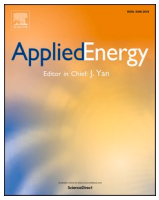
Applied Energy 376 (2024) 124273

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Can we trust explainable artificial intelligence in wind power forecasting?

Wenlong Liao [a,*], Jiannong Fang [a], Lin Ye [b], Birgitte Bak-Jensen [c], Zhe Yang [d], Fernando Porte-Agel [a]

[a] Wind Engineering and Renewable Energy Laboratory, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne 1015, Switzerland
[b] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[c] AAU Energy, Aalborg University, Aalborg 9220, Denmark
[d] Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong

HIGHLIGHTS

- Four explainable artificial intelligence techniques are tailored to provide interpretability for machine learning models.
- Multiple metrics are defined to evaluate the trustworthiness of the interpretability.
- Explainable artificial intelligence techniques are extensively investigated on real datasets and machine learning models.

ARTICLE INFO

Keywords:
Wind power forecast
Explainable artificial intelligence
Trustworthy
Neural network
Time series

ABSTRACT

Advanced artificial intelligence (AI) models typically achieve high accuracy in wind power forecasting, but their internal mechanisms lack interpretability, which undermines user confidence in forecast value and strategy execution. To this end, this paper aims to investigate the interpretability of AI models, which is crucial but usually overlooked in wind power forecasting. Specifically, four model-agnostic explainable artificial intelligence (XAI) techniques (i.e., Shapley additive explanations, permutation feature importance, partial dependence plot, and local interpretable model-agnostic explanations) are tailored to provide global and instance interpretability for AI models in wind power forecasting. Then, several metrics are proposed to evaluate the trustworthiness of interpretations provided by XAI techniques. Simulation results demonstrate that the proposed XAI techniques can not only identify important features from wind power datasets, but also enable the understanding of the contribution of each feature to the forecast power output for a specific sample. Furthermore, the proposed evaluation metrics aid users in comprehensively assessing the trustworthiness of XAI techniques in wind power forecasting, enabling them to judiciously select suitable XAI techniques for their AI models.

1. Introduction

Wind power is a highly effective renewable energy source for power generation in smart grids. The integration of wind power has surged globally over the past few decades [1]. In 2022, wind power generation in Switzerland increased by 5 % compared to the previous year. However, unlike traditional electricity generation (e.g., thermal power generation), wind power is subject to uncertainties [2]. To ensure the secure operation of power systems, it is imperative to forecast wind power accurately, particularly with the increasing integration of wind power.

Extensive studies have been conducted on wind power forecasting, and mainstream methods can be broadly classified into three major groups [3]: physical models, statistical models, and artificial intelligence (AI) models.

As far as the physical model is concerned, numerical weather prediction (NWP) data and environmental parameters like terrain features, are fed to complicated physical models to estimate wind speeds around blades. Then, wind power can be calculated by using the wind power curve, which maps the wind speed to power output with a wind power coefficient [4]. Despite the potential benefits of physical models for medium-term and long-term forecasting, they are hindered by

* Corresponding author.
E-mail addresses: wenlong.liao@epfl.ch (W. Liao), jiannong.fang@epfl.ch (J. Fang), yelin@cau.edu.cn (L. Ye), bbj@energy.aau.dk (B. Bak-Jensen), zhe1yang@polyu.edu.hk (Z. Yang).

https://doi.org/10.1016/j.apenergy.2024.124273
Received 10 May 2024; Received in revised form 22 June 2024; Accepted 16 August 2024
Available online 24 August 2024

computational complexity.

Regarding statistical models, the widely used methods in wind power forecasting involve linear regression, persistence model, autoregressive model, moving-average model, and gray model. For example, the work in [5] combines a seasonal moving average model with neural networks to capture the latent features of offshore wind power generation. In [6], a genetic algorithm is used to determine the optimal parameters of the gray model, which forecasts wind power and hydro power. To forecast the wind power of multiple wind farms, an autoregressive model is presented in [7] to account for the spatial dependencies in time series. Generally, statistical models are cost-effective for wind power forecasting, but their accuracy is relatively limited, especially at long look-ahead times.

Recently, there has been a surge of interest in the application of AI models for wind power forecasting. A few AI models (e.g., regression tree and k-nearest neighbors) are considered as glass-box models [8], meaning that their internal workings and decision-making processes are transparent and interpretable. However, these transparent AI models may not effectively model or extract intricate non-linear relationships. This limitation results in suboptimal performance, especially when dealing with complex wind power time series and NWP data. To improve the forecasting accuracy, recent publications have proposed more advanced and complex AI models [9], such as random forest, temporal convolutional network, transformer neural network, multi-layer perception (MLP) [10], and light gradient boosting machine (LightGBM) [11]. For example, to reduce forecast errors, the work in [12] designs an extreme gradient boosting (XGBoost) to forecast the short-term wind power considering extreme weather conditions. In [13], a long short-term memory (LSTM) and graph convolutional network are combined to capture spatio-temporal features from wind power. In [14], a gate recurrent unit (GRU) with skip connections is proposed to obtain the prediction intervals of wind power. Compared to statistical models and transparent AI models, these more recent and advanced AI models show superior performance. However, they are usually considered as black boxes, which have difficulty in understanding their forecasting mechanisms.

