Wind power prediction using bootstrap aggregating trees approach to enabling sustainable wind power integration in a smart grid

Fouzi Harrou\textsuperscript{a}, Ahmed Saidi\textsuperscript{b}, Ying Sun\textsuperscript{a}

\textsuperscript{a}King Abdullah University of Science and Technology (KAUST)
\hspace{0.5cm}Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal 23955-6900, Saudi Arabia
\hspace{0.5cm}E-mail: fouzi.harrou@kaust.edu.sa

\textsuperscript{b}ENERGARID Laboratory, Electrical Engineering Department, Tahri Mohammed University, BP 417 Route de Kenadsa, Béchar, Algeria

Abstract

Precise prediction of wind power is important in sustainably integrating the wind power in a smart grid. The need for short-term predictions is increased with the increasing installed capacity. The main contribution of this work is adopting bagging ensembles of decision trees approach for wind power prediction. The choice of this regression approach is motivated by its ability to take advantage of many relatively weak single trees to reach a high prediction performance compared to single regressors. Moreover, it reduces the overall error and has the capacity to merge numerous models. The performance of bagged trees for predicting wind power has been compared to four commonly know prediction methods namely multivariate linear regression, support vector regression, principal component regression, and partial least squares regression. Real measurements recorded every ten minutes from an actual wind turbine are used to illustrate the prediction quality of the studied methods. Results showed that the bagged trees regression approach reached the highest prediction performance with a coefficient of determination of 0.982. The result showed that the bagged trees approach is followed by support vector regression with Gaussian kernel, the same model when using a quadratic kernel, and the multivariate linear regression, partial least squares, and principal component regression gave the lowest prediction. The investigated models in this study can represent a helpful tool for model-based anomaly detection in wind turbines.

Keywords: Power prediction, Wind turbine, Ensemble Bagged Trees, regression models, Latent variable models, SVR.

1. Introduction

Wind power is one of the most potential energies and the major available renewable energy sources. It is one of the most competitive ways of adding new power generation to the grid in an expanding amount. The global capacity of wind power has reached 539 GW in 2017, with an increase of 52.5 GW compared to 2016 (Figure 1). According to the world wind energy association, the overall capacity of all wind turbines installed by the end of 2018 attained 600 Gigawatt. Hence, the requirement to operate these sources becomes
a mandatory obligation to reduce environment pollution and participate in sustainable energy development.

The main crucial and challenging issue in wind power production is its volatility intermittent due to mainly to weather conditions [1]. This fact usually makes the integration of wind turbines into the power grid, not an easy task. Hence, accurately predicting the wind power is of great importance to cope with the impacts of wind power fluctuation on power system operation. Furthermore, it is helpful for the management of power grid production, delivery, and storage, as well as for decision-making on the energy market. Accurate prediction of wind power plays a core role in sustainably integrating the wind power a smart grid. The need for short-term predictions is increased with the increasing installed capacity. Thus, several prediction methodologies have been designed in the literature and can generally be grouped into two classes that are physical and statistical-based methods [2]. Physical models can predict the long-term trend of the wind process, but their local precision is low. While statistical models employ data mining algorithms to train a model expressing the relation between wind power and other variables such as input and output. Generally speaking, the essence of physical models is the application of atmospheric motion equations to calculate future meteorological measurements, then forecasting wind power by using some predicted meteorological variables such as wind speed [3]. The typical physical model is generally performed using numerical weather estimation which is implemented into two steps: prediction of wind speed, then transformation of this wind speed to wind power [4]. This mechanism is usually accomplished by using wind power curve which could be modeled using parametric or non-parametric methods [5]. Although physical models provide acceptable features in predicting the long-term trend of wind variance, they are costly to develop and time-consuming and provide low prediction accuracy for a local area. For more details refer to [5].

Statistical prediction approaches build a model linking the inputs and outputs variables based on training datasets and then use the designed model to predict values of the output for new input data [6]. For instance,
in [7] proposed a method based on the least squares support vector machine for short-term prediction of wind power. The parameters of this model are optimally tuned using the gravitational search algorithm to further enhance the prediction accuracy. In [8], a hybrid approach merging the desirable characteristics of autoregressive integrated moving average and Kalman filter is introduced to predict the daily means wind speed. The parameters of this approach are optimized using the particle swarm optimization for enhancing the prediction performance. In [9], an approach merging Gaussian process regression and multiple imputations is proposed to predict wind power with missing data. To do so, the expectation-maximization algorithm is employed to estimate mixture components of the underlying data distribution and to handle missing data. In [10], a two steps hybrid modeling is introduced for enabling an enhanced prediction of wind power. After getting the trend of wind power using wind power curve, the deviations of this model can be adjusted using data-driven schemes. However, combining models may increase the time perform prediction of wind power [10].

