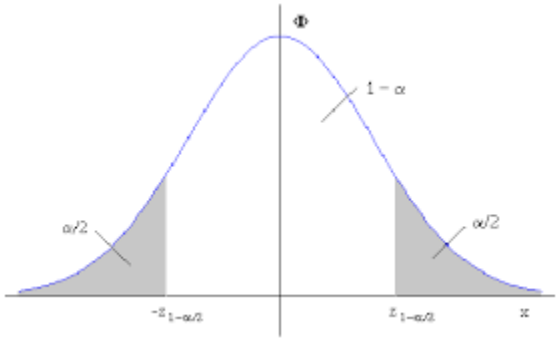


Figure 2: A normally distributed case for quantile specification

$$P[X < x] = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} e^{-\frac{t^2}{2}} dt \qquad (1)$$

The α-th quantile $\theta_\gamma(\propto), 0 <\propto< 1$ of a finite population vector $y = (y_1, \dots, y_N)$ is defined as

$$\theta_\gamma(\propto) = \inf\{t: F_\gamma(t) \geq \alpha\} \qquad (2)$$

where $F_\gamma(t)$ is the distribution function γ. In case $\dot{F}_\gamma(t)$, an estimator of $F_\gamma(t)$, is a monotonic non-decreasing function of *t*, the customary estimator of $\theta_\gamma(\propto)$ is obtained as

$$\theta_\gamma(\propto) = \inf\{t: \dot{F}_\gamma(t) \geq \alpha\} \qquad (3)$$

Let $\dot{F}_\gamma(t)$ be the customary estimator of $F_x(t)$. In case the population α-th quantile $\theta_x(\propto)$ of *x* is known, the ratio estimator of $\theta_\gamma(\propto)$ is given by

$$\ddot{\Theta}_{r\gamma}(\propto) = \frac{\ddot{\Theta}_\gamma(\propto)}{\ddot{\Theta}_x(\propto)} \theta_x(\propto) \qquad (4)$$

Similarly, a difference estimator of $\theta_\gamma(\propto)$ is given by:

$$\ddot{\Theta}_{d\gamma}(\propto) = \ddot{\Theta}_\gamma(\propto) - R\{\ddot{\Theta}_x(\propto) - \theta_x(\propto)\} \qquad (5)$$

where $R = \frac{\sum_{i \in s} \frac{y_i}{\pi_i}}{\sum_{i \in s} \frac{x_i}{\pi_i}}$ is a consistent estimator of the population ratio $R = Y/X$.

Both estimators $\ddot{\Theta}_{r\gamma}(\propto)$ and $\ddot{\Theta}_{d\gamma}(\propto)$ reduce to $\theta_\gamma(\propto)$ if $y_i \propto x_i \forall i \in U$. In this case, the variance become zero. The case is similar to a variety of distributions regardless of the nature, skewness or shape.

2.2 Gradient Boosting Regressor (GBR)

The concept of boosting aims at combining multiple base regressors to form a sequential ensemble for the purpose of developing a committee with better performance than any single regressor. Boosting is achieved by a step-wise training of a new learner, a weak learner and a base learner model with respect to the error realized at that step. Gradient Boosting Regressor (GBR) utilizes the concept of boosting for the development of an ensemble model that is a collection of tree models arranged sequentially. In this arrangement, the succeeding model learns from the errors of the preceding model, the performance of the preceding weak learning model is said to be boosted by the succeeding learner model. This ensemble is usually achieved by decision tree algorithm [60]. Considering a gradient boost regressor with N number of trees, the Equation (6) below can be stated.

$$f_N(x_j) = \sum_{n}^{N} \beta_n \, h_n(x_j) \tag{6}$$

Where h_n represents the weak learner model that has performed poorly on its own, β_n will represent the contribution of the model tree to the performance of the weak model, it is identified as a scaling factor. The loss function employed by XGBR to minimize errors is the gradient descent loss function. It achieves this by updating initial estimation with newer ones thereby tremendously improving performance of the final output.

2.3 Multi-layer Perceptron (MLP) Neural Networks

MLP networks are a type of feed forward neural networks that consist of three-layers; input layer, hidden layer and output layer. They have been widely used for regression and classification tasks. However, their efficacy is experienced in their ability to perform accurate regression analysis. A perceptron acquires a total of n features as input $x = x_1, x_1, \dots, x_n$, each of which has a weight associated to it. All features inputted into the network must be numeric in nature, hence, all non-numeric features must be first converted into numbers before being inputted into the network. The input features are passed on to an input function *u*, this function computes the weighted sum of the input features.

$$u(\boldsymbol{x}) = \sum_{i=1}^{n} w_i x_i \tag{7}$$

The result $u(\mathbf{x})$ is passed onto an activation function *f*, this function assists in producing the output of the perceptron. The activation function utilized in this step is a RELU.

$$y(x) = MAX(\mathbf{0}, x) \tag{8}$$

$$y(x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases} \tag{9}$$

Learning in MLPs, consist of adjusting the weights in order to reduce the error in predicting the training data. Learning is a back propagation task achieved by a back propagation algorithm (optimizer), that attempts to minimize the loss in predicting the ground truth.

2.4 Radial Basis Function

The RBF architecture was first proposed by Broom head, Lowe in their work captioned 'Radial Basis Functions, Multivariate Functional Interpolation and Adaptive Networks' 1988 [61]. RBFs consist of three layers by design: the input layer, the hidden layer, and the output layer. Figure (3) represents the typical structure of an RBF network with a single output node for single value tasks (e.g., regression etc.). The input node distributes the *k* input variables to the *m* nodes of the hidden layer. In the hidden layer, each node has a center with the same dimensions as the number of input variables. The hidden layer applies a non-linear transformation to the input space, transforming it into a higher-dimensional space. The activity $\mu_l(\mathbf{x}(\mathrm{f}))$ of the lth node is the Euclidean normal of the difference between the f-th input vector and the node center and is given in Equation (10) as:

$$\mu_l\big(\mathbf{x}(\mathrm{f})\big) = ||\boldsymbol{x}(f) - \dot{\mathrm{x}}_{\boldsymbol{l}}|| = \sqrt{\sum_{i=1}^{k} (\boldsymbol{x}(f) - \dot{\mathrm{x}}_{\boldsymbol{l},\boldsymbol{i}})^2}\,, \qquad f = 1, \dots, f \tag{10}$$

where f is the total number of available data, $\boldsymbol{x}^T(f) = [x_1(f), x_2(f), \dots, x_k(f)]$ is the input vector, and $\boldsymbol{x}_l^T = [\dot{\mathrm{x}}_{\mathbf{1}\,\boldsymbol{l}}, \dot{\mathrm{x}}_{\mathbf{2}\,\boldsymbol{l},\dots,}\dot{\mathrm{x}}_{\boldsymbol{k}\,\boldsymbol{l},}]$ is the centre of the lth node.

The activation function for each node is a radially symmetric function. In this work, we employ the sigmoid function (Equation (11)).

$$g(\mu) = \frac{1}{1 + e^{-\mu}} \tag{11}$$

The hidden node response is denoted by $\mathbf{z}(\mathrm{f})$ (Equation (12)):

$$\mathbf{z}(\mathrm{f}) = [g\left(\mu_1\big(\boldsymbol{x}(f)\big)\right), g\left(\mu_2\big(\boldsymbol{x}(f)\big)\right), \dots, g\left(\mu_m\big(\boldsymbol{x}(f)\big)\right)] \tag{12}$$

The output of an RBF network contains y unit, where y is the singular possible output value. The numerical output y(*f*) is produced by a linear combination of the hidden nodes' response (Equation (13)):

$$y(f) = \mathbf{z}(\mathrm{f}).\mathbf{w}_n = \sum_{i=1}^{m} w_{l,n} g\left(\mu_1\big(\boldsymbol{x}(f)\big)\right) \tag{13}$$

where $\mathbf{w}_n = [w_{1,n}, w_{2,n}, \dots, w_{m,n}]^T$ is a vector containing the synaptic weights corresponding to the output *n*.

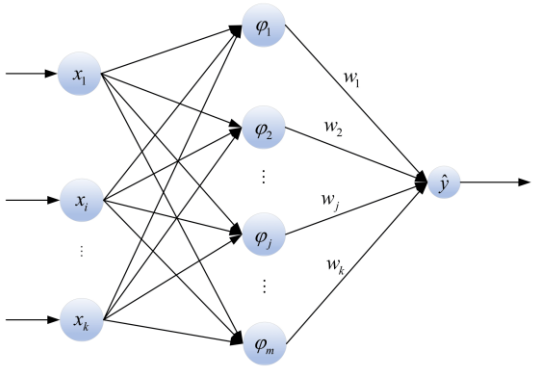


Figure 3: Radial Basis Function Network

The synaptic weights are commonly determined using linear regression of the hidden layer outputs to the real measured output after the RBF centers and non-linearities in the hidden layer have been fixed. In most cases, linear least squares in matrix form can be used to solve the regression problem.

The synaptic weights are commonly determined using linear regression of the hidden layer outputs to the real measured output after the RBF centers and non-linearities in the hidden layer have been fixed. In most cases, linear least squares in matrix form can be used to solve the regression problem.

$$\mathbf{W} = (\boldsymbol{Z}^T.\mathbf{Z})^{-1}.\boldsymbol{Z}^T.\boldsymbol{Y} \qquad (14)$$

where $\boldsymbol{Z} = [\boldsymbol{z}(1), \boldsymbol{z}(2), \dots, \boldsymbol{z}(F)]^T$ is a matrix containing the hidden layer responses for all input vectors. $W = [w_1, w_2, \dots, w_n]$ is a matrix containing all the synaptic weights for the output layer and converges to a scalar containing the target vector. The target vector y(f) carries the information of the value predicted by the f-th input vector.