The lack of interpretability in most AI models undermines user confidence in forecast values and strategy execution. According to the definition in related publications [15,16], interpretability means that the user knows the contribution of each feature to the forecasts. To ensure the security and transparency of decision-making in power systems, it is crucial to address the interpretability of AI models. In fact, analogous challenges have been explored outside wind power forecasting but within the broader field of artificial intelligence [17]. For example, the work in [18] uses Shapley values to determine influential attributes of images. To increase the transparency of graph neural networks in text categorization, a local interpretable model-agnostic explanation (LIME) is presented to provide interpretability [19]. In [20], sensitivity analysis, fidelity correlation and monotonicity metrics are used to evaluate the explainable artificial intelligence (XAI) techniques. In [21], the explanatory significance assessment is presented to evaluate the accuracy and spatial precision of different XAI techniques. In [22], both problem specific and agnostic evaluation metrics are developed to interpret a model. However, these works mainly focus on computer vision or text categorization, and the interpretability of AI models in wind power forecasting has barely been explored. Moreover, most works pay attention to providing interpretability for the behavior of AI models, but the trustworthiness of the interpretation remains an issue. In other words, it is difficult to determine whether the interpretation can be trusted.

In response to the above limitations, this paper aims to answer the following two research questions: 1) *how to extend the XAI techniques from computer vision into wind power forecasting?* 2) *How to evaluate the trustworthiness of interpretation provided by XAI techniques for AI models in wind power forecasting.*

In particular, four model-agnostic XAI techniques are tailored to provide interpretability for wind power forecasting, and then several metrics are proposed to evaluate the trustworthiness of interpretations. The proposed XAI techniques in this paper have two advantages over previous studies. Firstly, while most studies focus on developing advanced AI models that lack interpretability, the XAI techniques proposed in this paper provide interpretability for these AI models. Secondly, the evaluation of XAI techniques has been overlooked in many studies, whereas the second advantage of this paper is that the proposed metrics effectively evaluate the trustworthiness of different XAI techniques. The key contributions are as follows:

- Four XAI techniques are reformulated to be tailored to provide interpretability for AI models (i.e., black boxes) in wind power forecasting, from both global interpretability and instance interpretability. Also, these XAI techniques are made applicable to a variety of AI models in wind power forecasting, as they are model-agnostic.
- Four metrics are proposed to evaluate the trustworthiness of the interpretation provided by XAI techniques a topic that is barely explored in wind power forecasting.
- The XAI techniques are extensively investigated and evaluated on real datasets and AI models in wind power forecasting.

The remaining sections are organized as follows. Section 2 formulates four XAI techniques for AI models in wind power forecasting. Section 3 presents four metrics to evaluate the trustworthiness of interpretations provided by these XAI techniques. Simulation and analysis are performed in Section 4. Lastly, Section V draws the conclusions.

2. Explainable artificial intelligence techniques

In this section, four XAI techniques are set up and tailored to provide both global interpretability and instance interpretability of the AI models. Global interpretability means that the user knows the average contribution (i.e., feature importance) of each feature to forecast over all the samples. In other words, global interpretability represents the average importance of each feature in the dataset, which plays an important role in feature engineering.

On the other hand, the instance interpretability means that the user knows the contribution of each feature to the forecast value for a given sample. For different samples, the features determining the wind power output may vary. For instance, at low wind speeds, the wind speed may play a dominant role in influencing wind power output. However, when the wind speed is between the rated wind speed and the cut-off wind speed, other features, such as wind direction, may become more influential. In contrast to global interpretability that focuses on overall and average trends, instance interpretability can reveal the importance of each feature to forecasts for a specific sample.

The following sections will tailor four XAI techniques, including Shapley additive explanations (SHAP) [18], permutation feature importance (PFI) [23], partial dependence plot (PDP) [24], and LIME [19], to provide interpretability for AI models in wind power forecasting. The selection of these four XAI techniques is motivated by two factors. First, their effectiveness has been well established in the broader field of XAI, ensuring significant application potential in the field of wind power forecasting. Second, these techniques provide comprehensive coverage of different interpretability needs. For example, as shown in Fig. 1, SHAP provides both global and instance interpretability, while

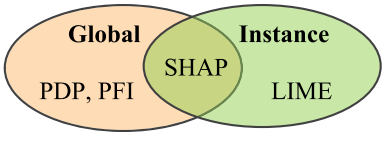


Fig. 1. The classification of four XAI techniques.