Machine learning is a remarkable multidisciplinary field, where methods could be implemented for wind power prediction. Several researches focused on applying artificial intelligence techniques for wind power prediction [11, 12, 13]. In [11], the k-nearest neighbor classifier (kNN) is used for predicting wind power on a short-term horizon using multi-tuples meteorological input measurements. Using genetic programming based ensemble of neural networks, authors in [13] proposed a robust approach short term prediction of wind power. In [14], a method integrating a data mining technique and enhanced support vector machine algorithm is introduced for a short-term of wind power. This method used data mining to explore the correlation between wind speed and wind powers output and then modify the invalid original data. However, this method is not verified for long-term prediction [14]. A combined approach using the wavelet transform and support vector machine is introduced in by Liu et al. [15]. In [16], a hybrid intelligent algorithm is introduced to directly modeling prediction intervals by using extreme learning machines and self-adaptive evolutionary extreme learning machines. In [17], an approach for probability density forecasting of wind power based on quantile regression neural network and kernel density estimation is introduced. It has been shown that this approach can uncover the entire conditional distribution of out variables, and handle with complex nonlinear problems. The effectiveness of machine learning approaches relies on the quality of data available.

The efficiency and precision of the methods used in predicting wind power are necessary for predicting the performance of wind turbines. The overarching goal of this study is to design an approach enabling an efficient prediction of wind power production based on times series supervisory control and data acquisition (SCADA) data from a wind turbine. In last years, ensembles models, which combine numerous single models to an ensemble modeling approach, present an effective alternative to the conventional machine learning approaches [18, 19]. The capacity of ensemble modeling approaches to decrease the model’s variance while obtaining a low bias makes them very attractive to improve further prediction accuracy [20]. Here, the
bagging ensembles of decision trees approach is adopted for enabling an efficient wind power prediction [20].

The choice of this regression approach is motivated by its ability to take advantage of many relatively weak single trees to reach a high prediction performance compared to single regressors. Moreover, it reduces the overall error and has the capacity to merge numerous models. Real measurements recorded every ten minutes from an actual wind turbine are used to illustrate the prediction quality of the studied methods. The performance of the bagged trees (BTs) prediction approach has been compared with four prediction models namely ordinary least squares, support vector machine regression (SVR), latent variables regression methods, i.e., partial least squares (PLS) and principal components regression (PCR). To the best of our knowledge, the LVR models have not been exploited for wind power prediction. Multivariate statistical projection methods such as PCR and PLS are commonly utilized to handle a high number of highly correlated process variables by conducting regression on a smaller number of transformed variables (i.e., latent or principal component), which are linear combinations of the raw measurements. After computing the latent variables in the process being investigated, then these fewer number of variables are used instead of using the raw data. These latent variables regression (LVR) approaches, generally result in well-conditioned parameter estimates and reliable model predictions. Another contribution in this paper is bringing this LVR method to the attention of the renewable energy community and show how it can be applied in a new field. Results showed the superior performance of the bagging trees approach compared to the four investigated prediction approaches.

This paper is structured as follows. The wind power data used in this study is described in Section 2. Then, the prediction models including SVR, PCR, PLSR, and BTs are briefly reviewed in Section 3. In Section 4, the experimental results are provided. Lastly, conclusions are offered in Section 5.

## 2. Wind turbine datasets and study site

### 2.1. Wind turbine description

Generally speaking, wind turbine transforms kinetic energy (wind speed) to electric energy (active power) via a generator. In other words, it permits transmitting the mechanical energy produced by the rotor in the form of a rotational force. In modern turbines, the shaft linked to a gearbox, which consists in increasing the rotation speed (e.g., from perhaps 20 to 1000 rpm or 1500 rpm (50 Hz) or 1200 rpm or 1800 rpm (60 Hz)), is appropriate to drive a synchronized generator. Figure 2 illustrates a typical drivetrain with a gearbox.

The wind power measurements used in this study are gathered from a high-wind-speed wind turbine Senvion MM82 located in France. The main characteristics of this wind turbine are summarized in Table 1. The MM82 wind turbine comprises a doubly-fed induction generator, frequency converter, and DC pitch system and has demonstrated its dependability for years. The central core of the wind turbines is the integrated drive train concept, where all the different components: rotor blade with 82m of diameter, 3
blades, planetary/spur gear, generator with a speed max of 1800 U/min, and cut-in wind speed from 3.5m/s to 25m/s of cut-out considering 14.5m of rated wind speed, converter and transformer, work in perfect harmony. The desirable characteristics of this wind turbine are its capacity to produce a high energy yield and also maintains noise emission low. The hub heights of this wind turbine of 80 meters make it a choice for sites with height restrictions (Table 1). The sampling frequency of the collected dataset is five minutes per point. The recorded dataframe contains twelve input variables and the active power as the response variable. In the model development stage, along with wind speed, torque, generator converter speed, generator speed, converter torque, rotor speed, pitch angle, gearbox oil sump temperature, absolute wind direction corrected, rotor bearing temperature, generator bearing 1 temperature, and generator bearing 2 temperature are included in the input matrix.