2.5 Mariano-Preve Test

An explicit test of the null hypothesis for the purpose of validating equal predictive ability (EPA) of two competing forecasting models was introduced in the field of econometrics by Diebold and Mariano in their work titled 'Comparing Predictive Accuracy, 1995' this test method doesn't require any symmetric or quadratic relationship for the loss function and can be applied when the error distribution is non-Gaussian, has a non-zero mean, is serially and contemporaneously correlated [62]. However, this proposed asymptotic and exact finite sample test only compared two competing forecast and in cases where competing forecasts were more than two inferior methods had to be employed. In 2012, a further work was published that aimed at expanding the concept for three or more competing forecast [63]. This test is model free in the sense that it assumes that the only information available to the analyst is a time series of forecast and actual values of the prediction. The task of such analysist is to ascertain if all the models perform equally in terms of a specific loss function, which could be squared error or absolute error. Let the Equation (15) represent forecast errors of *k* competing models.

$$\{f_{it}\} = \{\hat{\mathrm{Y}}_{it} - y_t\}, \quad i = 1, 2, 3, \dots k \qquad (15)$$

And if $g: R \to R$ represents the utilized loss function. The null hypothesis states that all the models have equal predictive ability under the specified loss function defined below.

$$Eg(f_{1t}) = Eg(f_{2t}) = \cdots = Eg(f_{kt}) \qquad (16)$$

Consider the loss differential series $\{d_{jt}\}$ as expressed in Equation (17), the null hypothesis requires that the expectation of the loss differential $E\ d_t = 0$.

$$d_{jt} = g(f_{it}) - g(f_{i+1,t}), i = 1, 2, 3, \dots, k \qquad (17)$$

The test statistics d_s is based on the vector of observed sample means. It is represented by Equation (18)

$$d_s = \frac{1}{s}\sum_{s=1}^{s} d_t \quad (18)$$

Where s represents the sample size.

2.6 Cochran's test

In statistics, it is usually of importance to represent a finite population by a sample that will possess characteristics approximate to that of the population. In this study, our concern is a sampling distribution asymptotic to that of the population. William Cochran [59] developed a formula to aid the calculation of the minimum sample size with the ability to imitate the population. Cochran's method assumes that the population is normally distributed and attempts to verify this after computing the minimum required sample size. Let n denote the minimum sample size, Cochran proposes the equation (19) below.

$$n = \frac{n_o}{1 + \frac{n_o}{N}} \quad (19)$$

Where N denotes the size of a finite population and n_o can be represented by the equation (20) below.

$$n_o = \frac{Z^2 P(1-P)}{e^2} \quad (20)$$

Here *Z* denotes the z-score at confidence interval *e* and P represents the portion of the population assumed to generally represent the population characteristics. in this study, P is taken to be fifty percent of the entire population.

2.7 Significance Level estimation

There exists a threshold value, usually relating to the degree of significance, that validates the rejection of a hypothesis in most statistical tests. For the case of Kolmogorov-Smirnov's test used in this study, this author attempted at defining a threshold value to serve as sufficient proof for the rejection of a certain assertion. In this approach, confidence levels are calculated separately for each wind speed bin based on the modeled plot of wind speed and power output for each turbine, this value is usually found to be approximately equal for similar brand of turbines. The process involves binning the two-dimensional power curve, developed by the WTPC model, on wind speed basis and obtaining the geometric median for each bin. Afterwards, the Euclidean distance between the median point and all other data points within the bin are computed. The percentage variation between the largest distance value and the Euclidean distance from each median point to the reference x-plane will be utilized in calculating the confidence level. The Figure (4) below depicts a modeled plot of wind speed and power output after binning.

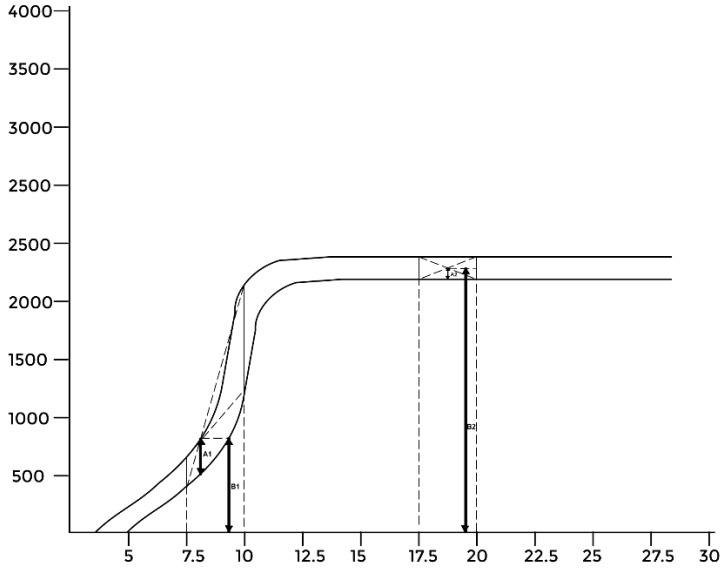


Figure 4: Significance level estimation from the modeled power curve

Let $\dot{X}$ represent the median point obtained after calculating the geometric median of data points within a specified bin (say. 7.5 m/s-10 m/s) and let the data points within the bin be represented by $X(P) = X_1, X_2, X_3, X_4, \dots, X_i$ the vector $Y(P)$ contains the distance of all data points within a bin to their bin-wise median point.

$$Y_i = EUCLIDEAN(|X_i - \dot{X}|) \tag{21}$$

Here Y_i denotes the distance between a data point and the median value $\dot{X}$. The distance between $\dot{X}$ and the reference x-plane is denoted by B. Therefore, the confidence level (CL) is expressed in the equation:

$$CL = \left[1 - (100 \times \frac{A_i}{B_i})\right] \tag{22}$$

A_i denotes the data point with the largest distance from the median point within a specific bin while B_i represents the distance between the median point and the reference x-plane for the specific bin.

2.8 Kolmogorov-Smirnov test

One of the ways of constructing limits for a set of probability distribution function and for taking the amount of statistical data into account is by using the Kolmogorov-Smirnov's test for the empirical cumulative distribution function (CDF) which is constructed for *m* observation and denoted by $F_m(x)$.

Consider the function $F(x)$ that represents a true probability distribution function of the observation, which in this case represents temporal instances of SCADA data and under an assumption that the PDF is unknown. If the observation set consists of *m* number of instances. A critical value of a test statistics $d_{m,1-\gamma}$ can be

calculated such that a width band $\pm d_{m,1-\gamma}$ as relating to $F_m(x)$ will entirely contain $F(x)$ under a significance of $(1-\gamma)$ interpreted as a confidence statement that signifies belief in a statistical framework. In such cases, a measure of the test statistics $D_m = \max|F_m(x) - F(x)|$ known as the Kolmogorov-Smirnov's test statistic is relevant for asserting the test under $\Pr\{D_m \geq d_{m,1-\gamma}\} = \gamma$. ways for computing $d_{m,1-\gamma}$, and for numerous values of m and γ , which are detailed in the work of Hubert [54]. A good approximation of the test statistics for $m > 10$ is shown by two equations according to [53]

$$d_{m,1-\gamma} \approx \frac{(1-\gamma)}{\sqrt{m}} \tag{23}$$

And

$$d_{m,1-\gamma} \approx (1-\gamma)(\sqrt{m} + 0.12 + 0.11\sqrt{m})^{-1} \tag{24}$$

In both cases the limits are the cumulative distribution functions (CDF) which are lower $F_m^l(x)$ and upper $F_m^u(x)$ bounds and are members of a known distribution function $F(x)$:

$$F_m^l(x) \leq F(x) \leq F_m^u(x) \tag{25}$$

Where

$$F_m^l(x) = \max\big(F_m(x) - d_{m,1-\gamma}, 0\big), \tag{26}$$

$$F_m^u(x) = \min\big(F_m(x) + d_{m,1-\gamma}, 1\big) \tag{27}$$

It is important to note that the K-S boundaries depend on the training examples *m.* it is seen from the inequality that the left of the upper boundary is $d_{m,1-\gamma}$ and the right of the lower boundary is $1 - d_{m,1-\gamma}$ these boundaries are located between boundary point of the sample space far apart.

2.9 Procedure

For a conceptual understanding of the technique proposed in this research, the author has defined the method by a sequence of six steps;

Step 1: from a SCADA data set of previous operations of a wind turbine, and containing necessary process parameters, use quantile based filtering technique to detect and remove anomalous data, the algorithm is stated below;

- Divide Data Frame into Sub-frames up to 50
- Define a single data-frame and set the power equal to the max power
- Define the probability distribution
- Apply quantiles to the probability distribution for each sub-frame in order to detect and remove outliers.
- Merge all data frames.
- END

Step 2: Use the filtered SCADA data set for training and validation of three choice model types.

- Multiple variables are inputted for model development other than wind speed only with an aim of capturing actual field conditions. The variables utilized are; Wind speed, density, blade pitch angle and temperature.
- For this study, the considered models are Radial Basis Function (RBF) architecture, multi-layer perceptron (MLP) architecture and gradient boost regressor (GBR). These models are considered because of their high estimation performance as recorded in the literature.