PDP and PFI primarily provide global interpretability and LIME focuses on instance interpretability.

A. *PFI*

As shown in Fig. 2, the PFI technique measures the feature importance by calculating the average difference in the forecast errors after permuting the feature [23]. If the model is highly dependent on a permuted feature, then the forecast errors will change significantly. Conversely, permuting a feature that is not important will not change the forecast errors of AI models.

First of all, an AI model (e.g., neural networks or tree models) is trained by using the original features, and then its average forecast error on the test set is evaluated by error metrics, such as mean square error and mean average error:

$$e_{\text{ori}} = G(Y, \text{AI}(X_{\text{ori}})) \tag{1}$$

where e_{ori} represents average forecast errors based on the original feature X_{ori}; Y represents real values; $\text{AI}(X_{\text{ori}})$ represents forecast values; G is a function to calculate the error metric.

Next, the feature i in the dataset is permuted to obtain the permuted feature $X_{\text{per},i}$, which serves as inputs of AI models to get the average forecast errors again:

$$e_{\text{per},i} = G\left(Y, \text{AI}\left(X_{\text{per},i}\right)\right) \tag{2}$$

where $e_{\text{per},i}$ represents average forecast errors based on the permuted feature.

Lastly, the average difference between forecast errors e_{ori} and $e_{\text{per},i}$ is considered as the importance of the feature i:

$$\text{FI}_i = \left|e_{\text{per},i} - e_{\text{ori}}\right| \tag{3}$$

where FI_i represents the average importance of the feature i.

B. *LIME*

As shown in Fig. 3, the LIME technique generates a set of perturbed samples near the sample to be interpreted, and then employs an interpretable glass-box model (e.g., linear model) to fit the AI model. In this case, the interpretable glass-box model is trained on the perturbed samples to provide insight into how each feature contributes to the forecasts for the specific sample under consideration [19].

Initially, an AI model (e.g., neural networks or tree models) is trained by using the original dataset.

Then, a specific sample X to be interpreted is selected. For example, a sample whose forecast error is large arouses the user's interest. A set of perturbed samples $(\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n)$ near this selected sample are generated by applying small random perturbations to each feature of this selected sample X.

After that, the perturbed samples $(\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n)$ are fed into the AI model and an interpretable glass-box model (e.g., linear model) to get their forecast values $(\text{AI}(\widetilde{X}_1), \text{AI}(\widetilde{X}_2), \ldots, \text{AI}(\widetilde{X}_n))$ and $(\text{IM}(\widetilde{X}_1), \text{IM}(\widetilde{X}_2), \ldots, \text{IM}(\widetilde{X}_n))$, respectively. IM represents an interpretable glass-box model. To provide insight into how each feature contributes to the forecasts for the specific sample, the interpretable glass-box model is trained to fit the AI model by minimizing their differences:

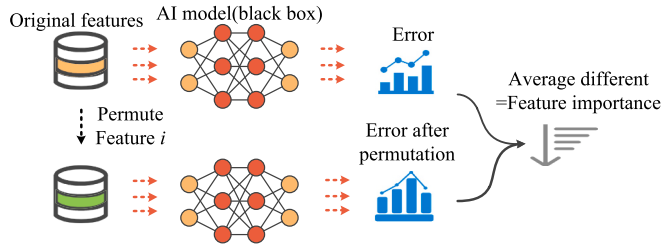


Fig. 2. A visual explanation of how to calculate the average feature importance with the PFI technique.

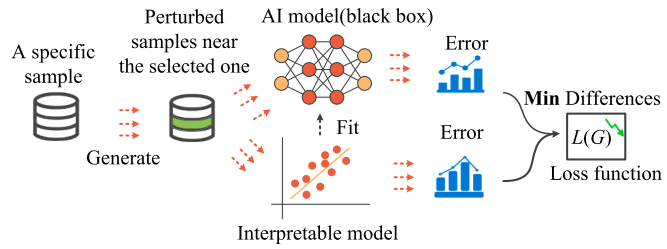


Fig. 3. A visual explanation of how to calculate the feature importance for a specific sample with the LIME technique.

$$\textit{argmin} \sum_{i=1}^{n} L(\text{AI}(\widetilde{X}_i), \text{IM}(\widetilde{X}_i)) \tag{4}$$

where L represents a loss function, such as mean square error.

Finally, the weight of the interpretable glass-box model can be used to understand the behavior of the AI model on this selected sample. In other words, the weights of the interpretable glass-box model help users know how each feature contributes to the forecasts.