### Table 1: Wind turbine main characteristics.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Senvion MM82</strong></td>
<td><strong>Senvion MM82</strong></td>
</tr>
<tr>
<td>Rated power</td>
<td>2,050.0 kW</td>
</tr>
<tr>
<td>Rotor diameter</td>
<td>82.0 m</td>
</tr>
<tr>
<td>Hub height</td>
<td>80 m</td>
</tr>
<tr>
<td>Number of blades</td>
<td>3</td>
</tr>
<tr>
<td>Rotor speed, max</td>
<td>7.1 U/min</td>
</tr>
<tr>
<td>Generator Speed, max</td>
<td>1,800.0 U/min</td>
</tr>
<tr>
<td>Cut-in wind speed</td>
<td>3.5 m/s</td>
</tr>
<tr>
<td>Rated wind speed</td>
<td>14.5 m/s</td>
</tr>
<tr>
<td>Cut-out wind speed</td>
<td>25.0 m/s</td>
</tr>
</tbody>
</table>

3. Methods

This section describes the prediction methodologies used for predicting wind power. Four prediction methods have been investigated in this study for wind power prediction: multivariate linear regression, support vector regression, principal component regression and partial least squares.
3.1. Ordinary least squares regression

Consider an input with \( n \) samples and \( m \) variables \( \mathbf{X} \in \mathbb{R}^{n \times m} \) and an output vector with \( n \) samples \( \mathbf{y} \in \mathbb{R}^{n \times 1} \) (i.e., the active power). The linear regression model linking the input and the output variables is

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} =
\begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1m} \\
x_{21} & x_{22} & \cdots & x_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_m
\end{bmatrix} +
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{bmatrix},
\]

(1)

It can expressed in the compact form as,

\[
y = \mathbf{X}\beta + \epsilon.
\]

(2)

Ordinary least square (OLS) regression is one of the most commonly used estimation approaches, in which the model parameters are obtained by the minimization of the following cost function \([21, 22]\),

\[
\min_{\beta} \left( \|\mathbf{X}\beta - \mathbf{y}\|^2_2 \right),
\]

(3)

The estimated \( \hat{\beta} \) coefficient vector of the least-squares is given by:

\[
\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.
\]

(4)

3.2. Principal Component Regression (PCR)

To predict \( \mathbf{y} \) from \( \mathbf{X} \), the PCR method is performed into two steps: first, the input data matrix \( \mathbf{X} \) is decomposed using principal component analysis (PCA), and then linear model linking the retained principal component and the output variable, \( \mathbf{y} \), is established \([21, 22]\). Here, \( \mathbf{y} \) is the active power and \( \mathbf{X} \) is the input (predictor) variables, such as wind speed, pitch angle, and absolute wind direction corrected. So, based on the PCA model, the input data matrix \( \mathbf{X} \) can be expressed as a sum of the approximated matrix, \( \hat{\mathbf{X}} \), and residual data, \( \mathbf{E} \).

\[
\mathbf{X} = \mathbf{T}\mathbf{W}^T = \sum_{i=1}^{k} t_i w_i^T + \sum_{i=k+1}^{m} t_i w_i^T = \hat{\mathbf{X}} + \mathbf{E}
\]

(5)

where \( \mathbf{T} \in \mathbb{R}^{n \times m} \) represent a matrix of the principal components (PCs) and \( \mathbf{W} \in \mathbb{R}^{m \times m} \) is the loading matrix. In the presence of cross-correlated multivariate data, \( \mathbf{X} \), the first \( 'k' \) PCs (where \( k < m \)) are sufficient for preserving relevant information in the original data. Indeed, the PCs, \( \mathbf{T} \), which represents a linear combination of the original variables (\( \mathbf{X} \)), are used to represent the variance in the data as efficiently as possible and they don’t have any physical meaning. The loading matrix, \( \mathbf{W} \), is frequently calculated through Singular Value Decomposition of the covariance matrix of the data \( \mathbf{X} \) \([23, 24]\). Here cumulative
percentage variance (CPV) procedure is utilized to select the number of PCs to retain in the model, due to its simplicity and accuracy [25]. Let $\hat{T} \in \mathbb{R}^{n \times l}$ be the matrix of the $k$ retained principal components. In the PCR approach, the linear regression between $\hat{T}$ and the response variable $y$ is obtained by solving the following quadratic optimization problem.

$$\hat{\beta} = \arg \min_{\beta} \left( \| \hat{T}\beta - y \|_2^2 \right)$$  

(6)

The least squares solution is expressed as:

$$\hat{\beta} = (\hat{T}^T \hat{T})^{-1} \hat{T}^T y.$$  

(7)

Note that the PCR model when using all PCs becomes equivalent to OLS.

3.2.1. Partial Least Square (PLS)

Consider an input with $n$ samples and $m$ variables $X \in \mathbb{R}^{n \times m}$ and output with $n$ samples and $p$ variables $Y \in \mathbb{R}^{n \times p}$. PLS extracts the principal components iteratively by maximizing the covariance of the extracted principal components. PLS model development has two components, one is to develop inner models and the other one is to develop outer models [26, 27, 28]. Outer models have a relationship with the inner model as:

$$\begin{cases} 
T = \sum_{i=1}^{l} t p_i^T = TP^T + G \\
U = \sum_{i=1}^{q} u q_i^T = UQ^T + F,
\end{cases}$$  