Step 3: Compare the loss distribution of the competing models for equal predictive ability (EPA).

- The Mariano-Preve's test for EPA is utilized in this step. The null hypothesis states that all competing models have EPA, the alternate hypothesis states that the expected loss differential between the models is not equal to zero under a 0.05 significance level.
- If the null hypothesis is rejected, an algorithm compares the correlation coefficient of the models and select the model with correlation coefficient closest to one as exhibiting superior predictive ability (SPA).

Step 4: The superior model is used in generating power values corresponding to various wind speed to serve as a modeled population from which inference about a normally functioning turbine can be drawn.

- The population is separated into bins of wind speed intervals to serve as distributions representing normal operating condition (NOC) of the wind turbine.
- The plot of wind speed and power output is used in generating bin specific significance levels to aid in Kolmogorov-Smirnov's test decision making.

Step 5: The most suitable sample size is calculated using Cochran's formula. This sample size represents the number of SCADA temporal instances required for the K-S test of similar distributions.

Step 6: The Kolmogorov-Smirnov's test ascertains if a sample is drawn from a certain distribution or not, this is used to differentiate normal and abnormal operation.

3 Results and Discussion

3.1 Anomaly Detection

The Figures (5a, 5b) represents data obtained from wind turbine SCADA systems. It can be seen from the plot that data obtained directly from SCADA systems contains numerous erroneous readings of various types: (Type I errors) errors generated when no power output is recorded at times when wind speed in significantly greater than the cut-in wind speed, (Type II errors) errors generated when the output power is constrained at higher values of wind speed and (Type III errors) errors typically generated by unsteady readings and are usually close to the designed value. Hence, there is need for an anomaly detection and filtering approach before being utilized for modelling. This study makes use of a quantile based approach which is established upon a hypothesis about the probability distribution of the SCADA population. It was discovered that faulty data appeared with less frequency compared to normal ones, if represented by a distribution, we could set quantiles to aid in the separation of normal from abnormal data. Using the quantile based filtration technique proper filtration results was achieved as can be seen form Figures (5c, 5d).

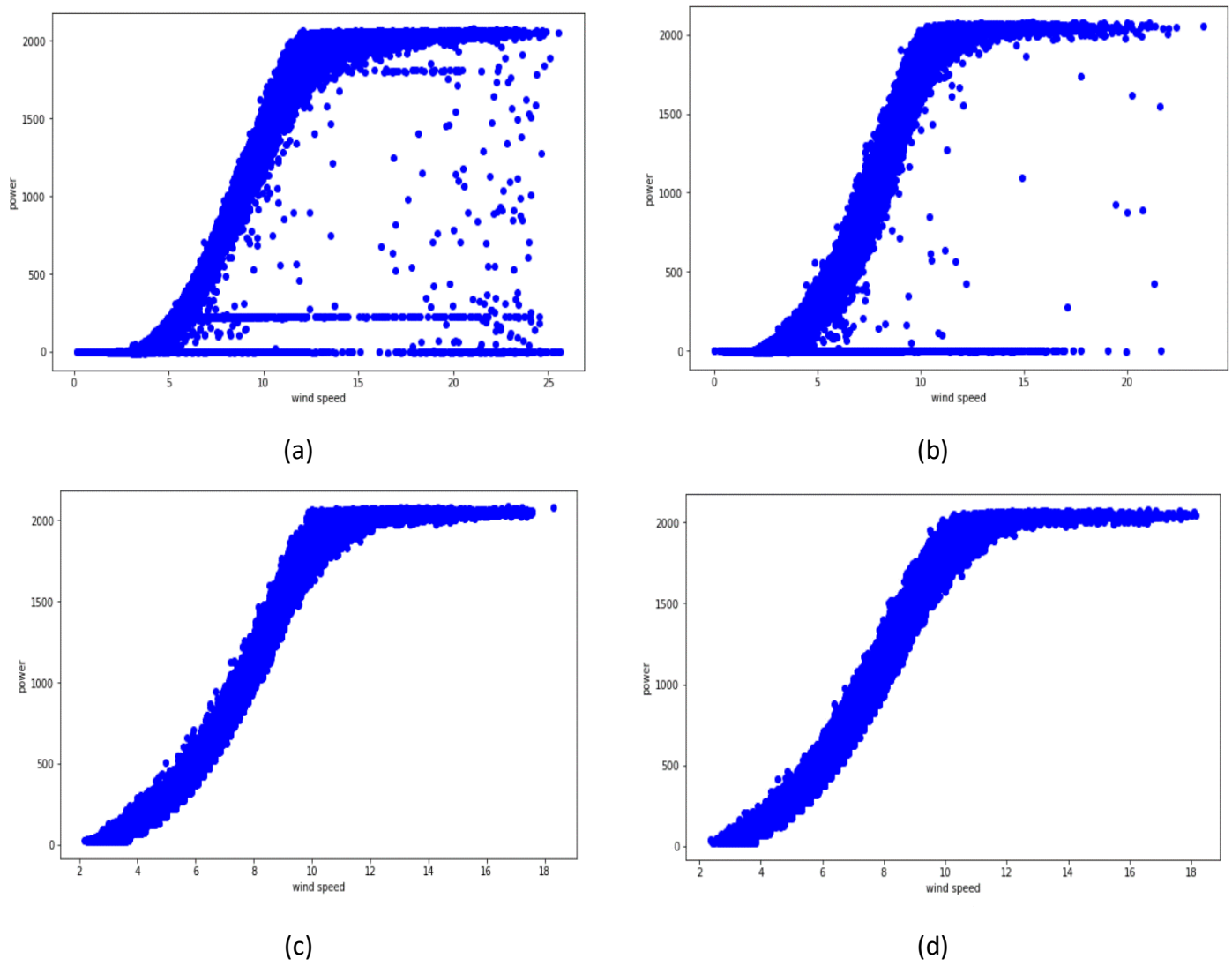


Figure 5: Plot (a) and (b) represent unfiltered SCADA data, plot (c) and (d) depicts filtered representations of the SCADA case study.

Visualization of a filtered plot usually isn't enough evidence that a filtration technique is performing optimally or at least comparative to alternatives developed and used in the literature. One method for validating the efficiency of a filtration technique is to compare its elimination rate with that of its numerous alternatives found in the literature. The Table (1) below details a comparison of the elimination rates of the filtration techniques used in this research with those found in the literature.

Method	Elimination rate (%)
QF	16
iForest	27.47
GMM	15
LOF	12

Table 1: The elimination rates of various filtration techniques are compared. This is performed with the aim of providing sufficient evidence about the performance of quantile based filtration technique used in this research. The results show that the elimination rate of the QF based method is comparative to those utilized in the literature. This further confirms the efficiency of the utilized filtration technique.

Results highlighted can be found in [64]. they include the utilized quantile based filtering (QF), isolation forest (iForest), Gaussian Mixture Modelling (GMM) and Local outlier factor (LOF).

3.2 Developed Power Curves

Two test metrics are used in evaluating SPA among the utilized models. The equations below detail their method. They are mean absolute error (MAE) and coefficient of determination (R^2).

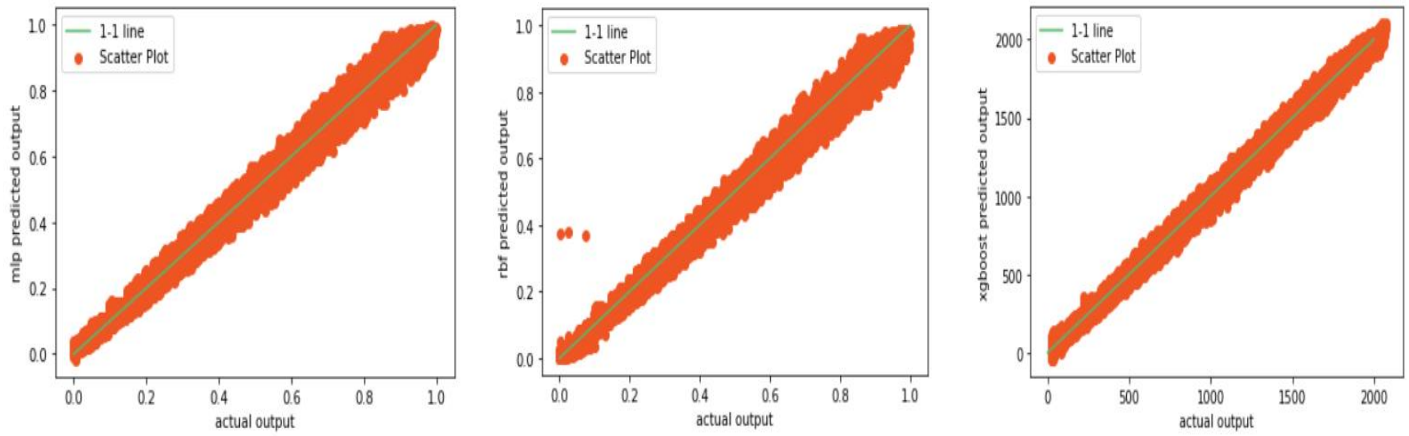
$$MAE = median(\ |\hat{Y}_n - Y_n|) \tag{28}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(\hat{Y}_n - Y_n)^2}{\sum_{n=1}^{N}(\hat{Y}_n - Y_n)^2} \tag{29}$$

The Table (2) details the average performance of all the models utilized for this study. The MLP network recorded an average MAE and R^2 score of 20.07 and 0.995 respectively, the RBF network resulted in MAE and R^2 values of 21.00 and 0.9949, while the gradient boosting method displayed 19.53 and 0.9953 in MAE and R^2 respectively. In a case where only the MAE and R^2 scores are analyzed, no consideration is made about the distribution of model losses. This analysis may be sufficient when there is considerable difference between the observed MAE and R^2 values. However, a problem arises when inference about the strength or weakness of the compared model is made when there isn't a significant difference in their observed MAE and R^2 values. In such cases, the forecast error distribution should be considered. This study considers the MAE values to assert that the choice models are competitive, SPA is validated by Mariano-Preve's test of the null hypothesis.

