Correlation analysis of factors affecting wind power based on machine learning and Shapley value

Chuanjun Pang \(^1\)\(^2\) | Jianming Yu \(^1\)\(^2\) | Yan Liu \(^1\)\(^2\)

\(^1\)NARI Group Corporation Co., Ltd (State Grid Electric Power Research Institute Co., Ltd), Nanjing, China
\(^2\)Beijing KeDong Electric Power Control System Co., Ltd, Beijing, China

Abstract
An analysis of the impact of various factors on wind power can help grid dispatchers understand the characteristics of wind power output and improve the accuracy of wind power forecasting. A correlation analysis method of factors affecting wind power is proposed based on machine learning and the Shapley value. First, factors affecting wind power and the method of constructing wind power models based on machine learning are introduced. Then, to measure the influence of factors on wind power, the Shapley value is proposed based on the wind power model. In addition, calculation methods, properties, and application scenarios of the Shapley value are introduced. Finally, based on the actual data of a wind farm, the method is used to analyse environmental factors affecting wind power, and the main factors affecting the wind farm are determined. The experimental results show that the method can identify important factors affecting wind power and measure the complex non-linear relation between each environmental factor and wind power.

1 | INTRODUCTION
With the continuous expansion of the capacity of wind power, the intermittence and randomness of wind power caused by the characteristics of wind resources have an increasingly obvious impact on the power grid. For example, they pose many operational and control challenges that hamper the reliable and stable operation of power grids [1]. Wind power generation often faces difficulties regarding reliability in terms of the generation, planning, and scheduling of the supply of electricity [2, 3]. In addition, the uncertainty of power load further increases the difficulty of power grid dispatch and operation [4]. There are two measures to solve the impact of wind power and improve the system’s ability to absorb wind power. First, increase the reserve capacity of the power system to improve the flexibility of the power system [5]; and second, improve the accuracy of wind power forecasting. In the former, an ample and flexible reserve capacity is bound to be detrimental to the economic operation of power grids [6]. Wind power forecasting can help power grid dispatchers to grasp the characteristics of wind power and provide a basis for adjusting the power grid dispatching schedule [7, 8], improving the economic operation level of the power grid. Wind power forecast methods are mainly divided into two categories [9–11]: one directly predicts future wind power based on historical wind power, the other first predicts meteorological factors such as wind speed and direction, and then makes predictions based on wind power models.

However, wind power is affected by many complex factors. How to analyse the impact of various factors on wind power output quantitatively, so as to improve the accuracy of wind power forecasting and improve the awareness of wind power operation characteristics, has become a major difficulty faced by power grid dispatching operations. It includes methods based on physical models and those based on mathematical statistics. The physical model method establishes a physical model of wind power output based on the aerodynamic principles of wind turbines, and then analyses the influence of external factors on wind power based on the physical model [12–14]. However, because of the complex relation between various factors and wind power output, it is difficult to establish an accurate physical model [15]. Statistical-based methods are based on actual historical wind power data, using correlation analysis and other methods to analyse the impact of various factors on wind power [16–18]. However, there is a highly complex non-linear relation between wind power and various factors, and there are also interactive correlations between various factors [19]. It is difficult
to analyse the non-linear effects of factors on wind power quantitatively using statistical analysis methods. In addition, the correlation between factors and wind power has strong temporal characteristics, and such methods cannot measure the influence of changes in factors on the trend of wind power.

This work combines machine learning technology with the Shapley value that comes from cooperative game theory to analyse the complex non-linear effects of environmental factors on wind power. Machine learning technology has powerful feature extraction and non-linear modelling capabilities and has achieved good results in the fields of wind power prediction and wind power modelling [20]. However, the wind power model trained with machine learning technology has a black-box characteristic and cannot explain the non-linear relation between factors and wind power. This research uses the Shapley value to explain the implicit relation of the model, to realize the quantitative analysis of the influence of environmental factors on wind power. The contributions of this work are thus:

1) The importance of environmental factors' non-linear impact on wind power can be measured based on the Shapley value, to identify important factors affecting power output.
2) The complex non-linear relation between environmental factors and wind power can be analysed based on the Shapley value. And the interaction effect of two environmental factors' impact on wind power can also be analysed by the Shapley interaction value of factors.
3) The black-box characteristic of the wind power model based on machine learning is solved using the Shapley value, to gain insight into the prediction process of the wind power model and to realize the interpretability of the prediction results.

The purpose of this work is not to improve the performance of the wind power model algorithm, but based on the wind power model, to use the Shapley value to analyse the complex non-linear relation between the influencing factors and wind power. The remainder of this work is organised as follows. Section 2 analyses the main factors and characteristics that affect wind power output and the machine learning method of wind power model construction. Section 3 gives the definition, properties and calculation method of the Shapley value of environmental factors affecting wind power output. In Section 4, we use actual wind farm power data to validate the method. Conclusions are given in Section 5.