(8)

where, $T \in \mathbb{R}^{n \times l}$ and $U \in \mathbb{R}^{n \times q}$ represent a matrix of the transformed uncorrelated variables. The loading matrices of input and output space are $P \in \mathbb{R}^{m \times l}$ and $Q \in \mathbb{R}^{p \times q}$, respectively. The model residuals are the $G$ and $F$. The essence of PLS is to factorize $X$ and $Y$ using PCA and then performing regression between the PCs of $X$ and $Y$ (i.e., $T$ and $U$, respectively). As shown above, the PCs in PCR model are computed in such a way that they only describe the variability in $X$, while in PLS the PCs are determined so that they describe $Y$ and contains a maximum correlation among $X$ and $Y$. The PCs in PLS are just mathematical constructs to explain the variance in the data, and lack of physical interpretation. The number of principal components ($l$) is determined by a well-known method called cross-validation [28]. The retained latent variables of the input and output space i.e., $T$ and $U$, are related by the linear model as:

$$U = T \beta + E,$$  

(9)

where $\beta$ is a regression matrix, and $E$ represents a residual matrix. The response $Y$ can now be expressed as:

$$Y = T \beta Q^T + F^*,$$  

(10)
PLS model is constructed in a sequential manner by iteratively adding the estimated principal component to the model [29, 30]. The Nonlinear Iterative Partial Least Squares (NIPALS) is the most commonly used algorithm to solve PLS problem [31, 32].

3.3. Support vector regression (SVR) approach

This section describes the general concept of SVR modeling and for more details refer to [33, 34]. The support vector machine was originally designed by Vapnik [33], and has been extensively employed in classification [35] and regression [36, 37]. It has been shown that SVR has good properties in learning limited samples [34]. Also, it is an approach to uncover nonlinear features in multivariate input-output data by using kernel trick. Generally speaking, in $\epsilon$-SVR, the aim is to come up with a function $f(x)$ that could deviate from the current target $y_i$ at most by $\epsilon$ for the entire training dataset, and it is as flat as possible [34]. Here, the errors remain acceptable since they are less than $\epsilon$, but any deviation larger than $\epsilon$ is not acceptable. Indeed, in ordinary least squares regression, the aim is to minimize the error, whereas in SVR modeling the goal is to fit the error within a certain threshold.

Basically, the SVR model projects the training data $(x_i, y_i), i = 1, 2, \ldots, m$, where $x_i \in \mathbb{R}^n$ is the input vector and $y_i$ is the target value into a high dimensional feature space based on a nonlinear mapping $\phi$ and then apply a linear regression for estimating the following function in this feature space.

$$f(X) = w \cdot \phi(x) + b,$$  \hspace{1cm} (11)

The weight vector $w$ and the bias $b$ can be obtained by solving the following minimization of convex quadratic programming problem [34].

$$\min_{w, \xi, \xi^*} \left\{ \frac{1}{2} w^T w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \right\}$$  \hspace{1cm} (12)

under constraints

$$\begin{cases} y_i - (w^T \phi(x_i) - b) \leq \epsilon + \xi_i, \\ w^T \phi(x_i) + b - y_i \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, i = 1, \ldots, l. \end{cases}$$

where the constant $C > 0$ defines the trade-off between the flatness of $f$ and the margin of tolerance (i.e., the maximum deviations larger than $\epsilon$ to be tolerated. $\xi$ and $\xi^*$ are the slack variables that measure the distance of the point to its marginal hyperplane. They are used to quantify the amount of error. Indeed, only the points which have the least error rate (within the margin of tolerance) are taken in order to get a better fitting model.
The dual of Equation (12) is obtained as [34],

$$\begin{align*}
\text{max} \quad & \frac{1}{2} \sum_{i,j=1}^{m} (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(x_i, x_j) \\
& - \epsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i (\alpha_i - \alpha_i^*) \\
\text{s.t.} \quad & \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0, \ 0 \leq \alpha_i, \ \alpha_i^* \leq C\text{.}
\end{align*}$$

Where $\alpha_i$ and $\alpha_i^*$ are Lagrange multipliers [34].

Finally, the equation (11) can be expressed using the kernel function as

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) k(x_i, x_j) + b, \tag{13}$$

where $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel, which is a function employed for mapping a lower-dimensional data into a higher-dimensional space. Here training vectors $x_i$ are mapped into a higher (maybe infinite) dimensional space by the function $\phi$. In practice, the kernel function has a core role in the performance of the SVM classifier. As the classification performance of the SVM algorithm depends on the selected kernel function, in this study, SVM with three frequently used kernel functions (linear, Radial Basis Function (RBF), and polynomial kernels) are compared.

- Linear kernel:
  $$K(x_i, x_j) = x_i . x_j,$$

- Polynomial kernel:
  $$K(x_i, x_j) = (x_i . x_j + 1)^d,$$

  where $d$ is the degree of polynomial kernel.

- RBF kernel:
  $$K(x_i, x_j) = \exp\left(\frac{||x_i - x_j||^2}{2\sigma^2}\right),$$

  where $\sigma$ is the width of Gaussian kernel.