C. *PDP*

The PDP can effectively reveal the nature of the association between the target variable (i.e., wind power) and a specific feature by visualizing the selected features and forecast value (i.e., wind power). Normally, the partial dependence function $\widehat{f}$ for wind power forecasting can be formulated as:

$$\widehat{f}_S(X_S) = E_{X_C}[\widehat{f}(X_S, X_C)] \tag{5}$$

where X_S represents the selected feature you are interested in (normally, the set S only includes one or two features); X_C represents the remaining features; E_{XC} represents the expectation with respect to X_C; and $\widehat{f}$ represents an AI model.

The essence of partial dependence is to marginalize the output of an AI model across the distribution of features in set C. This process enables the depiction of the connection between the features in set S, which is our focus, and the forecast value. By marginalizing the effects of the remaining features, we can obtain a function that depends solely on the features in set S, including their interactions with other features.

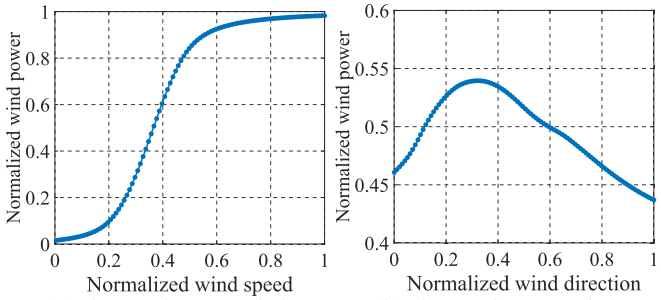
Further, the Monte Carlo method is adapted to approximate the above partial dependence function with training set [24]. This involves randomly sampling from the training set to approximate the expected value the AI model's output, thereby providing an estimation of the partial dependence function:

$$\widehat{f}_S(X_S) = \frac{1}{N} \sum_{i=1}^{N} \widehat{f}(X_S, X_C^i) \tag{6}$$

where N is the number of training samples; and X_C^i is the feature of the sample i in set C.

The PDP provides insight into the average marginal effect on the forecast value (i.e., wind power) for selected features. It tells us how changes in the features within set S affect the overall forecast results.

For example, we use the PDPs to reveal the nature of the association between the wind power and a specific feature when employing an MLP to forecast wind power on the dataset from GEFCom 2014 [25]. Fig. 4 shows the PDPs between wind power and selected features (e.g., wind speed and direction at 100 m).



(a) The PDP between wind speed at 100 m and wind power

(b) The PDP between wind direction at 100 m and wind power

Fig. 4. The PDPs between wind power, wind speed, and wind direction.

The initial increase in wind power output with increasing wind speed is evident from the PDPs. However, as the wind speed continues to increase, there comes a point at which the turbines reach their rated operational capacity, resulting in a fixed wind power output. This complex interplay between wind speed and wind power output as revealed by the PDP shows the physical characteristics inherent in the wind power curve.

Similarly, as the wind direction increases, there is an initial increase followed by a subsequent decrease in wind power output. In particular, the wind power output reaches its maximum when the wind direction is around 0.3. This phenomenon is attributed to the directional characteristics of the wind and its effect on the efficiency of the wind turbines. The initial increase in wind power output may be due to the optimal layout of the turbines with regard to the prevailing wind direction, resulting in decreased overall wake loss. However, as the wind direction continues to change, deviations from the optimal layout may occur, leading to a reduction in wind power output.

While direct quantitative measures of feature importance cannot be obtained from PDP, the slopes of the curves serve as indicative measures of feature importance. To measure the importance of a feature, we divide the PDP curve into m intervals, and compute the slope of each interval, and take average, which represents the feature importance:

$$\mathrm{SL}_i = \frac{\mathrm{PDP}_{i+1} - \mathrm{PDP}_i}{x_{i+1} - x_i}, i = 1, 2, \ldots m \tag{7}$$

$$\mathrm{FI}_i = \frac{1}{m}\sum_{i=1}^{m} |\mathrm{SL}_i| \tag{8}$$

where SL_i represents the slope of the interval i; and PDP_i represents the wind power when the feature value is x_i in the PDP curve.

D. *SHAP*

The SHAP estimates the feature importance by weighing and combining their contributions to all combinations with other features [18]:

$$\mathrm{FI}_i = \sum_{S\subseteq\{1,2,\ldots N\}\setminus\{i\}} \frac{(N - |S| - 1)!|S|!}{N!}(\mathrm{AI}(S\cup\{i\}) - \mathrm{AI}(S)) \tag{9}$$

where N is the number of features; S is a subset of the features; and $\mathrm{AI}(S)$ is the forecast power output using the features as inputs in set S.