2 | INFLUENCING FACTORS AND MODELLING OF WIND POWER

2.1 | Analysis of factors affecting wind power

The physical law of wind power generation states that [21]:

$$ y = \frac{1}{2} C_p \rho \pi R^2 V^3 $$

(1)

where $y$ is the output power; $R$ is the radius of the rotor; $V$ is wind speed; $\rho$ is air density; and $C_p$ is the power coefficient, which is believed to be a function of (at least) the blade pitch angle and the turbine's tip speed ratio.

What else might affect $C_p$ is still under debate. Currently, no formula exists to express $C_p$ analytically in terms of its influencing factors [22]. Therefore, $C_p$ is empirically estimated. Turbine manufacturers provide the nominal power curve for a specific turbine with the corresponding $C_p$ values under different combinations of wind speed, $V$, and air density, $\rho$. This expression also provides the rationale for why temperature, $T$, and air pressure, $P$, are converted into air density, $\rho$, to explain wind power, rather than used individually.

Although on the surface, the expression in Equation (1) suggests that the electrical power a wind turbine extracts from the wind is proportional to $V^3$, an actual power curve may exhibit a different non-linear relation. This happens because the tip speed ratio is a function of wind speed, $V$, making $C_p$ also a function of $V$ and adding complexity to the functional relation between wind speed and wind power.

The underlying physics of wind power generation expressed in Equation (1) provide some clues concerning a preferable power curve model. Thus [23]:

1) There appear to be at least three important environmental factors that affect wind power generation: wind speed, wind direction, and air density. This does not exclude the possibility that other environmental factors may also influence the power output.
2) The functional relations between the environmental factors and the power response are generally non-linear. The complexity comes, in part, from the lack of understanding of power coefficient $C_p$, which is affected by many environmental factors. Because there is no analytical expression linking the power coefficient to any influencing factors, the functional form of the power curve is unknown.
3) The environmental factors appear in a multiplicative relationship in the power law equation.

2.2 | Wind power model based on machine learning

A wind power model describes the corresponding relation between environmental factors and the output power of the wind turbine. The first principal use of the wind power model is for wind power forecasting. Wind power forecasting can be done by forecasting wind speed first and then converting a speed forecast to a power forecast through the use of a power model. The second principal use of the wind power model is for turbine performance assessment and turbine health monitoring, in which the wind power model is used to characterise a turbine’s power production efficiency [24, 25]. In both applications, accurate modelling of the power is essential, because it underlies subsequent decision making.

However, there is complicated non-linear relation between environmental factors and wind output power, and it is difficult
to use mathematical analysis to establish an accurate wind power model. Therefore, the use of machine learning methods to build wind power models based on actual data has become an option. From the perspective of machine learning, wind power modeling is a typical supervised learning problem. It is the process of building a wind power model based on historical wind power data and historical environmental factor data. The process of model establishment is the process of training with historical data to minimise the expected value of loss, as in Equation (2):

$$\hat{F}(X) = \arg \min_{\hat{F}} E_{X,Y} \left[ L \left( Y, \hat{F}(X) \right) \right]$$

(2)

where $\hat{F}(\cdot)$ is the wind power model; $L(\cdot)$ is the loss function; $Y = \{y_i\}_{i=1}^{n}$ are output wind power; $X = \{x_i\}_{i=1}^{n}$ are data of factors affecting wind power; $x_i = \{x_i^1, x_i^2, ..., x_i^{j-1}, x_i^j, x_i^{j+1}, ..., x_i^n\}$ is the $i$th instance in the data; $x_i^j$ is the $j$th environmental factor in the $i$th sample; $p$ is the number of environmental factors affecting wind power; and $n$ is the number of training data instances.

Many supervised machine learning algorithms can be used to build wind power models, including artificial neural networks (ANNs) [26, 27], support vector machines (SVMs) [28], and gradient boosting decision trees (GBDTs) [29]. Although the machine learning algorithm has achieved good results, it has a black-box characteristic and cannot explain the complex non-linear relation between environmental factors and wind power. If we can quantitatively analyse the complex relation between factors and wind power in the wind power model, we can quantitatively analyse the impact of environmental factors on wind power output.