3.4. Bagged regression trees

The essence of bootstrap aggregating (bagging) trees, which has been primarily designed by Breiman [20], is to construct multiple similar independent predictors and average the outputs of these predictors to get the final prediction. This allows for reducing the variance error [38]. In bagging trees or bagging ensembles of decision trees approach, several trees (individual models) are combined jointly for enhancing the prediction quality of the model (Figure 3). Indeed, the BTs predictive model is employed for reducing the variance of regression trees and remedy the overfitting problem in the single tree. Figure 3 illustrates the basic principle behind BTs predictive model.
The first step in BTs is to create \( N \) new training datasets of the same size \( n \) as the original data by the selection of \( n \) out of \( n \) samples uniformly with replacement from the original training dataset. Then, train each tree in the ensemble individually on the corresponding training new sets. For instance, in this study 30 trees are used in the bagging trees models. Lastly, compute the average of all predictions to get a final prediction. The prediction of the bagging trees model is expressed as:

\[
\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(X),
\]

(14)

where each tree model \( f_i \) is trained on bootstrap data \( i \). The main steps utilized for computing the bagging trees prediction are outlined in Algorithm 1.

It can be theoretically shown that the variance of prediction can be reduced to \( 1/n \) (\( n \) is the number of learners) of the original variance (single learner). Thus, using a large number of learners results on reduced or similar variance compared to prediction performance when using lower numbers of learners. To understand the core idea of how bagging could reduce the mean squared prediction error, consider a regression problem with base regressors \( b_1(x), \ldots, b_n(x) \). Suppose that there is an ideal target function of true answers \( y(x) \) resulting from a given set of inputs and that the distribution \( p(x) \) is defined. Then, the error for each
The regression function can be computed as

$$
\varepsilon_i(x) = b_i(x) - y(x), \quad i = 1, \ldots, n
$$

(15)

The mean squared error can be expressed as:

$$
\mathbb{E}_x [(b_i(x) - y(x))^2] = \mathbb{E}_x [\varepsilon_i^2(x)].
$$

(16)

Then, the mean error over all the regression functions can be computed as

$$
\mathbb{E}_1 = \frac{1}{n} \mathbb{E}_x [\varepsilon_i^2(x)].
$$

(17)

Let us assume that the errors are unbiased and uncorrelated, that is

$$
\begin{align*}
\mathbb{E}_x [\varepsilon_i(x)] &= 0, \\
\mathbb{E}_x [\varepsilon_i(x)\varepsilon_j(x)] &= 0, \quad i \neq j.
\end{align*}
$$

(18)

The regression function computed by averaging the individual functions is given as

$$
a(x) = \frac{1}{n} \sum_{i=1}^{n} b_i(x).
$$

(19)
Then, its mean squared error is computed as,

\[
E_n = \mathbb{E}_x \left[ \frac{1}{n} \sum_{i=1}^{n} b_i(x) - y(x) \right]^2
= \mathbb{E}_x \left[ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \right]^2
= \frac{1}{n^2} \mathbb{E}_x \left[ \sum_{i=1}^{n} \varepsilon_i^2(x) + \sum_{i \neq j} \varepsilon_i(x) \varepsilon_j(x) \right]
= \frac{1}{n} \mathbb{E}_1. \tag{20}
\]

Thus, by averaging the individual answers, we reduced the mean squared error by a factor of \( n \). Overall, the desirable characteristic of bagging is its capacity for reducing the variance of prediction error by training the model on different datasets. The advantage of bagging resides in the fact that individual models trained based on different training data and their errors are reduced during the averaging to get the final prediction output. In addition, in some of the training datasets, outliers observations are likely omitted. When combining several trees in the BTs model, it could be challenging to interpret the obtained model. Of course, BTs enhances the prediction performance at the expense of interpretability.

4. Experimental results and discussion

The wind power prediction based on regression models (i.e., OLS, PCR, PLSR, SVR, and BTs) is performed into three stages as summarized in Figure 4. In general, the aim of prediction is first to explore the correlation between the input process variables and the output variable (or response) in the training phase, then, provide us the predicted values of the output for the given unseen input data. More specifically, first, the predictive model (e.g., BTs, SVR and LVR models) is constructed using the training input-output data, \((X, y)\). Second, for testing unseen input data, the build model is used to predict the input variable (wind power). Finally, the performance of the prediction is checked. R-squared \((R^2)\), mean absolute error (MAE) and root mean squared error (RMSE) are the most commonly used metrics to check the quality of prediction. These were calculated as follows:

\[
MAE = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|,
\tag{21}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2},
\tag{22}
\]

\[
R^2 = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \text{mean}(Y))^2},
\tag{23}
\]

where \( y_t \) are the measured values, \( \hat{y}_t \) are the corresponding predicted values by the regression model and \( n \) is the number of observations.
Here, the k-fold cross-validation (CV) technique, which is frequently performed to give a suitable estimation of a model’s prediction error, is used to build the prediction models [39]. The training dataset is divided into k portions (folds) where every portion is utilized as a testing data at some point (Figure 5). This permits the construction of a reliable prediction model. Figure 5 illustrates an example of a K-fold cross-validation with $K = 5$. The available dataset is randomly divided into $K$ equal subsets. As illustrated in Figure 5, in the first iteration, the first fold is used for testing and the rest for training. In the second iteration, the second fold is utilized to test the model and the rest folds are used for training. This procedure is repeated until ensuring that all folds are utilized as testing dataset. The mean square error is computed for each testing sub-datasets, $MSE_i$. The cross-validation error is calculated as the mean of the prediction errors as:

$$CV_{error} = \frac{1}{k} \sum_{i=1}^{k} MSE_i.$$  

(24)