In practice, the feature importance is usually calculated by using approximation techniques (e.g., Gaussian kernel-based SHAP), because the consideration of all the feature combinations in Eq. (9) would impose a significant computational burden [18]. Note that the Gaussian kernel-based SHAP is model-agnostic, and thus can be used in different AI models, such as neural networks and tree models. Specifically, the Gaussian kernel can be formulated as follows:

$$K(x_i, x_j) = exp\left(\frac{\|x_i - x_j\|^2}{-2\sigma^2}\right) \tag{10}$$

where $K(\cdot)$ is the Gaussian kernel between feature x_i and feature x_j, in which bandwidth parameter is σ.

Then, the Gaussian kernel is applied to Eq. (9) to obtain a Gaussian kernel-based feature importance:

$$\mathrm{FI}_i = \sum_{S\subseteq\{1,2,\ldots N\}\setminus\{i\}} \frac{K(S,\widehat{S})}{Z}(\mathrm{AI}(S\cup\{i\}) - \mathrm{AI}(S)) \tag{11}$$

where $\widehat{S}$ is a randomly sampled subset from the feature set; and Z is a normalization term that ensures the weights of all sampled subsets sum to 1.

In addition to the Gaussian kernel, there are other kernels (e.g., the tree kernel and the deep kernel designed for tree models and neural networks) that can also estimate the feature importance, and more details can be found in [18].

3. Evaluation metrics

Section II tailors four XAI techniques to provide interpretability for AI models, but it is still difficult to determine whether the interpretation to AI models can be trusted. Therefore, this section proposes four metrics to evaluate the trustworthiness of interpretations provided by XAI techniques for AI models in wind power forecasting.

(1) The construction of real feature importance across all the samples (i.e., real global interpretability)

As previously mentioned, global interpretability means that the user knows the average feature importance across all samples. In other words, the output of XAI techniques for global interpretability is a vector, representing the average feature importance. Before evaluating the average feature importance provided by the XAI techniques, it is necessary to define the real feature importance. Note that the calculation of real importance to be defined below is time-consuming and tedious, especially when dealing with a large number of features. This is one of the reasons why affordable XAI techniques are needed.

For wind power forecasting, we can define the real feature importance by iteratively removing each feature (i.e., feature ablation). As shown in Fig. 5, we first train an AI model with all features, and obtain the forecast error on the test set. Then, we systematically remove each feature, and retrain the model with the remaining features to measure the forecast error on the test set again. The difference between the two forecast errors indicates the importance of the removed feature. A larger gap indicates higher importance.

(2) The construction of real feature importance for a given specific sample (i.e., real instance interpretability)

Similarly, the output of XAI techniques for instance interpretability is also a vector, representing the feature importance for a specific sample.

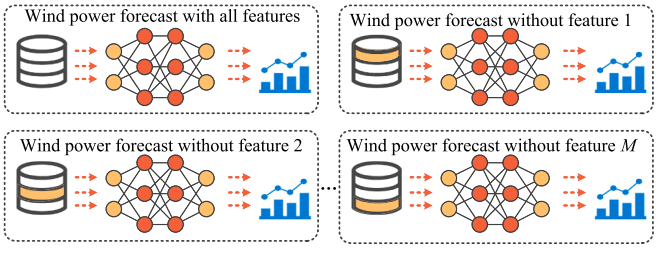


Fig. 5. A visual explanation of feature ablation.

Before evaluating the feature importance provided by the XAI techniques, it is necessary to define the real feature importance for this specific sample.

For wind power forecasting, we can define the real feature importance for a specific sample by introducing small perturbations. As shown in Fig. 6, we feed the specific sample into an AI model to obtain the forecast error. To analyze the importance of feature *i*, we keep other features fixed and inject small perturbations to feature *i*, generating a series of new samples around the original one. Subsequently, we feed these new samples into the AI model to obtain their average forecast errors. The difference between the average forecast error of the new samples and the forecast error of original sample is considered the importance of feature *i* for the specific sample. If feature *i* is unimportant, the error change after introducing small perturbations will be minimal. Conversely, if feature *i* is important, the change in error will be significant.

(3) **The evaluation of the feature importance**

After defining the real feature importance, the normalized mean squared error (NMSE) and normalized mean absolute error (NMAE) can be used to measure the difference between real feature importance and the estimated feature importance provided by the XAI techniques:

$$\mathrm{NMSE} = \frac{1}{M}\sum_{i=1}^{M}\left(\mathrm{FI}_{\mathrm{real},i} - \mathrm{FI}_{\mathrm{xai},i}\right)^2 \tag{12}$$

$$\mathrm{NMAE} = \frac{1}{M}\sum_{i=1}^{M}\left|\mathrm{FI}_{\mathrm{real},i} - \mathrm{FI}_{\mathrm{xai},i}\right| \tag{13}$$

where M is the number of features; $\mathrm{FI}_{\mathrm{real},i}$ is the real importance of feature i; and $\mathrm{FI}_{\mathrm{xai},i}$ is the estimated importance of feature i provided by the XAI techniques.