Model	No of Nodes	MAE	R^2
MLP-NN	1	20.07	0.9950
RBF-NN	1	21.00	0.9949
XGBoost	-	19.53	0.9953

Table 2: Comparison of the models based on MAE and R2. This signifies that the compared models are close to each other in terms of accuracy, with very little calculated differences. In this case, considerations have to be made regarding the distribution of model loss instead of visual comparison.



(a) (b) (c)

Figure 6: Actual vs Predicted power output for (a) MLP based model, (b) RBF based model and (c) gradient boosting based model. The y = x(1:1) identity line represents perfect correlation.

The Figure (6) indicates fitting of actual power values to estimated values of all for the purpose of comparison. The 1:1 line indicates perfect correlation between actual power and model prediction. As corroborated by the R^2

values in Table (2), it can be seen that there exists a high correlation between actual power and model prediction.

3.3 Test for Superior Predictive Ability

The multivariate Diebold and Mariano test can be used in comparing more than two competing forecasts for EPA. The original test properties inherited by the multivariate version includes no quadratic or symmetric requirement for the loss function, the errors realized during prediction should have a non-zero mean and can be non-Gaussian. The test statements are detailed below;

Null hypothesis:	H_0 will denote the null hypothesis. The null hypothesis states that all the models are performing equally or that all forecasting models have equal predictive ability. In statistics terms, the expectation for a *k-th* loss differential series $\boldsymbol{E}\ d_t$ will equal zero to confirm EPA.
Alternate hypothesis:	H_1 denotes the alternate hypothesis. When the expectation of the *K-th* loss differential series is not equal to zero, the alternate hypothesis is confirmed. This means the forecasting models do not have EPA.
If Null:	If the null hypothesis is confirmed, an algorithm randomly selects a model.
If Alternate:	If the null hypothesis is rejected, an algorithm compares the correlation coefficient (R^2) of all competing forecasts and select the model with an R^2 value closest to one.

In this study, we utilized a significance level of 95 percent to serve as sufficient deviation that indicates a rejection of the null hypothesis. The *q* value is obtained experimentally. In the research of Mariano *et al.* [63] q values from 1 to 4 were utilized, the most optimum observed value was 3, which we utilized in this research.

Dataset	S_c	q	Significance (α)	P-value	Decision
DATA 1	0.28449	3	0.05	0.8672	accept H_0
DATA 2	3.0217	3	0.05	0.2207	Reject H_0

Table 3: The Mariano-Preve's table of values. It indicates the test statistics (Sc), q-value, significance level (α), P-value and decision.

The test statistic is denoted by S_C. the values of the test statistic and p-value are calculated and utilized to confirm the acceptance or rejection of a hypothesis. A rejection of the null hypothesis is validated under significance level α whenever $S_C > \chi^2_{k,1-\alpha}$ where $\chi^2_{k,1-\alpha}$ is the (1- α) quantile of the chi-squared distribution, represented by (p-values).

The table (3) presents the results from Mariano-Preve's test. For Data 1 the test statistic is 0.28449, which is significantly lower than the obtained p-value of 0.8672. In this case, we accept the null hypothesis, leading to random selection of a forecasting model from the list. Data 2 presents a different result. The test statistic of 3.0217 is greater than the p-value of 0.2207, validating the rejection of the null hypothesis. In this case, the model with R^2 value closest to one is selected as having SPA.

3.4 Sample Size Estimation

A sample composed of data instances from real time operation of a wind turbine is needed for the validation of a fault by Kolmogorov-Smirnov's test. A larger sample size can also be used for K-S assessment however it will take a long time for such sample to be generated from real time instances to aid the K-S test. At this point the concern is acquiring the smallest sample that can inherit population characteristics in order to reduce the time taking for asserting faults. Cochran's formula for obtaining the smallest sample size that will retain all the information of its population was utilized in this study, it is explained in section (2.6). To utilize this method, certain parameters must be obtained; a data percentage that will have characteristics very close to that of the population, denoted by P, is selected to be 0.5 as it is obvious that a sample size half that of the population will inherit all the population characteristics to a large extent.

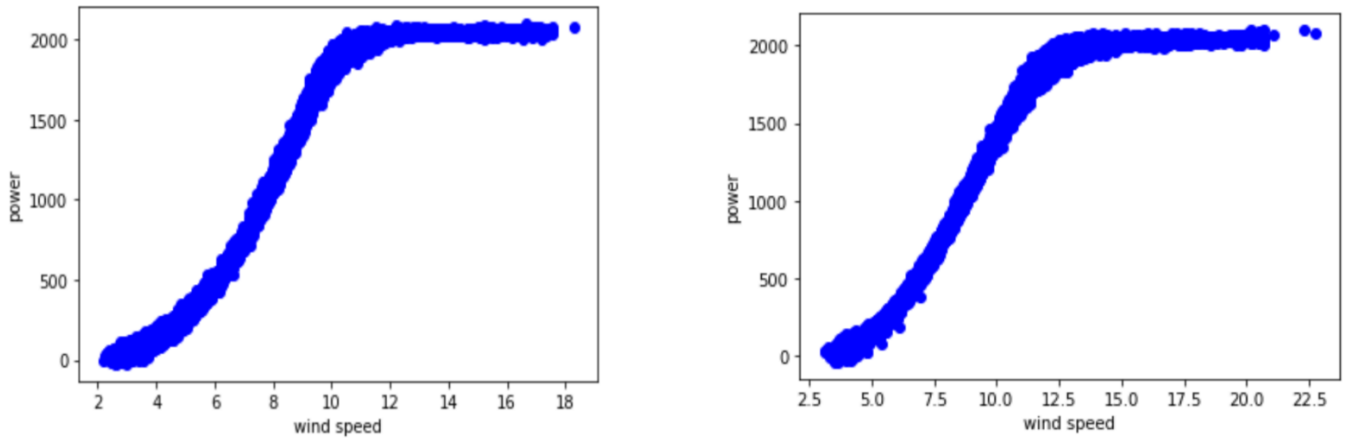
Dataset	Population	P	α	Z	ε	n_0	Sample size
DATA 1	56, 560	0.5	0.1	1.65	0.1	68.2	65
DATA 2	56, 500	0.5	0.1	1.65	0.1	67.4	64

Table 4: Cochran's test for obtaining the minimum sample size with the capacity to possess all the information of the population

The sample characteristics n_0 is estimated from confidence and error values, taken to both be 0.1. the test is carried out for the two turbine SCADA data set. It is found that the minimum sample size with the ability to inherit population characteristics with error ε and under confidence level α is 65, as shown in the Table (4).

3.5 Estimating confidence level

There exists a level of significance above which the K-S test should sufficiently reject the null hypothesis. Because the magnitude and frequency of the data varies between bins, the required significance level should be bin specific. This study presents a method for obtaining bin-wise confidence levels based on the modeled plot of wind speed and power output as is shown in the figure (7)



(a) (b)

Figure 7: Modeled power curve of most optimal model for data 1 (a) and data 2 (b), utilized for development of bin-wise confidence level and bin-wise ground truth for validating the K-S test.

A detailed explanation of this process is provided in section (2.7). bins contain wind speed intervals of 2.5 m/s. The farthest data point from the middle a (watt) is used alongside the distance from the middle point to the reference x-plane b(watt) for calculating the level of significance for each bin. It can be seen from Table (4 and 5) that the significance value α_s reduces with an increase in bin wind speed range corresponding to increased output power however, from the rated wind speed to the cut-out wind speed, the significance level remains the same.

	Dataset 1			
s/n	Bin (m/s)	a (watt)	b(watt)	α_s
1	2.5-5.0	10	10	100
2	5.0-7.5	150	500	0.3
3	7.5-10.0	150	1050	0.14
4	10.0-12.5	150	1750	0.085
5	12.5-15.0	120	2000	0.06
6	15.0-17.5	120	2000	0.06
7	17.5-20.0	120	2000	0.06

8	20.0-22.5	120	2000	0.06
9	22.5-25.0	120	2000	0.06

Table 5: Bin-wise confidence level estimates for data 1

	Dataset 2			
s/n	Bin (m/s)	a	b	α_s
1	2.5-5.0	5	5	100
2	5.0-7.5	100	500	0.200
3	7.5-10.0	135	1000	0.121
4	10.0-12.5	140	1550	0.090
5	12.5-15.0	125	1950	0.064
6	15.0-17.5	125	1950	0.064
7	17.5-20.0	125	1950	0.064
8	20.0-22.5	125	1950	0.064
9	22.5-25.0	125	1950	0.064

Table 6: Bin-wise confidence level estimate for data 2.