In practice, cross-validation with $K = 5$ or $K = 10$ is typically used in model construction [39]. Indeed, selecting the number of folds, $k$, in cross-validation depends on the size of the data. For instance in 10-fold cross-validation, only 10% of the training dataset is used for validation. It should be guaranteed that the training and validation datasets are from the same distribution. In other words, both training and validation datasets comprise sufficient variability so that the underlining distribution is described.
4.1. Preliminary data analysis

Generally speaking, wind turbines are designed for operating within an interval of wind speeds and with maximum power. This power value is the nominal value of the generator, but it is rarely achieved. The operation of the wind turbine depends on the wind speed. This wind speed will vary constantly as shown in Figure 6(a), which represents the evolution of wind speed and power production from the 2,05 MW Senvion MM82 wind turbine. The wind power curve, which represents the process from wind speed to wind power (Figure 6(a)), is usually used to detect abnormalities in wind turbines by comparing the empirical wind power curve with the theoretical one provided by the constructor. The red full circle in Figure 6(a) represented observations corresponding to nil power when where wind speed is greater than the cut-in speed. These observations require more investigation for understanding if there was any unplanned downtime. Figure 6(b) illustrates the plot of rotor speed against wind speed. We can see an increase in the rotor speed in accordance with wind speed after surpassing the cut-in wind speed. Another important curve to characterize the operating condition of a wind turbine is the pitch curve which plot winds speed against pitch angle (Figure 6(c)). Similarly to the power curve analysis, deviations in characteristic operation in the pitch curve can be identified and further analysis for anomaly detection can be performed. For instance, the generated power is below the desired performance as wind increases because of faults in the adjustment system. Essentially, to bypass this problem, variable speed turbines use active pitch control to adjust the blade pitch angle to augment the catch of power given wind speed. Figure 6(c) displays a wind speed variation as a function of blade pitch angle degree in under normal conditions.

Figure 6: (a) Active power, (b) rotor speed, and (c) blade pitch angle vs the wind speed for the 2,05 MW Senvion MM82 wind turbine obtained from SCADA data.

Figure 7 displays the produced power as a function of rotor torque speed. Generally, the rotor torque rises when the wind speed rises after reaching the cut-in wind speed. Accordingly, wind power increases with the increase of the torque until reaching the maximum power (Figure 7). Thus, any change impacting the ratio between rotor torque and the wind power produced over time should be investigated to uncover possible abnormalities in the wind turbine. It is essential to augment or decrease the rotor speed system such that more or less torque is used for obtaining a similar amount of power should be produced.
4.2. Models design

The training dataset from September 1st, 2013 to May 14, 2014, are employed to build the prediction models. Instead of displaying all observations of a dataset or only mean with standard deviation, Table 2 presents summary values of the training dataset: the minimum, the first quartile, the median, the third quartile, the maximum, kurtosis, and skewness. Skewness and kurtosis are calculated to show the symmetry and shape of the studied time series distributions. From Table 2, we can conclude that several variables were negatively skewed with relatively large kurtosis. The training set is normalized then used to construct regression models.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>Q.25</th>
<th>Q.5</th>
<th>Q.75</th>
<th>max</th>
<th>skewness</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind speed (m/s)</td>
<td>5.697</td>
<td>2.512</td>
<td>0</td>
<td>4.29</td>
<td>5.70</td>
<td>7.12</td>
<td>16.48</td>
<td>0.27</td>
<td>3.88</td>
</tr>
<tr>
<td>Torque (Nm)</td>
<td>2206.75</td>
<td>2310.69</td>
<td>-734.41</td>
<td>350.24</td>
<td>1625.31</td>
<td>3223.94</td>
<td>10875.70</td>
<td>1.37</td>
<td>4.72</td>
</tr>
<tr>
<td>Generator converter speed (rpm)</td>
<td>1137.29</td>
<td>596.29</td>
<td>1.86</td>
<td>971.80</td>
<td>1257.59</td>
<td>1640.26</td>
<td>1804.78</td>
<td>-0.81</td>
<td>2.43</td>
</tr>
<tr>
<td>Generator speed (rpm)</td>
<td>1135.81</td>
<td>596.57</td>
<td>-0.07</td>
<td>970.07</td>
<td>1256.11</td>
<td>1659.08</td>
<td>1803.15</td>
<td>-0.81</td>
<td>2.43</td>
</tr>
<tr>
<td>Converter torque (Nm)</td>
<td>2342.14</td>
<td>2265.35</td>
<td>-9.32</td>
<td>641.63</td>
<td>1838.41</td>
<td>3288.11</td>
<td>10915.70</td>
<td>1.36</td>
<td>4.82</td>
</tr>
<tr>
<td>Rotor speed (rpm)</td>
<td>10.83</td>
<td>5.70</td>
<td>0</td>
<td>9.23</td>
<td>11.96</td>
<td>15.64</td>
<td>17.21</td>
<td>-0.81</td>
<td>2.42</td>
</tr>
<tr>
<td>Pitch angle (deg)</td>
<td>11.01</td>
<td>25.23</td>
<td>-37.06</td>
<td>-0.98</td>
<td>-0.95</td>
<td>0.11</td>
<td>92.68</td>
<td>2.10</td>
<td>6.38</td>
</tr>
<tr>
<td>Gearbox oil sump temperature (°C)</td>
<td>52.95</td>
<td>8.84</td>
<td>11.47</td>
<td>52.08</td>
<td>56.14</td>
<td>57.89</td>
<td>63.16</td>
<td>-2.53</td>
<td>9.69</td>
</tr>
<tr>
<td>Absolute wind direction corrected (deg)</td>
<td>172.43</td>
<td>100.22</td>
<td>0.01</td>
<td>66.39</td>
<td>190.51</td>
<td>249.15</td>
<td>359.99</td>
<td>-0.11</td>
<td>1.84</td>
</tr>
<tr>
<td>Rotor bearing temperature (°C)</td>
<td>27.13</td>
<td>6.18</td>
<td>5.67</td>
<td>23.53</td>
<td>27.94</td>
<td>31.30</td>
<td>41.51</td>
<td>-0.57</td>
<td>3.25</td>
</tr>
<tr>
<td>Generator bearing 1 temperature (°C)</td>
<td>39.70</td>
<td>7.24</td>
<td>6.09</td>
<td>37.61</td>
<td>41.40</td>
<td>44.28</td>
<td>61.56</td>
<td>-1.78</td>
<td>6.87</td>
</tr>
<tr>
<td>Generator bearing 2 temperature (°C)</td>
<td>39.17</td>
<td>7.36</td>
<td>4.28</td>
<td>37.23</td>
<td>40.73</td>
<td>43.66</td>
<td>60.52</td>
<td>-1.90</td>
<td>7.30</td>
</tr>
<tr>
<td>Active power (kW)</td>
<td>564.01</td>
<td>626.06</td>
<td>-12.38</td>
<td>79.15</td>
<td>320.11</td>
<td>873.02</td>
<td>2266.90</td>
<td>1.20</td>
<td>3.34</td>
</tr>
</tbody>
</table>