In addition to evaluating the difference in feature importance, wind power forecasting is also concerned with the ranking of the feature importance. Therefore, we employ Kendall rank correlation (KRC) coefficient and Spearman rank correlation (SRC) coefficient to evaluate the difference between the real and generated rankings [26]:

$$\mathrm{KRC} = \frac{n_c - n_d}{0.5M(M-1)} \tag{14}$$

$$\mathrm{SRC} = 1 - \frac{6\sum_{i=1}^{M} d_i^2}{M\left(M_s^2 - 1\right)} \tag{15}$$

where n_c and n_d are the numbers of concordant and discordant pairs, respectively; and d_i is the difference in ranking for the feature i.

The smaller the NMSE and NMAE, the more trustworthy the interpretability provided by XAI techniques. The larger the KRC and SRC, the more trustworthy the interpretability provided by the XAI techniques.

4. Case study on datasets from GEFCom 2014

A. *Simulation Settings*

To investigate the global interpretability and instance interpretability of the AI models in wind power forecasting, simulations are conducted on a publicly available wind power dataset from GEFCOM 2014 [25]. The time resolution is 1 h, and the look-ahead time is 24 h. The features of the AI models include the forecast values (i.e., NWP data) of wind speed at 10 m (WS10), wind direction at 10 m (WD10), wind speed at 100 m (WS100), and wind direction at 100 m (WS100). The dataset spans from January 2012 to December 2013.

Fig. 6. A visual explanation of how to define the real feature importance for a specific sample.

The first 80 % of the data is used as the training set, the following 10 % is used as the validation set, and the last 10 % is designated as the test set. In addition, we will test the interpretability of the AI models, including neural networks (e.g., MLP in [10], LSTM in [13], and GRU in [14]) and tree models (e.g., LightGBM in [11], XGBoost in [12], and RF in [27]). Their suitable parameters are determined by using the Bayesian optimization in [12], and the specific structures and parameters of each AI model can be found in [15].

The programming language is Python. Machine learning libraries include TensorFlow 2.0 and scikit-learn 1.6. The computer configuration is as follows: Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz, 8GB of RAM.

In the following sections, we will evaluate the effectiveness of XAI techniques from three perspectives: First, we will discuss the global interpretability of XAI techniques in Section B. Next, we will explore their instance interpretability in Section C. Finally, we will compare the time complexity of different XAI techniques in Section D.

B. *Global Interpretability and Evaluation*

Global interpretability means that the user knows the average feature importance across all samples. To observe the global interpretability of XAI techniques for AI models in wind power forecasting, we train several neural networks and tree models, including MLP, LSTM, GRU, RF, LightGBM, and XGBoost. Then, SHAP, PFI, and PDP techniques are used to estimate the average feature importance over all the samples, and the feature ablation technique as mentioned in Section III is adopted to obtain the real feature importance. Finally, Fig. 7 and Fig. 8 present the results.

(1) ***Importance Analyses of Features***

For the wind power dataset from GEFCOM 2014, WS100 is the main factor affecting wind power forecasting, highlighting its pivotal role. Following closely is WD100, which also occupies a significant position, underlining the combined importance of both wind speed and direction at this altitude. However, the influence of WD diminishes when considering it at 10 m. WD10 retains some impact, indicating its secondary role in affecting wind power forecasting. WS10 has the least effect, demonstrating that the wind speed at this altitude has a relatively minor impact on wind power forecasting.

(2) ***Comparison with Existing Methods***

In fact, the linear regression model and Pearson correlation coefficient are widely used in previous publications to estimate the average feature importance over all the samples [28,29]. To show the superiority of the proposed XAI techniques, we compare the proposed XAI techniques with two existing methods (e.g., linear regression model and Pearson correlation coefficient). In particular, the first method is to use the weights of each feature in a linear regression model as the feature importance [28]. The second method is to use the Pearson correlation coefficient between the feature and the output as the feature importance [29]. Fig. 9 presents the average feature importance over all the samples provided by the existing methods.