One caveat to this method is that the first and second bins cannot be used in asserting a fault. This is because their wind and power values are small and relatively close to zero. Hence do not possess deviations reasonable enough to confirm a claim, this can be seen from their recorded α_s values, which are relatively high. It is also worth noting that significance levels are similar regardless of the turbine investigated.

3.6 Fault Detection

In this research, it was hypothesized that abnormality (faults) could be detected by comparing the distribution of modeled power estimates within a specified bin with a sample of real time SCADA temporal instances of minimum sample size and wind speed values within the specified interval. One such test method applied to this kind of problem is the Kolmogorov-Smirnov's goodness of fit test. This test compares the cumulative distribution functions (CDF) of the modeled power output with the CDF of a sample of real time SCADA data, within a specified bin, by calculating the deviation in CDF and comparing it to a critical value defined by significance levels.

Null hypothesis H_0 will denote the null hypothesis. The null hypothesis states that the population and sample were drawn from similar distributions or that the test statistic D is a value less than Kolmogorov-Smirnov's critical value D_α defined by significance level α_s In this case, the test asserts that the turbine is performing normally.

Alternate hypothesis H_1 will denote the alternate hypothesis. Whenever the test statistics D reports a value that is greater than the Kolmogorov-Smirnov's critical value D_α the null hypothesis is rejected in favor of the alternate. This signals a fault or abnormality.

For Kolmogorov-Smirnov's test, one can assert a claim by comparing the test statistic D and K-S critical value D_α or by comparing the p-value and significance level α_s. In either cases, the result will be the same: Fault is confirmed whenever $(D \geq D_\alpha)$ or $(\alpha_s \geq p - value)$. The table (7) details the application of Kolmogorov-Smirnov's test for validating normal and faulty operations. For each data sample, three tests were performed; two of which were normal operation without any associated faulty feedback, one of which is a sample taken few days before SCADA system recorded faults that led to downtime. In cases where normal samples were analyzed, the Kolmogorov-Smirnov test displayed results that validated an acceptance of the null hypothesis while in cases where the test sample was taken few days before the emergence of a fault that caused downtime,

the results obtained from Kolmogorov-Smirnov's validation indicated the rejection of the null hypothesis in favor of the alternate.

Dataset	Bin(m/s)	Nature of Sample	Sample Time	α_s	D	D_α	P-value	Decision
Data 1	10-12.5	Normal	05/06/2020	0.090	0.0876	0.115	0.6788	Accept H_0
Data 1	12.5-15	Normal	01/07/2020	0.064	0.0459	0.1455	0.9986	Accept H_0
Data 1	10-12.5	Faulty	03/11/2020	0.090	0.178	0.115	0.0248	**Reject** H_0
Data 2	10-12.5	Normal	15/03/2020	0.090	0.0831	0.115	0.7394	Accept H_0
Data 2	12.5-15	Normal	19/09/2020	0.064	0.1235	0.1455	0.2360	Accept H_0
Data 2	12.5-15	Faulty	01/06/2020	0.064	0.1787	0.1455	0.0248	**Reject** H_0

Table 7: Result table detailing the outcomes of Kolmogorov-Smirnov test. It can be seen that the Kolmogorov-Smirnov test accurately detects Normal and Faulty SCADA samples by either accepting the null hypothesis when a normal sample is tested, or by rejecting the null hypothesis when the sample is faulty.

The power distribution as specified within wind speed intervals is that which is analyzed by the Kolmogorov-Smirnov test. That being said, Phong B. Dao [25] identified two other process parameters, asides output power, that provides optimal fault detection result out of seven parameters investigated, they are; gearbox bearing temperature and generator speed. With the aim of providing further validation for faults detected using K-S analysis of the output power, an assessment of both gearbox bearing temperature and generator speed is provided. For this investigation, SCADA data corresponding to four process parameters, and captured few days before a fault is detected, is utilized. The data point captured is of minimum sample size as required for K-S evaluation.

The Figure (8) shows that the wind speed at the captured time interval is stochastic in nature, falling between (18.0 m/s and 24 m/s). it is also observed that after the 24th instance, the wind speed exhibits higher frequency around 20 m/s.

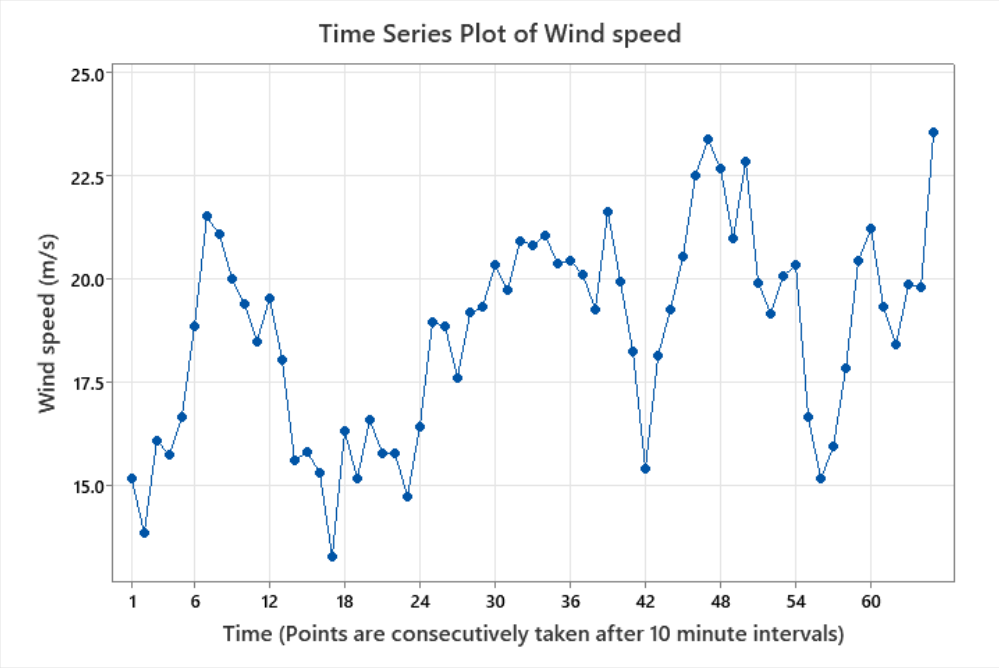


Figure 8: Temporal plot of wind speed, of minimum sample size and collected few days before fault was realized in data 2.

As the wind speed in Figure (8) frequently exhibits high values around 20 m/s, from the 24th instance, the output power in Figure (9) is seen to decrease slightly. It is also observed that the output power attempts to correct itself back to the rated power range. This change in power behavior within specified bins causes a significant shift of the frequency distribution that leads to large computed distance between the observed CDFs as calculated by K-S test statistics.

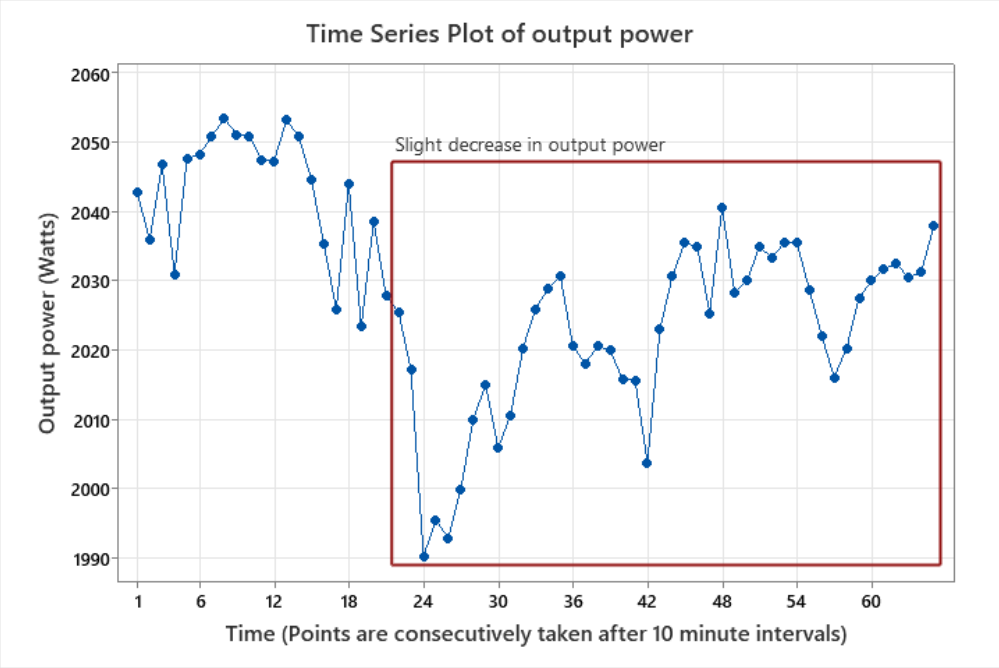


Figure 9: Temporal plot of power output, of minimum sample size and collated few days before a fault was detected in data 2. It corresponds to the wind speed plot.

The anomaly in power distribution from Figure (9) can be related to an increase in gearbox bearing temperature as shown in Figure (10). It is observed that as the output power decreased slightly, there was a simultaneous increase in gearbox bearing temperature, from an interval between 64 ^{o}C and 74 ^{o}C to an interval between 71 ^{o}C and 80 ^{o}C. This increase is rather minimal and may not be detected by a wind farm operator. Additionally, it was identified that the K-S test could be performed for gearbox bearing temperature and will yield similar results as obtained from the K-S test performed on the output power.