The training measurements are utilized to design prediction methods. As discussed above in this study five prediction methods are investigated and compared for predicting wind power of an actual wind turbine.
In this paper, the 5-fold cross-validation technique is used to find the model parameters. Here, the SVR procedure is used for wind power prediction due to its flexibility to deal with linear and nonlinear data by using nonlinear kernels. There are numerous kernel functions that can be used in the SVR prediction approach. In this study, we tested three commonly used kernels: linear, polynomial, and Gaussian function. It should be noted that one important step in PCR and PLSR models development is to select the number of PCs. For this purpose, the cumulative percentage variance (CPV) procedure is utilized due to its simplicity and accuracy. Figure 8(a-b) shows respectively the CPV explained in the input data X for both PLSR and PCR and the CPV explained in the output for the PLSR model. From Figure 8(a-b), it can be seen that two PCs are sufficient to describe around 99% of the variability in X in the two models. As the prediction purpose, it to predict the response, Figure 8(b) indicates that three components in PLSR are able to describe 94.8% of the variability in y. Thus, two PCs are used in designing the PCR model while three PCs are used for the PLSR model.

Figure 8: Selection of the number of PCs in PLSR and PCR models. (a) Percent variance accounted in X, and (b) Percent variance accounted in y (PLSR model) vs the number of PCs.

Figure 9(a-g) shown the prediction results of the OLS, PCR, PLSR, SVR and BTs models using training dataset. These results illustrate the goodness of bagging trees predictive model for the prediction of wind power.

Here, the prediction quality of the six models is evaluated in terms of their RMSE, MAE and $R^2$. Qualitatively, Table 3 summarizes the prediction accuracy of the six models based on the training dataset using 5-fold and 10-fold cross-validation. This permits to assess the models’ robustness, to exploit the whole datasets for training and for validation and helps avoid overfitting. In the BTs approach, we started with 20 trees and leaf size of 8, and we iteratively tested with different number of trees (learners) and found that BTs with 30 trees produced high accuracy. The results in Table 3 indicate that the prediction performances of OLS, linear SVR, PCR, and PLSR are relatively comparable. Indeed, these models are able to extract only linear features in the data. The results in Table 3 also show that SVR with quadratic kernel provides a slight advantage over OLS, linear SVR, PCR, and PLSR by reaching an $R^2$ of 0.954. The use of a
quadratic kernel in the SVM approach allows extracting nonlinear features in data. The results indicate
that the SVM model with Gaussian kernel achieved improved prediction accuracy with an $R^2$ of 0.971. This
result confirms that the SVM model with Gaussian kernel is appropriate in predicting wind as compared to
the OLS, PCR, and PLSR models. After testing the three kernel functions with SVM, the best prediction
performance is realized when using the Gaussian function, which achieved an $R^2$ of 0.98 (Table 3). The BT
model achieved the smallest values of MAE and RMSE, which indicates that this model is more appropriate
for wind prediction compared to the other studied models. As the bagging decreases the variance and the
mean squared errors [40, 20], it can be seen that both the RMSE and the MAE decreased significantly in
the bagged model compared to the traditional models. Also, the improvement is clearly noticeable when
using the evaluation metrics RMSE and MAE compared to $R^2$. This is mainly because $R^2$ is conveniently
scaled between 0 and 1, whereas RMSE and MAE are not scaled to any particular values. Also, when using
the bagged model, the MSE is decreased which directly reflected by RMSE and MAE. Moreover, $R^2$ can
be viewed as normalized MSE by the total sum of squares error, thus, the improvement is less significant
compared to RMSE and MAE. Furthermore, results in Table 3 indicate that the bagged trees model archives
almost comparable result when using 10-fold cross-validation and 5-fold cross-validation. Here the 5 fold
cross-validation is chosen in this study because using more folds increase the computational time, which is
not advantageous.