3. ANALYSIS OF FACTORS AFFECTING WIND POWER BASED ON SHAPLEY VALUE

3.1 Shapley value of environmental factors affecting wind

The Shapley value is a method of distributing expenditures for players based on their contribution to total expenditures in cooperative game theory [30]. Wind power is affected by many environmental factors, and the contribution of each factor value to wind power output can be measured by the Shapley value. Do not become confused by the word ‘value’: The factor value is the numerical or categorical value of an environmental factor; the Shapley value is the contribution of environmental factor to the wind power. After the wind power model training is complete, wind power output based on the power model can be explained by assuming that each environmental factor is a ‘player’ in a game where the output is the payout. The Shapley values tell us how to distribute the ‘pay out’ fairly among environmental factors of wind power. Using the Shapley value, we can measure the contribution of each factor to the wind power output.

The Shapley value of an environmental factor value is its contribution to the wind power output based on the wind power model, weighted and summed over all possible environmental factor combinations. For an instance of historical data of factors affecting wind power $x_i = \{x_i^1, x_i^2, ..., x_i^{j-1}, x_i^j, x_i^{j+1}, ..., x_i^n\}$, the Shapley value of the $j$th environmental factor value is defined as in Equation (3) [30, 31]:

$$\phi_{i}^{(j)}(\text{val}) = \sum_{S \subset \{x_1^1, x_1^2, ..., x_1^{j-1}, x_1^j, x_1^{j+1}, ..., x_1^n\} \setminus \{x_i^j\}} \frac{|S|!(p - |S| - 1)!}{p!} \left( \text{val}(S \cup \{x_i^j\}) - \text{val}(S) \right)$$

(3)

where $\phi_{i}^{(j)}(\text{val})$ is the Shapley value of the $j$th environmental factor value of the $i$th data instance; $\{x_i^1, x_i^2, ..., x_i^{j-1}, x_i^j, x_i^{j+1}, ..., x_i^n\}$ is the $i$th data instance; $p$ is the number of environmental factors affecting wind power; $S$ is subset of the remaining environmental factors after excluding the $j$th factor in the $i$th data instance; $|S|$ is the number of environmental factors in $S$; and $\text{val}(S)$ is the value function that indicates the degree of influence of the environmental factors in $S$ on the result of wind power output through ‘cooperation’; it can use the environmental factors in $S$ to calculate the marginal output of the environmental factors not included in $S$ on the wind power model, as in Equation (4):

$$\text{val}(S) = \int \hat{F}(x_1^1, ..., x_i^j, ..., x_i^n) dP_{x \not\in S}$$

$$-\frac{1}{n} \sum_{i=1}^{n} \hat{F}(\{x_1^1, x_i^2, ..., x_i^j, ..., x_i^n\})$$

(4)

where $\hat{F}(\cdot)$ is the wind power model; $p$ is the number of environmental factors; and $n$ is the number of data instances; $x = \{x_1^1, x_i^2, ..., x_i^j, ..., x_i^n\} \setminus S$ indicates factors that are not included in $S$.

We actually perform multiple integrations for each environmental factor that is not contained in $S$. A concrete example follows: The wind power model works with four environmental factors $x_1^1, x_1^2, x_1^3, x_1^4$, and we evaluate the prediction for coalition $S$ consisting of environmental factor values $x_1^1$ and $x_1^3$ according to Equation (5):

$$\text{val}(S) = \text{val}(\{x_1^1, x_1^3\}) = \int \int \hat{F}(x_1^1, x_1^2, x_1^3, x_1^4) dP_{x \not\in S}$$

$$-\frac{1}{n} \sum_{i=1}^{n} \hat{F}(\{x_1^1, x_i^2, x_i^3, x_i^4\})$$

(5)

3.2 Properties of the Shapley value

The Shapley value is the only attribution method that satisfies the properties Efficiency, Symmetry, and Dummy,
which together can be considered a definition of a fair payout [30].

1) Efficiency: The contributions of factors in a data instance must add up to the difference of predicted value for this instance and the average of the predicted values of all data instances:

\[ \sum_{j=1}^{p} \phi_{i}^{(j)} = \bar{F}(x_{i}^{1}, \ldots, x_{i}^{j-1}, x_{i}^{j+1}, \ldots, x_{i}^{p}) \]

\[ - \frac{1}{n} \sum_{i=1}^{n} \bar{F}(x_{1}^{1}, \ldots, x_{j}^{j-1}, x_{j}^{j+1}, \ldots, x_{p}^{p}) \]

where \( \bar{F}(\cdot) \) is the wind power model; \( p \) is the number of influencing factors; and \( n \) is the number of data instances.

2) Symmetry: The contributions of two factor values, \( j \) and \( k \), should be the same if they contribute equally to all possible coalitions. Given a subset of factors affecting wind power, \( S \subseteq \{ x_{1}^{1}, x_{1}^{2}, \ldots, x_{p}^{p} \} \setminus \{ x_{k}^{j}, x_{k}^{k} \} \), if:

\[ \text{val}(S \cup \{ x_{i}^{j} \}) = \text{val}(S \cup \{ x_{i}^{k} \}) \]

then \( \phi_{i}^{(j)} = \phi_{i}^{(k)} \).