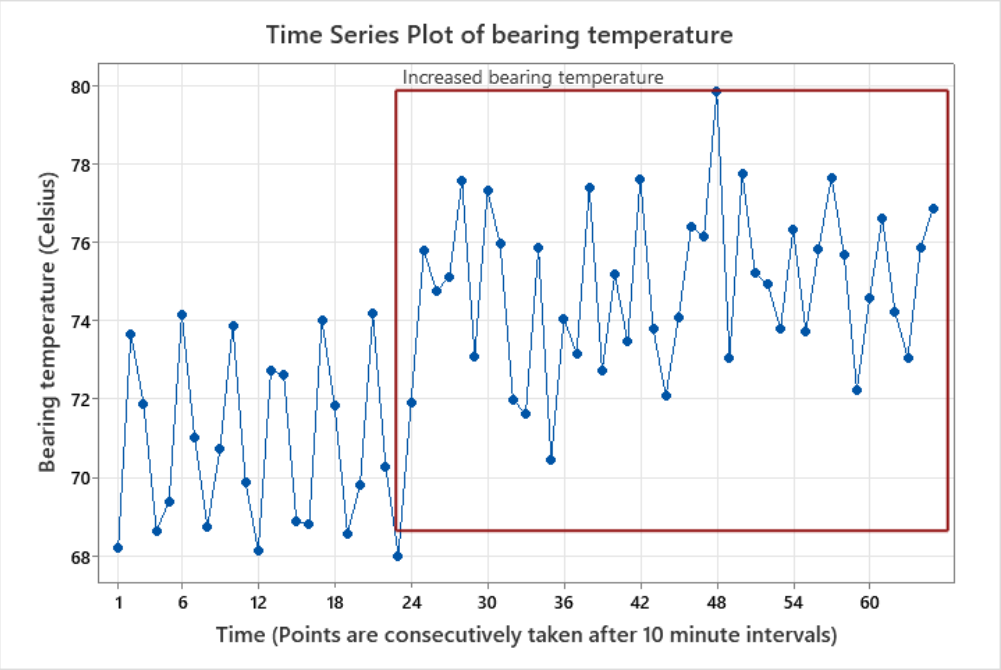


Figure 10: Temporal plot of bearing temperature. Data points are of minimum sample size and collated a few days before fault was detected In data 2.

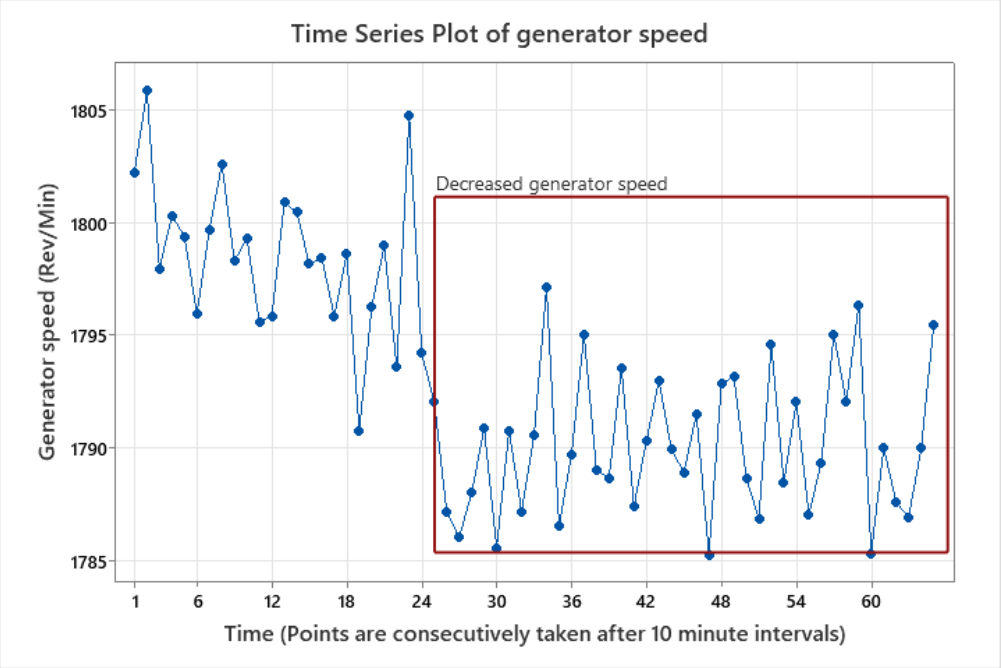


Figure 11. Temporal plot of generator speed containing instances equivalent to the minimum sample size and taken a few days before a fault was detected.

The sight decrease in output power and increase in gearbox bearing temperature is as shown in Figures (9 and 10) is accompanied by a decrease in generator speed as shown in Figure (11). The generator speed generally exhibits a downward trend, the rate of which increases after the 24th instance. The analysis entails 65 instances corresponding to the minimum sample size that will possess the ability to retain population information as

proposed by William Cochran. the generator speed is observed to decrease from 1806 rpm at the 2nd instance to about 1798 rpm at the 20th instance. The rate of this decrease is observed to improves around the 20th instance, from about 1804 rpm at the 23rd instance to 1785 rpm at the 60th instance. This is proof that the results obtained from an analysis of generator speed further corroborate the result obtained from performing K-S test on the output power. Generator speed will also yield favorable results if utilized as variable for the K-S test.

3.7 Discussion

Given the results from each of the sections in this paper, a discussion is necessary in order to highlight the benefits and limitations identified during the case study:

- The quantile based filtration technique as found in sections 2.1 and 3.1, detects and eliminates faulty data by defining quantiles on a distribution of SCADA data, on the basis of a process parameter. Output power was used in this research. It is important to note that the most optimal quantile must be obtained experimentally and is user defined. In other words, the efficiency of this technique depends on several iterations performed with different quantile values alongside an assessment of the filtered plot of wind speed and power output and the elimination rate. If a good quantile value is realized, this method could yield better performance than other techniques employed for data filtration in the literature.
- This study employs RBF, MLP and GBR for developing power curves that are compared for the extraction of bin-wise information required for K-S test. MLP and RBF are neural network based methods while GBR is based on a decision tree algorithm. While it is possible that one of the competing models was disadvantaged due to the selection of suboptimal hyper-parameters, the authors attempted using the most optimal hyper-parameter to the best of their knowledge.
- The superior forecasting model will possess better bin information than any inferior alternative. In turn, this will reduce the tendency of realizing a type 1 (incorrect rejection of the null hypothesis) or type 2 (failure to reject the null hypothesis when true) error. In cases where forecasting options do not have EPA, it was found that by analyzing the extent to which the predicted power linearly correlates to the observed values for each of the forecasting options, SPA could be confirmed.
- Cochran's method for sample size estimation aids the acquisition of a minimum sample size with sampling distribution similar to that of the population under a specified significance level. A significance level (α) of 0.05 indicates that 1 out of 20 samples will not possess adequate information from the population. If this aberrant sample is used for K-S test, it will result in a false positive or type 1 error. An analysis of other relating process parameters can be used to identify such samples. If the fault obtained from K-S test isn't validated by the other process parameters, the sample should be regarded as a false positive.
- A bin comprises parameters corresponding to wind speed of 2.5 m/s intervals. The first bin contains parameters corresponding to wind speed values between 2.5 m/s and 5.0 m/s, the last bin is defined by wind speed values between 22.5 m/s and 25.0 m/s. all bins situated after the rated wind speed possess the same K-S significance value (α_s) while the value of α_s decreases rapidly for bins before the rated wind speed. in cases where the wind speed is highly dynamic and intersects several bins located after the rated wind speed, analysis could be carried out by merging them together because their α_s values are the same. On the other hand, the first and second bin represented with intervals of (2.5 m/s-5.0 m/s) and (5.0 m/s -7.5 m/s) respectively should not be used for K-S analysis as results obtained using them will be erroneous due to the nature of their α_s.

- Kolmogorov-Smirnov's goodness of fit test proves to be an efficient method for comparing the CDF of samples. K-S evaluation is performed on the output power in this work, this is because α_s was generated based on the wind turbine power curve. It is observed that K-S evaluation can be performed on gearbox bearing temperature as well as generator rpm, achieving optimal results. However, it isn't clear whether α_s obtained from analyzing the power curve could be applied for such analysis.

4 Conclusion

This study presents a new method for operational state monitoring and fault detection of wind turbines based on the modeled power curve and Kolmogorov-Smirnov's nonparametric goodness of fit test. Machine learning (ML) has been applied in recent times for conditional monitoring and fault detection of wind turbines however, they are known to possess some limitations; long training time, enormous computational cost and large data requirement. Additionally, ML methods utilized in the literature decrease in performance over time due to turbine ageing, change of critical components and sensor recalibration. These reasons present the need for more reliable and efficient solutions to the problem of CM and automatic fault detection of wind turbines. The method investigated in this study attempts the validation of faults by a comparison of distributions. Filtered SCADA from previously operating turbines are used for the development of a WTPC model of optimal performance. Three modelling architectures namely; radial basis function, multi-layer perceptron and gradient boosting are compared in order to identify the most optimal alternative. Comparison is performed using Mariano-Preve and linear correlation test. distribution information from each wind speed interval (bin) alongside confidence levels are extracted from the most optimal model alternative, to be applied for Kolmogorov-Smirnov's test of the similarity between CDFs. the minimum sample size to minimize erroneous results is calculated using Cochran's formula and used to specify the amount of SCADA data temporal instances needed to assert a fault or abnormality. The null hypothesis assumes that both samples are drawn from similar distributions or have similar CDFs, if the null hypothesis is rejected, a fault claim is asserted. this claim is assessed in relation to other process parameters known to provide useful information on the imminence of a fault before it is confirmed, they are; gearbox bearing temperature and generator speed.