From Fig. 9, the two most important features identified by existing methods are WS100 and WS10, while WD100 and WD10 are considered unimportant. However, this differs from the real feature importance ranking provided by the feature ablation technique. As shown in Fig. 7 and Fig. 8, the most important feature is WS100, followed by WD100, WD10, and WS10. The feature importance rankings provided by the

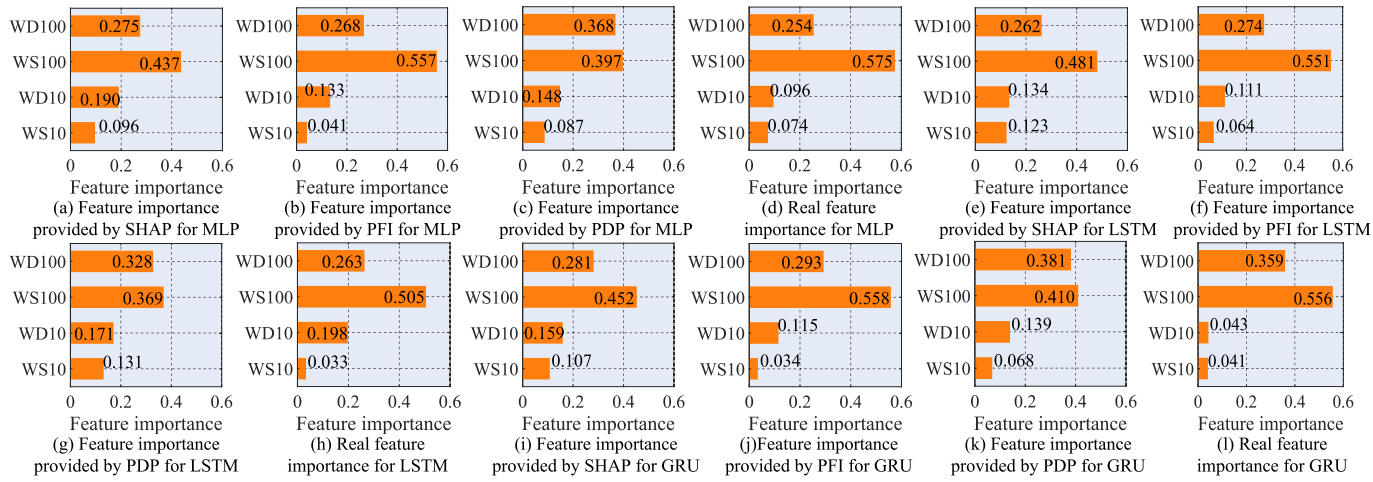


Fig. 7. The global interpretability of XAI techniques for neural networks in wind power forecasting.

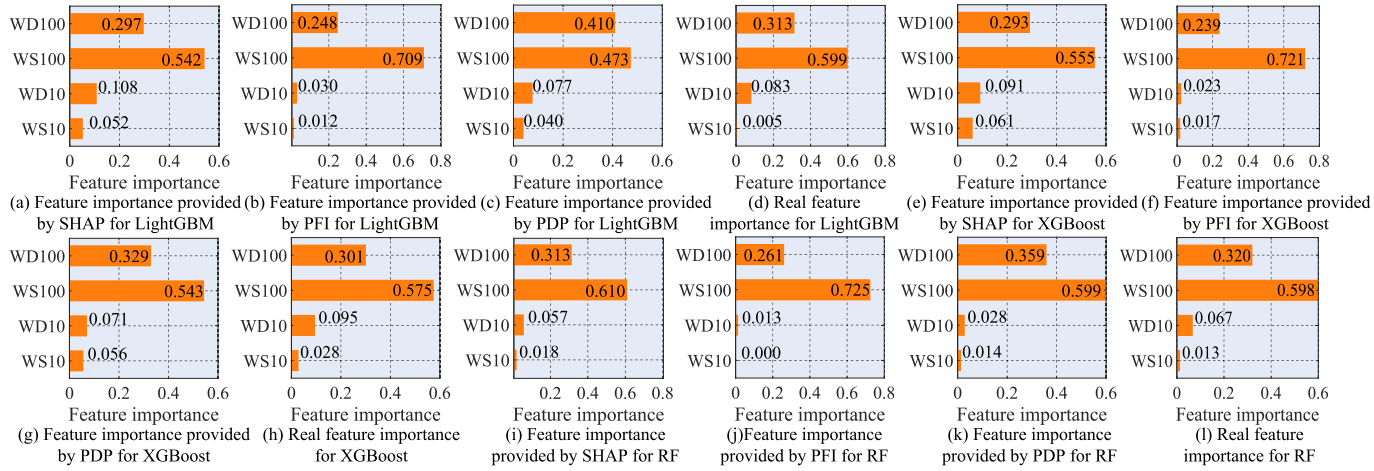


Fig. 8. The global interpretability of XAI techniques for tree models in wind power forecasting.