3) Dummy: A factor \( j \) that does not change the predicted wind power, regardless of which coalition of factor values to which it is added, should have a Shapley value of 0. Given a subset of factors affecting wind power \( S \subseteq \{ x_{1}^{1}, x_{1}^{2}, \ldots, x_{p}^{p} \} \), if:

\[ \text{val}(S \cup \{ x_{i}^{j} \}) = \text{val}(S) \]

then \( \phi_{i}^{(j)} = 0 \).

The properties of the Shapley value can ensure that the contribution of each environmental factor to the wind power is fairly distributed, thus reflecting the impact of each environmental factor on wind power.

### 3.3 Method of Shapley value calculation

All possible coalitions (sets) of environmental factors have to be evaluated with and without the \( j^{th} \) factor to calculate the exact Shapley value. For more than a few factors, the exact solution to this problem becomes problematic because the number of possible coalitions exponentially increases as more factors are added [31]. After the wind power model training is completed, the number of model input influence factors is fixed, and the influence factors cannot be truly excluded from the data set to calculate the Shapley value. We use an approximation with Monte-Carlo sampling to calculate the Shapley value [32]. The approximate Shapley estimation for single factor value is [33, 34]:

**Steps for calculating the Shapley value**:

1. For all \( m = 1, \ldots, M \): Perform steps 2–7.
2. Draw random instance \( z \) from the data matrix \( X \).
3. Choose a random permutation \( o \) of the feature values.
4. Order instance: \( x_{\varphi_{z}} = (x_{1}, x_{2}, \ldots, x_{n}) \).
5. Order instance: \( z_{\varphi_{z}} = (z_{1}, z_{2}, \ldots, z_{n}) \).
6. Construct two new instances:
   a. With feature \( j \): \( x_{\varphi_{zj}} = (x_{1}, \ldots, x_{j}, z_{j}, x_{j+1}, \ldots, x_{n}) \).
   b. Without feature \( j \): \( x_{\varphi_{zj}} = (x_{1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{n}) \).
7. Compute marginal contribution:

\[ \phi_{j} = \bar{F}(x_{\varphi_{zj}}) - \bar{F}(x_{\varphi_{zj}}) \]

8. Compute Shapley value as the average:

\[ \phi_{j} = \frac{1}{M} \sum_{m=1}^{M} \phi_{j}^{m} \]

First, select an instance of interest \( x \), a feature \( j \) and the number of iterations \( M \). For each iteration, a random instance \( z \) is selected from the data and a random order of the features is generated. Two new instances are created by combining values from the instance of interest \( x \) and the sample \( z \). The first instance \( x_{\varphi_{zj}} \) is the instance of interest, but all values in the order before and including value of feature \( j \) are replaced by feature values from the sample \( z \). The second instance \( x_{\varphi_{zj}} \) is similar, but has all of the values in the order before but excluding feature \( j \) replaced by values of feature \( j \) from sample \( z \). The difference in the prediction from the wind power model is computed according to Equation (9). All of these differences are averaged and get the Shapley value of factor \( j \) according to Equation (10). This procedure has to be repeated for each factor to get all Shapley values.

**3.4 Measure the importance of impact of environmental factors on wind power based on Shapley value**

Based on the physical understanding hinted by the power generation law in Equation (1), it is apparent that wind speed, direction, and air density are important factors affecting wind power. The question is how to measure the importance of these factors. The Shapley value indicates the degree of influence of environmental factors on the wind power output. The greater the absolute value of Shapley, the greater the impact on wind power. The average absolute value of the Shapley value of each environmental factor in all data instances can measure the importance of impact of environmental factors on wind power, as in Equation (11):

\[ I_{j} = \frac{1}{n} \sum_{i=1}^{n} |\phi_{i}^{(j)}| \]
where \( I_j \) is the importance of environmental factors \( j \) measured by the average absolute value of the Shapley value; \( \phi^{(i)}_j \) is the Shapley value of the \( j^{th} \) environmental factor of the \( i^{th} \) instance; and \( n \) is the number of data instances.