4.3. Prediction results

Now, the constructed prediction models will be tested using new test dataset. The test dataset from May 15 to May 22 2014 is employed to test the quality of the prediction models. This dataset is collected every 5 minutes. Figure 10(a-g) illustrates the measured and predicted wind power using OLS, SVR models, PLSR, PCR, and the bagging decision tree model.

Also, to show clearly the accordance of the measured and predicted wind power from the explored models, the scatter plots are displayed in Figure 11. It can be seen that from Figures 10 and 11 that the bagging trees predict well the wind power data.

The accuracy metrics of the six models previously designed using 5 and 10 folds in cross-validation, when applied to unseen testing datasets, are summarized in Table 4. In this study, linear models including

<table>
<thead>
<tr>
<th>Model</th>
<th>5-Fold Cross Validation</th>
<th>10-Fold Cross Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R^2</td>
<td>RMSE</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.950</td>
<td>144.374</td>
</tr>
<tr>
<td>Linear SVR</td>
<td>0.946</td>
<td>150.545</td>
</tr>
<tr>
<td>Quadratic SVR</td>
<td>0.971</td>
<td>110.650</td>
</tr>
<tr>
<td>Gaussian SVR</td>
<td>0.980</td>
<td>91.818</td>
</tr>
<tr>
<td>PCR</td>
<td>0.949</td>
<td>146.102</td>
</tr>
<tr>
<td>PLSR</td>
<td>0.949</td>
<td>146.007</td>
</tr>
<tr>
<td>Bagged Trees</td>
<td>0.980</td>
<td>79.909</td>
</tr>
</tbody>
</table>

Table 3: Summary of models quality when using 5 and 10 folds cross-validation.
OLS, linear SVR, PLSR, and PCR reached the lowest accuracy of around 0.93. These models are not able to capture nonlinear features in the data and result in low prediction accuracy compared to the nonlinear models (Quadratic and Gaussian SVR, and BTs). The main benefit of SVRs consists in their capacity to suitably uncover important patterns in the data by transforming problems into higher dimensions using kernel functions, enabling a non-linear relationship to appear approximately linear. The results from Table 4 demonstrated that the BTs prediction approach clearly outperformed OLS, linear SVR, PLSR, and PCR methods. Indeed, OLS, linear SVR, PLSR models omitted the presence of nonlinear features in modeling wind power data. In other words, they were not designed to model linear process. Accordingly, important features and pattern in the data were lost when these predictors are used. Also, it can be highlighted that the BTs exhibited superior prediction performance compared to SVR with Gaussian kernel, which is a linear predictor (Table 4). Indeed, BTs outperformed the other methods (Table 4) by achieving a $R^2$ up to 0.98, which means that the model is able to explain 98% of the total variation in the testing data about the average and 2% is in the residuals. Furthermore, results in Table 4 indicate a significant improvement of the prediction performance of the bagged model over our individual models is reached in term of MAE and RMSE. This flags the ability of the bagged for appropriately predicting wind power production. This fact was attributed to the flexibility of BTs as an ensemble learning model and its ability to reduce the variance of prediction error. Also, the results in Table 4 show that results obtained with models constructed using 5 and 10 folds in cross-validation are almost comparable, which suggests that 5 folds could be selected.

As discussed above, bagging trees model as an ensemble employs multiple predictors (decision trees) and combine their output values to get the final prediction, which permits improving the prediction accuracy.
and reducing the required computation time. Also, ensemble approaches need less tuning than conventional individual machine learning models.

5. Conclusion

A reliable prediction of wind power production may be a tool for facilitating wind turbine integration in the smart grid. In this paper, we exploited the desirable characteristics of bagged trees approach to enabling an improved prediction of wind power. As expected, using bagged trees as an ensemble model helps in reducing the variance error and enhance the prediction quality. Furthermore, we provided a comparison of the proposed model with four prediction models namely ordinary least square regression, support vector regression, principal components regression and partial least squares regression. This presented a comparative study for predicting wind power has been performed using real SCADA dataset. The results revealed that the bagged trees model achieved superior prediction quality. The designed prediction models when using fault-free data mimics the nominal behavior of wind turbine can be very useful in designing monitoring schemes to prevent faults before they occur and thus save money and unnecessary downtime. As future work, it would be interesting to develop a statistical approach by combining these prediction models with statistical quality control techniques to supervise wind turbine performances.

Acknowledgement

This publication is based upon work supported by King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR) under Award No: OSR-2019-CRG7-3800.

References