Two SCADA data, associated with two fault events, were utilized as case studies for confirming the technique proposed in this research. One major advantage of this method, asides fast computation, is that it is invariant to performance degradation over time as is the case with ML and other parametric and nonparametric approaches such as; neural networks, decision trees, support vector machines, co-integration method, CUSUM-based approach, Chow test and Wilcoxon rank sum test approach. The monitoring and fault detection process is also very simple. This method can find application in a plethora of sectors in engineering asides wind energy, provided there is sufficient historical data for the formation of a frequency distribution and a system for acquiring real time data to be used for comparison.

The certainty of a fault is confirmed after analyzing other process parameters known to provide acceptable fault signals. A system can be developed where Kolmogorov-Smirnov's test is performed on the three significant process parameter simultaneously to completely eliminate type 1 and type 2 errors. Currently, there exists no method for obtaining the significance levels needed for decision making for gearbox bearing temperature and generator speed data. Hence, future works should focus on developing a method for obtaining confidence levels that will form decision boundaries for the acceptance or rejection of a hypothesis.

Supplementary materials

Supplementary material for this this article is found at:

[A] Kelmarsh wind farm data. https://zenodo.org/record/5841834#.YsGnSHbMLIU

[B] Penmanshiel Wind Farm Data. https://zenodo.org/record/5946808#.YsGnSXbMLIU

[C] Code Link: https://github.com/henrii1/WTPCM-using-multivariate-DM-and-K-S-test

Reference

[1] T. Guo *et al.*, "Nacelle and tower effect on a stand-along wind turbine energy output—A discussion on field measurements of a small wind turbine," *Appl. Energy*, vol. 303, no. September, 2021, doi: 10.1016/j.apenergy.2021.117590.

[2] R. He, H. Yang, H. Sun, and X. Gao, "A novel three-dimensional wake model based on anisotropic Gaussian distribution for wind turbine wakes," *Appl. Energy*, vol. 296, no. April, p. 117059, 2021, doi: 10.1016/j.apenergy.2021.117059.

[3] "Global Wind Energy Council. Global Wind Report: Annual Market Update 2021." https://gwec.net/global-wind-report-2021/.

[4] B. Akay, D. Ragni, C. S. Ferreira, and G. J. W. Van Bussel, "Investigation of the root flow in a Horizontal Axis," *Wind Energy*, pp. 1–20, 2013, doi: 10.1002/we.

[5] M. De Prada Gil, O. Gomis-Bellmunt, and A. Sumper, "Technical and economic assessment of offshore wind power plants based on variable frequency operation of clusters with a single power converter," *Appl. Energy*, vol. 125, pp. 218–229, 2014, doi: 10.1016/j.apenergy.2014.03.031.

[6] A. Kusiak and A. Verma, "A data-mining approach to monitoring wind turbines," *IEEE Trans. Sustain. Energy*, vol. 3, no. 1, pp. 150–157, 2012, doi: 10.1109/TSTE.2011.2163177.

[7] A. Kusiak and W. Li, "The prediction and diagnosis of wind turbine faults," *Renew. Energy*, vol. 36, no. 1, pp. 16–23, 2011, doi: 10.1016/j.renene.2010.05.014.

[8] P. Sun, J. Li, C. Wang, and X. Lei, "A generalized model for wind turbine anomaly identification based on SCADA data," *Appl. Energy*, vol. 168, pp. 550–567, 2016, doi: 10.1016/j.apenergy.2016.01.133.

[9] F. P. García Márquez, A. M. Tobias, J. M. Pinar Pérez, and M. Papaelias, "Condition monitoring of wind turbines: Techniques and methods," *Renew. Energy*, vol. 46, pp. 169–178, 2012, doi: 10.1016/j.renene.2012.03.003.

[10] F. Castellani, D. Astolfi, P. Sdringola, S. Proietti, and L. Terzi, "Analyzing wind turbine directional behavior: SCADA data mining techniques for efficiency and power assessment," *Appl. Energy*, vol. 185, pp. 1076–1086, 2017, doi: 10.1016/j.apenergy.2015.12.049.

[11] A. Stetco *et al.*, "Machine learning methods for wind turbine condition monitoring: A review," *Renew. Energy*, vol. 133, pp. 620–635, 2019, doi: 10.1016/j.renene.2018.10.047.

[12] E. Artigao, S. Martín-Martínez, A. Honrubia-Escribano, and E. Gómez-Lázaro, "Wind turbine reliability: A comprehensive review towards effective condition monitoring development," *Appl. Energy*, vol. 228, no. May, pp. 1569–1583, 2018, doi: 10.1016/j.apenergy.2018.07.037.

[13] S. Soua, P. Van Lieshout, A. Perera, T. H. Gan, and B. Bridge, "Determination of the combined vibrational and acoustic emission signature of a wind turbine gearbox and generator shaft in service as a pre-requisite for effective condition monitoring," *Renew. Energy*, vol. 51, pp. 175–181, 2013, doi: 10.1016/j.renene.2012.07.004.

[14] Z. Feng, X. Chen, and M. Liang, "Iterative generalized synchrosqueezing transform for fault diagnosis of wind turbine planetary gearbox under nonstationary conditions," *Mech. Syst. Signal Process.*, vol. 52–53, no. 1, pp. 360–375, 2015, doi: 10.1016/j.ymssp.2014.07.009.

[15] C. Q. Gómez Muñoz, F. P. García Márquez, and J. M. Sánchez Tomás, "Ice detection using thermal

infrared radiometry on wind turbine blades," *Meas. J. Int. Meas. Confed.*, vol. 93, pp. 157–163, 2016, doi: 10.1016/j.measurement.2016.06.064.

[16] W. Liu *et al.*, "Thermal degradation mechanism of poly(hexamethylene carbonate)," *Polym. Degrad. Stab.*, vol. 112, pp. 70–77, 2015, doi: 10.1016/j.polymdegradstab.2014.12.013.

[17] G. de Novaes Pires Leite *et al.*, "Alternative fault detection and diagnostic using information theory quantifiers based on vibration time-waveforms from condition monitoring systems: Application to operational wind turbines," *Renew. Energy*, vol. 164, no. August 2019, pp. 1183–1194, 2021, doi: 10.1016/j.renene.2020.10.129.

[18] M. Xu, G. Feng, Q. He, F. Gu, and A. Ball, "Vibration characteristics of rolling element bearings with different radial clearances for condition monitoring of wind turbine," *Appl. Sci.*, vol. 10, no. 14, 2020, doi: 10.3390/app10144731.

[19] S. Dey, P. Pisu, and B. Ayalew, "A Comparative Study of Three Fault Diagnosis Schemes for Wind Turbines," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 5, pp. 1853–1868, 2015, doi: 10.1109/TCST.2015.2389713.

[20] P. F. Odgaard and J. Stoustrup, "A benchmark evaluation of fault tolerant wind turbine control concepts," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 1221–1228, 2015, doi: 10.1109/TCST.2014.2361291.

[21] J. Maldonado-Correa, S. Martín-Martínez, E. Artigao, and E. Gómez-Lázaro, "Using SCADA data for wind turbine condition monitoring: A systematic literature review," *Energies*, vol. 13, no. 12, 2020, doi: 10.3390/en13123132.

[22] P. B. Dao, W. J. Staszewski, T. Barszcz, and T. Uhl, "Condition monitoring and fault detection in wind turbines based on cointegration analysis of SCADA data," *Renew. Energy*, vol. 116, pp. 107–122, 2018, doi: 10.1016/j.renene.2017.06.089.

[23] E. Gonzalez, B. Stephen, D. Infield, and J. J. Melero, "Using high-frequency SCADA data for wind turbine performance monitoring: A sensitivity study," *Renew. Energy*, vol. 131, pp. 841–853, 2019, doi: 10.1016/j.renene.2018.07.068.

[24] K. S. Wang, V. S. Sharma, and Z. Y. Zhang, "SCADA data based condition monitoring of wind turbines," *Adv. Manuf.*, vol. 2, no. 1, pp. 61–69, 2014, doi: 10.1007/s40436-014-0067-0.

[25] P. B. Dao, "On Wilcoxon rank sum test for condition monitoring and fault detection of wind turbines," *Appl. Energy*, vol. 318, no. May, p. 119209, 2022, doi: 10.1016/j.apenergy.2022.119209.

[26] J. Tautz-Weinert and S. J. Watson, "Using SCADA data for wind turbine condition monitoring - A review," *IET Renew. Power Gener.*, vol. 11, no. 4, pp. 382–394, 2017, doi: 10.1049/iet-rpg.2016.0248.

[27] P. B. Dao, "Condition monitoring and fault diagnosis of wind turbines based on structural break detection in SCADA data," *Renew. Energy*, vol. 185, pp. 641–654, 2022, doi: 10.1016/j.renene.2021.12.051.