### 3.5 Shapley interaction value of two factors affecting wind power

The interaction effect is the additional combined factors effect after accounting for the individual factor effects on wind power. The Shapley interaction value from game theory is defined after accounting for the individual factor effects on wind power:

\[
\phi_{ij} = \sum_{S \subseteq \{x, x_j\}} \frac{|S|!(p-|S|-2)!}{2(p-1)!} \delta_{ij}(S)
\]  

where \( i \neq j \) and:

\[
\delta_{ij}(S) = \hat{F}(S \cup \{x_i, x_j\}) - \hat{F}(S \cup \{x_i\}) - \hat{F}(S \cup \{x_j\}) + \hat{F}(S)
\]  

where \( S \) is a subset of the features used in the wind power model; \( x \) is the vector of environmental factors values of the instance to be explained; \( x_i \) and \( x_j \) are the \( i^{th} \) and \( j^{th} \) factors of interest; \( \hat{F}(\cdot) \) is the wind power model; and \( p \) is the number of factors affecting wind power.

Equation (13) subtracts the main effect of the features so that we get the pure interaction effect after accounting for the individual effects. We average the values over all possible feature coalitions \( S \), as in the Shapley value computation.

### 4 CASE STUDIES

4.1 Process of the experiment

The method proposed here is used to analyse environmental factors affecting the power output of a wind farm. The detailed process of this experiment is shown in Figure 1:

1) Collect data and normalise wind power based on the rated power of the wind farm and preliminarily analyse the influence of environmental factors on wind power during exploratory data analysis.
2) Divide the data into a training set, validation set, and test set. The training set is used to train the wind power model, the validation set is used to find the optimal hyper-parameters of the model, and the test set is used to evaluate the performance of the model.
3) ANNs, SVMs, and GBDTs are used to train wind power models separately, and Bayesian optimization is used to find the optimal hyper-parameters based on the validation set [35], so that the prediction performance of each algorithm is optimal. The mean squared loss is used as the loss function, as shown in Equation (14):

\[
L(Y, \hat{F}(X)) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \hat{F}(x_i) \right)^2
\]  

where \( L(\cdot) \) is the loss function; \( \hat{F}(\cdot) \) is the wind power model; \( Y = \{y_i\}_{i=1}^{n} \) are real wind power; \( X = \{x_i\}_{i=1}^{n} \) are data of factors affecting wind power; and \( n \) is the number of training instances.

4) Calculation of the Shapley value is based on the wind power model, and the performance of the model directly affects the accuracy of the Shapley value. Therefore, use the test set to evaluate the three models and select the model with the best performance based on performance evaluation metric defined as the absolute error:

\[
e_i = \left| \hat{F}(x_i) - y_i \right|
\]  

where \( \hat{F}(\cdot) \) is the wind power model; \( y_i \) are real wind power of the \( i^{th} \) instance; and \( x_i \) are environmental factor values of the \( i^{th} \) instance.

5) Calculate the Shapley values of each environmental factor value of each data instance based on the wind power model. Then, analyse the relation between factors and wind power and observe the forecasting process of wind power model based on the Shapley value.

### 4.2 Exploratory data analysis of experimental data

The datasets used in this case study come from an inland wind farm used in Gwhyun et al. [23]. The turbine specifications in this wind farm are shown in Table 1.

Based on the historical environmental factors and historical power data of this wind farm, the impact of its important environmental factors on wind power is analysed. It can be seen from Equation (1) that the output power of a wind turbine is directly related to the characteristics of the wind. In addition, output power is affected by environmental factors such as the wind’s direction, air density, turbulence intensity, and wind shear. Thus, the selected environmental factors are shown in Table 2.

The correlation between the normalized wind power of the wind farm and its environmental factors is shown in Figures 2 and 3. Figure 2a–d are scatterplots between wind power and wind speed, air density, turbulence intensity, and wind shear, respectively. Figure 2 shows that between the cut-in wind speed and the rated wind speed, the wind power...
and the wind speed have a linear positive correlation. On the surface, there is no obvious correlation between the other three factors and wind power. These scatterplots in Figure 2 are unconditional for wind speed and wind direction. In this setting, these environmental factors show no obvious effect on the power output. Giwhyun et al. [23] analyses the same variables but under different wind speeds and wind directions; the result shows that interaction effects exist among wind speed, wind direction, and other environmental factors. Figure 3 shows that wind direction also affects wind power.

### Table 1 Specifications of turbines in wind farm used in experiment

| Parameter                  | Configuration |
|----------------------------|---------------|
| Number of wind turbines    | 200           |
| Hub height (m)             | 80            |
| Rotor diameter (m)         | About 80      |
| Cut-in wind speed (m/s)    | 3.5           |
| Cut-out wind speed (m/s)   | 20            |
| Rated wind speed (m/s)     | Around 13     |
| Rated power (MW)           | 1.5-2.0       |

### Table 2 Main factors affecting wind power considered in the experiment

| Environmental factors | Description |
|-----------------------|-------------|
| Wind speed (m/s)      | Wind speed at height of wind turbine hub |
| Turbulence intensity  | Ratio of standard deviation of fluctuating wind velocity to mean wind speed |
| Air density (Kg/m³)   | Air density where wind turbine is located |
| Wind direction (°–360°) | The direction from which it originates |
| Wind shear (m/s)      | Changes in wind direction and speed under wind turbine hub |

## 4.3 Comparison of performance of different models

ANNs, SVMs, and GBDTs are used to train the wind power model based on the training set and validation set, and the test set is used to evaluate model performance. A box plot of the absolute error of each type of algorithm is shown in Figure 4. Figure 4 shows that the mean of the absolute error of the GBDT algorithm on the test set is relatively small and the
performance is the best. GBDT has strong feature extraction capabilities and generalisation capabilities; it can achieve better prediction results on a smaller training set. Therefore, the

4.4 | Analysis of factors affecting wind power based on Shapley value

Based on the wind power model $\tilde{F}(\cdot)$ trained by GBDT, and the entire data set, the Shapley value calculation method is used
Table 3 Importance of impact of environmental factors on wind power

| Environmental factors | Importance of impact on wind power (average absolute value of Shapley value) |
|-----------------------|--------------------------------------------------------------------------------|
| Wind speed            | 30                                                                              |
| Turbulence intensity  | 0.55                                                                            |
| Air density           | 0.54                                                                            |
| Wind direction        | 0.44                                                                            |
| Wind shear            | 0.23                                                                            |

to calculate the Shapley value of each environmental factor of each instance in the data set. We analyse environmental factors affecting wind power based on the Shapley value of environmental factors.

4.4.1 Identify important environmental factors affecting power output

The importance of the impact of environmental factors on wind power is calculated according to Equation (11) based on the Shapley value. The result is shown in Table 3.

The result shows that wind speed is the main factor affecting wind power. Other environmental factors also have an impact on wind power. Their order of importance is turbulence intensity, air density, wind direction, and wind shear.

4.4.2 Analyse the relation between environmental factors and wind power

The distribution of the Shapley value of each environmental factor of wind power is shown in Figure 5.

Figure 5 combines the Shapley value of environmental factors with factor values. Each point in Figure 5 is a Shapley value for a factor and an instance. The position on the vertical axis is determined by environmental factors, and on the horizontal axis, by the Shapley values. The colours represent the values of environmental factors from low to high. Overlapping points are jittered in vertical axis direction, so we get a sense of the distribution of the Shapley values per factor.

Figure 5 shows that the higher the wind speed, the greater the impact on the increase of wind power. This is consistent with the commonsense view that wind speed is the main factor affecting wind power. In addition to wind speed, wind direction, air density, turbulence intensity, and wind shear all affect wind power. The greater the turbulence intensity and the air density, the greater the impact on the decrease in wind power.

Indications of the relation between the value of an environmental factor and the impact on the wind power are shown in Figure 5. However, the form of the relation is not shown. To measure the impact of the changes of environmental factors values on wind power clearly, wind speed and its corresponding Shapley value are combined with the values of wind direction, air density, turbulence intensity, and wind shear in the form of a scatterplot, as shown in Figure 6.

In Figure 6, the horizontal axis represents wind speed, the vertical axis represents the Shapley value of wind speed, and the colours represent the values of corresponding environmental factors.

Figure 6 shows that sometimes when the wind speed exceeds the rated wind speed (around 13 m/s), the impact of wind speed on wind power will be reduced. What is the reason? Figure 6a,b,d shows that when the corresponding influencing factors change, the impact on wind power is small. However, Figure 6c shows that when the wind speed is above the rated wind speed, a reduction in the impact of wind speed on wind power is caused by an increase in turbulence intensity.

4.4.3 Analyse the interaction effect of two factors affecting wind power

The interaction effect of two factors affecting wind power can be measured by the Shapley interaction value of factors. The calculated Shapley interaction value of wind speed and other environmental factors are shown in Figure 7.

Compared with Figure 6, it can be seen that the increase in turbulence intensity is the main reason for the reduction in the impact of wind speed on wind power when the wind speed is above the rated wind speed.

4.4.4 Comparison with partial dependence method

To verify these results, based on the wind power model \( \hat{F}(\cdot) \) trained by GBDT, partial dependence [34] is used to measure the influence of wind speed and turbulence intensity on wind power. The result is shown in Figure 8.

Figure 8 shows that when the wind speed is above the rated wind speed, there is a collapsed area when the turbulence intensity is high. It also shows that the increase in turbulence intensity is the main reason for the reduction in the impact of wind speed on wind power when the wind speed is above the rated wind speed. However, compared with the Shapley value, it is not obvious.

4.5 Analysis of wind power forecasting process based on Shapley value

The Shapley value can be used to gain insight into the process of wind power model to predict wind power. Observe the impact of each type of environmental factors on the forecast results of wind power during forecasting.