[28] A. Kusiak and Z. Zhang, "Analysis of wind turbine vibrations based on SCADA data," *J. Sol. Energy Eng. Trans. ASME*, vol. 132, no. 3, pp. 0310081–03100812, 2010, doi: 10.1115/1.4001461.

[29] W. Yang, R. Court, and J. Jiang, "Wind turbine condition monitoring by the approach of SCADA data analysis," *Renew. Energy*, vol. 53, pp. 365–376, 2013, doi: 10.1016/j.renene.2012.11.030.

[30] M. Schlechtingen, I. F. Santos, and S. Achiche, "Wind turbine condition monitoring based on SCADA data

using normal behavior models. Part 1: System description," *Appl. Soft Comput. J.*, vol. 13, no. 1, pp. 259–270, 2013, doi: 10.1016/j.asoc.2012.08.033.

[31] A. Meyer, "Multi-target normal behaviour models for wind farm condition monitoring," *Appl. Energy*, vol. 300, p. 117342, 2021, doi: 10.1016/j.apenergy.2021.117342.

[32] F. Qu, J. Liu, H. Zhu, and B. Zhou, "Wind turbine fault detection based on expanded linguistic terms and rules using non-singleton fuzzy logic," *Appl. Energy*, vol. 262, no. November 2019, p. 114469, 2020, doi: 10.1016/j.apenergy.2019.114469.

[33] D. F. Lin, P. H. Chen, and M. Williams, "Measurement and analysis of current signals for gearbox fault recognition of wind turbine," *Meas. Sci. Rev.*, vol. 13, no. 2, pp. 89–93, 2013, doi: 10.2478/msr-2013-0010.

[34] S. Pei and Y. Li, "Wind turbine power curve modeling with a hybrid machine learning technique," *Appl. Sci.*, vol. 9, no. 22, 2019, doi: 10.3390/APP9224930.

[35] I. Abdallah *et al.*, "Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data," *Saf. Reliab. - Safe Soc. a Chang. World - Proc. 28th Int. Eur. Saf. Reliab. Conf. ESREL 2018*, no. iii, pp. 3053–3062, 2018, doi: 10.1201/9781351174664-382.

[36] M. Morshedizadeh, M. Kordestani, R. Carriveau, D. S. K. Ting, and M. Saif, "Improved power curve monitoring of wind turbines," *Wind Eng.*, vol. 41, no. 4, pp. 260–271, 2017, doi: 10.1177/0309524X17709730.

[37] A. Pliego Marugán, A. M. Peco Chacón, and F. P. García Márquez, "Reliability analysis of detecting false alarms that employ neural networks: A real case study on wind turbines," *Reliab. Eng. Syst. Saf.*, vol. 191, no. May, p. 106574, 2019, doi: 10.1016/j.ress.2019.106574.

[38] Z. Kong, B. Tang, L. Deng, W. Liu, and Y. Han, "Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units," *Renew. Energy*, vol. 146, pp. 760–768, 2020, doi: 10.1016/j.renene.2019.07.033.

[39] A. P. Marugán, F. P. G. Márquez, J. M. P. Perez, and D. Ruiz-Hernández, "A survey of artificial neural network in wind energy systems," *Appl. Energy*, vol. 228, no. June, pp. 1822–1836, 2018, doi: 10.1016/j.apenergy.2018.07.084.

[40] M. Schlechtingen and I. Ferreira Santos, "Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection," *Mech. Syst. Signal Process.*, vol. 25, no. 5, pp. 1849–1875, 2011, doi: 10.1016/j.ymssp.2010.12.007.

[41] P. B. Dao, "A CUSUM-based approach for condition monitoring and fault diagnosis of wind turbines," *Energies*, vol. 14, no. 11, 2021, doi: 10.3390/en14113236.

[42] P. B. Dao, W. J. Staszewski, T. Barszcz, and T. Uhl, "Condition monitoring and fault detection in wind turbines based on cointegration analysis of SCADA data," *Renew. Energy*, vol. 116, pp. 107–122, 2018, doi: 10.1016/j.renene.2017.06.089.

[43] P. B. Dao, "Analysis of gearbox and generator temperature data," *Diagnostyka*, vol. 19, no. 1, pp. 63–71, 2018, doi: 10.29354/diag/81298.

[44] A. Llombart, S. J. Watson, D. Llombart, and J. M. Fandos, "Power curve characterization I: Improving the bin method," *Renew. Energy Power Qual. J.*, vol. 1, no. 3, pp. 367–371, 2005, doi: 10.24084/repqj03.304.

[45] J. Gottschall and J. Peinke, “Stochastic modelling of a wind turbine’s power output with special respect to turbulent dynamics,” *J. Phys. Conf. Ser.*, vol. 75, no. 1, 2007, doi: 10.1088/1742-6596/75/1/012045.

[46] M. Marčiukaitis, I. Žutautaitė, L. Martišauskas, B. Jokšas, G. Gecevičius, and A. Sfetsos, “Non-linear regression model for wind turbine power curve,” *Renew. Energy*, vol. 113, pp. 732–741, 2017, doi: 10.1016/j.renene.2017.06.039.

[47] A. Kusiak and A. Verma, “Monitoring wind farms with performance curves,” *IEEE Trans. Sustain. Energy*, vol. 4, no. 1, pp. 192–199, 2013, doi: 10.1109/TSTE.2012.2212470.

[48] D. Villanueva and A. Feijóo, “Comparison of logistic functions for modeling wind turbine power curves,” *Electr. Power Syst. Res.*, vol. 155, pp. 281–288, 2018, doi: 10.1016/j.epsr.2017.10.028.

[49] D. Karamichailidou, V. Kaloutsa, and A. Alexandridis, “Wind turbine power curve modeling using radial basis function neural networks and tabu search,” *Renew. Energy*, vol. 163, pp. 2137–2152, 2021, doi: 10.1016/j.renene.2020.10.020.

[50] F. Pelletier, C. Masson, and A. Tahan, “Wind turbine power curve modelling using artificial neural network,” *Renew. Energy*, vol. 89, pp. 207–214, 2016, doi: 10.1016/j.renene.2015.11.065.

[51] B. Manobel, F. Sehnke, J. A. Lazzús, I. Salfate, M. Felder, and S. Montecinos, “Wind turbine power curve modeling based on Gaussian Processes and Artificial Neural Networks,” *Renew. Energy*, vol. 125, pp. 1015–1020, 2018, doi: 10.1016/j.renene.2018.02.081.

[52] R. K. Pandit, D. Infield, and A. Kolios, “Gaussian process power curve models incorporating wind turbine operational variables,” *Energy Reports*, vol. 6, pp. 1658–1669, 2020, doi: 10.1016/j.egyr.2020.06.018.

[53] G. Zhang, X. Wang, Y. C. Liang, and J. Liu, “Fast and robust spectrum sensing via Kolmogorov-Smirnov test,” *IEEE Work. Local Metrop. Area Networks*, vol. 58, no. 12, pp. 3410–3416, 1993, doi: 10.1109/TCOMM.2010.11.090209.

[54] F. Baselice, G. Ferraioli, V. Pascazio, and A. Sorriso, “Denoising of MR images using Kolmogorov-Smirnov distance in a Non Local framework,” *Magn. Reson. Imaging*, vol. 57, no. November 2018, pp. 176–193, 2019, doi: 10.1016/j.mri.2018.11.022.

[55] D. Dos Reis, P. Flach, S. Matwin, and G. Batista, “Fast unsupervised online drift detection using incremental kolmogorov-smirnov test,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, no. 1, pp. 1545–1554, 2016, doi: 10.1145/2939672.2939836.

[56] M. S. Kovalev and L. V. Utkin, “A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov–Smirnov bounds,” *Neural Networks*, vol. 132, pp. 1–18, 2020, doi: 10.1016/j.neunet.2020.08.007.

[57] P. P. Oktaviana and Irhamah, “Kolmogorov-Smirnov Goodness-of-Fit test for identifying distribution of the number of earthquakes and the losses due to earthquakes in Indonesia,” *J. Phys. Conf. Ser.*, vol. 1821, no. 1, 2021, doi: 10.1088/1742-6596/1821/1/012045.

[58] A. Reinhart, V. Ventura, and A. Athey, “Detecting changes in maps of gamma spectra with Kolmogorov-Smirnov tests,” *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 802, pp. 31–37, 2015, doi: 10.1016/j.nima.2015.09.002.

[59] W. G. Cochran, “Samling Techniques,” p. 448, 1977.

[60] H. Rao *et al.*, “Feature selection based on artificial bee colony and gradient boosting decision tree,” *Appl. Soft Comput. J.*, vol. 74, pp. 634–642, 2019, doi: 10.1016/j.asoc.2018.10.036.

[61] D. L. D.S. Broomhead, “Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Nctworks.,” *Emerg. Technol. Situ Process.*, vol. 139, no. 4148, p. 221, 1988.

[62] F. X. Diebold, “Comparing Predicitve accuracy,” vol. 20, no. 1, pp. 134–144, 2008.

[63] R. S. Mariano and D. Preve, “Statistical tests for multiple forecast comparison,” *J. Econom.*, vol. 169, no. 1, pp. 123–130, 2012, doi: 10.1016/j.jeconom.2012.01.014.

[64] R. Morrison, X. Liu, and Z. Lin, “Anomaly detection in wind turbine SCADA data for power curve cleaning,” *Renew. Energy*, vol. 184, pp. 473–486, 2022, doi: 10.1016/j.renene.2021.11.118.