Use the Shapley value to observe the wind power prediction result at a certain moment, as shown in Figure 9.

In Figure 9, the upper and lower horizontal axes represent output wind power of the wind power model, the middle vertical axis represents the average predicted wind power of
FIGURE 5 Distribution of Shapley values for each environmental factor

FIGURE 6 Shapley value of wind speed combined with values of other factors. (a) Wind direction. (b) Air density. (c) Turbulence intensity. (d) Wind shear

FIGURE 7 Shapley interaction value of wind speed and other environmental factors combined with values of corresponding factors: (a) wind direction; (b) air density; (c) turbulence intensity; (d) wind shear
the all data instances, and the left vertical axis represents the environmental factors of wind power. The curve in Figure 9 represents the process of predicting wind power using the model. The value in brackets is the value of environmental factors. The curve starts from the position of the average predicted wind power on the lower horizontal axis and is affected by environmental factors to reach the predicted power on the upper horizontal axis. The magnitude of the impact is measured by the Shapley value, as in Table 4.

The impact of various environmental factors on wind power at this moment is as follows: wind shear has little effect on wind power, air density and wind direction have a negative impact on wind power, and turbulence intensity and wind speed have a positive impact on wind power. The negative and positive results are for the difference in predicted value for this moment (data instance) and the average of the predicted values of all data instances. According to the efficiency properties of the Shapley value, the sum of the Shapley value of all environmental factors at this moment is equal to the difference between the predicted wind power at that moment and the average of the predicted powers of all data instances.

5 | CONCLUSIONS

This work combines machine learning technology with the Shapley value in cooperative game theory to analyse the correlation between environmental factors and wind power. This method can be used to measure the influence of environmental factors on wind power and to identify main factors affecting wind power. It can also be used to analyse the complex nonlinear relation between each environmental factor and wind power and to gain insight into the prediction process of the wind power model. Based on the historical data of a wind farm, this method is used to analyse environmental factors affecting wind power. The experimental results found that in addition to wind speed, turbulence intensity, air density, wind shear, and wind direction all affect the output of this wind farm. Moreover, when the wind speed is greater than the rated wind speed, the
reduction in the influence of wind speed on wind power is mostly caused by an increase in turbulence intensity.

The core idea of this research is to use Shapley values to reveal the complex relations hidden in the wind power model trained by machine learning techniques. Therefore, the performance of the wind power model directly affects the calculation effect of the Shapley value. Before using this method, an accurate reduction in the influence of wind speed on wind power is mostly different, and the analysis of the data of different wind farms using this method may lead to different conclusions.

ACKNOWLEDGEMENTS
This work was supported by State Grid Science and Technology Project (Project No. 5700-20205368A-0-0-00). The research content of this project is power grid operation data prediction, mining and analysis.

ORCID
Chuanjun Pang https://orcid.org/0000-0003-0766-8831

REFERENCES
1. Ahmed, S.D., et al.: Grid integration challenges of wind energy: a review. IEEE Access. 8(5), 10857–10878 (2020)
2. Temitope, R.A., et al.: Challenges of grid integration of wind power on power system grid integrity: a review. Int. J. Renew. Energy Res. 2(4), 618–626 (2012)
3. Baghace, H.R., et al.: Fuzzy unscented transform for uncertainty quantification of correlated wind/ PV microgrids: possibilistic-probablistic power flow based on RBFNNs. IET Renew. Power Gener. 11(6), 867–877 (2017)
4. Noradzin, G., et al.: Two stage forecast engine with feature selection technique and improved meta-heuristic algorithm for electricity load forecasting. Energy, 161(15), 130–142 (2018)
5. Yang, Y., et al.: Impact of down spinning reserve on operation reliability of power systems. J. Mod. Power Syst. Clean Energy, 8(4), 709–718 (2020)
6. Xiao, S., et al.: Analysis of factors affecting wind farm output power. In: 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, pp. 1–5 (2018)
7. Xue, Y., et al.: A review on short-term and ultra-short-term wind power prediction. Autom. Electr. Power Syst. 39(6), 141–151 (2015)
8. Shi, K., et al.: Short-term wind power prediction based on entropy association information mining of historical data. Autom. Electr. Power Syst. 41(5), 15–18 (2017)
9. Mao, Y., Shaoshuai, W.: A review of wind power forecasting & prediction. In: International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Beijing, pp. 1–7 (2016)
10. Iltani, C., Seref, S., Mehmet, Y.: Data mining and wind power prediction: a literature review. Renew. Energy, 46, 241–247 (2012)
11. Zhao, Q., et al.: A review and discussion of decomposition-based hybrid models for wind energy forecasting applications. Appl. Energy, 235(1), 939–953 (2019)
12. El-Ahm, M.H., El-Sayed, A.M., Hemeida, A.M.: Evaluation of factors affecting wind turbine output power. In: Proceedings of 2017 Nineteenth International Middle East Power Systems Conference (MEPCON), Cairo, USA, pp. 1471–1476 (2017)
13. Wiratama, I.K., Mara, I.M., Nuarsa, I.M.: Investigation of factors affecting power curve wind turbine blade. J. Eng. Appl. Sci. 11(4), 2759–2762 (2016)
14. Muthu, M., Victor, K.: A critical review of factors affecting wind turbine and solar cell system power production. Int. J. Adv. Eng. Res. Stud. 3, 2249–8974 (2014)
15. Yang, M., Yang, Q.: Review of modelling of wind speed-power characteristic curve for wind turbine. Electr. Power Autom. Equip. 38(2), 34–43 (2018)
16. Jafariyan, M., Soroudi, A., Elsman, M.: The effects of environmental parameters on wind turbine power pf curve. In: Proceedings of Canadian Conference on Electrical and Computer Engineering, Canadian, pp. 001193–001198 (2008)
17. Jayanna, K.: Aerodynamic factors affecting wind turbine power generation. Int. J. Eng. Res. Technol. 11(9), 2278-0181 (2012)
18. Jeon, J., Taylor, J.W.: Using conditional kernel density estimation for wind power density forecasting. J. Am. Stat. Assoc, 10(7), 66–79
19. Ding, Y., Tang, J., Huang, J.Z.: Data analytics methods for wind energy applications. In: Proceedings of ASME Turbo Expo 2015: Turbine Technical Conference and Exposition, Montreal, Canada, pp. 1–9 (2015)
20. Lydia, M., et al.: A comprehensive review on wind turbine power curve modelling techniques. Renew. Sustain. Energy Rev. 30(2), 452–460 (2014)
21. Lin, P., et al.: Wind power curve modelling based on measured data and uncertainty estimation. Electr. Power Autom. Equip. 35(4), 90–95 (2015)
22. Belgazi, O., Cherkaoui, M.: Pitch angle control for variable speed wind turbines using genetic algorithm controller. J. Theor. Appl. Inf. Technol. 39(1), 6–10 (2012)
23. Giwihyuu, L., et al.: Power curve estimation with multivariate environmental factors for inland and offshore wind farms. J. Am. Stat. Assoc. 110(509), 56–67 (2015)
24. Eunsun, B.: Wind turbine operations and maintenance: a tractable approximation of dynamic decision making. IESE Trans. Taylor & Francis J. 45(11), 1188–1201 (2013)
25. Byron, E., Ntario, L., Ding, Y.: Optimal maintenance Strategies for wind turbine systems under stochastic weather conditions. IEEE Trans. Reliab. 59(2), 393–404 (2010)
26. Gong, L., Jing, S.: On comparing three artificial neural networks for wind speed forecasting. Appl. Energy, 87(7), 2313–2320 (2010)
27. Medina, S.V., Ajenjo, U.P.: Performance improvement of artificial neural network model in short-term forecasting of wind farm power output. J. Mod. Power Syst. Clean Energy, 8(3), 484–490 (2020)
28. Maria, Grazia De Giorgi., et al.: Comparison between wind power prediction models based on wavelet decomposition with least-squares support vector machine (LS-SVM) and artificial neural network (ANN). Energies, 7(8), 5251–5272 (2014)
29. Li, G., Jiang, X.: Wind power generation ensemble forecasting method based on boosted regression tree and random forest. Proc. CSU-EPSA. 30(11), 70–74 (2018)
30. Lloyd, S.S., A value for n-person games, pp. 307–317. RAND Corpora- tion, (1952)
31. Lundberg, Scott, Su-In, Lee.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Chicago, USA, pp. 4768–4777 (2017)
32. Consistent individualized feature attribution for tree ensembles. https://arxiv.org/abs/1802.03888. Accessed 12 August 2020
33. Lundberg, S.M., et al.: From local explanations to global understanding based on boosted tree models. Proc. Interspeech, 2019, 1577–1581 (2018)
34. Lundberg, S.M., et al.: Consistent individualized feature attribution for tree ensembles. https://christophm.github.io/interpretation-ml-book/. Accessed 12 August 2020
35. Jia, W., et al.: Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. 17(1), 26–40 (2019)

How to cite this article: Pang C., Ju Y., Liu Y. Correlation analysis of factors affecting wind power based on machine learning and Shapley value. IET Energy Syst. Integr. 2021,1–11. https://doi.org/10.1049/est2.12022