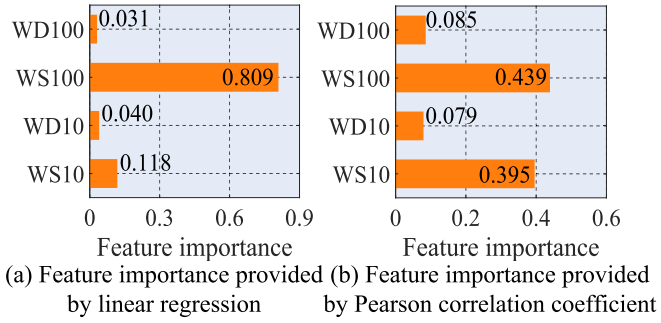


Fig. 9. The feature importance provided by existing methods.

proposed XAI techniques (e.g., SHAP, PFI, and PDP) are consistent with the real feature importance ranking. The feature importance rankings provided by the proposed XAI techniques are closer to the real rankings compared to those provided by existing methods (i.e., linear regression model and Pearson correlation coefficient), indicating that the interpretations from the proposed XAI techniques are more reliable.

(3) ***Trustworthy Analyses***

To quantitatively evaluate the trustworthiness of global interpretability provided by SHAP, PFI, and PDP techniques, we calculate the metrics (i.e., NMAE, NMSE, KRC and SRC) between estimated feature importance and real feature importance, as presented in Tables 1 and 2.

When interpreting neural networks in wind power forecasting, the difference between the real feature importance and the estimated feature importance provided by the PFI technique is the smallest, resulting in the smallest MAE and MSE compared to the SHAP and PDP techniques in estimating feature importance. Conversely, in the case of tree-based models, the SHAP technique closely approximates the real feature importance obtained by feature ablation, followed by PDP and PFI techniques.

The observed discrepancy in the performance of XAI techniques between neural networks and tree-based models may be due to their inherent differences in model structure and decision processes. Neural networks, which are highly nonlinear and complex, may benefit more from the perturbation-based PFI technique, which takes into account the change in model performance when features are shuffled. On the other hand, the higher performance of the SHAP technique in tree-based models may be due to its ability to capture complex interactions and dependencies within decision trees. PDP, which provides a simpler approximation, may face challenges in encapsulating the nuanced relationships within complex models, contributing to its comparatively higher MAE and MSE.

Regardless of whether it is a neural network or a tree-based model,

Table 1
The evaluation metrics of global interpretability provided by XAI techniques for neural networks in wind power forecasting.

Evaluation metrics	Global interpretability for MLP			Global interpretability for LSTM			Global interpretability for GRU		
	SHAP	PFI	PDP	SHAP	PFI	PDP	SHAP	PFI	PDP
NMAE	0.069	0.025	0.089	0.045	0.043	0.081	0.073	0.037	0.091
NMSE	0.007	0.001	0.012	0.003	0.003	0.008	0.008	0.002	0.009
KRC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SRC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2
The evaluation metrics of global interpretability provided by XAI techniques for tree models in wind power forecasting.

Evaluation metrics	Global interpretability for LightGBM			Global interpretability for XGBoost			Global interpretability for RF		
	SHAP	PFI	PDP	SHAP	PFI	PDP	SHAP	PFI	PDP
NMAE	0.033	0.076	0.076	0.011	0.094	0.028	0.010	0.080	0.026
NMSE	0.001	0.008	0.006	0.000	0.010	0.001	0.000	0.008	0.001
KRC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SRC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

the feature importance rankings provided by XAI techniques are entirely consistent with the actual rankings, as shown in Fig. 7 and Fig. 8. Therefore, both the KRC and SRC are equal to 1.

In summary, the PFI technique is more suitable for providing global interpretability for neural networks in wind power forecasting compared to both SHAP and PDP techniques. On the other hand, SHAP technique stands out as the optimal choice for producing global interpretability of tree models.

C. Instance Interpretability and Evaluation

Global interpretability focuses on the model behavior over all the samples, while instance interpretability looks at the behavior of AI models on a specific sample.

To observe the instance interpretability of XAI techniques for AI models in wind power forecasting, a sample is selected from the test set randomly. Then, SHAP and LIME techniques are used to estimate the feature importance of this selected sample, as shown in Fig. 10.

(1) *Importance Analyses of Features*

Regardless of the AI model used, both the SHAP and LIME techniques consistently identify the WS100 as the most important feature for this selected sample. However, there are differences in the importance rankings provided by the SHAP and LIME techniques for the other three features. For example, in terms of LSTM, the instance interpretability provided by SHAP suggests a descending ranking of importance for the features as WS100, WS10, WD100, and WD10, while the instance interpretability provided by LIME presents a descending ranking of importance as WS100, WD100, WD10, and WS10.

(2) *Trustworthy Analyses*

Further, to quantitatively evaluate the trustworthiness of instance interpretability provided by SHAP and LIME techniques, we calculate the metrics (i.e., NMAE, NMSE, KRC and SRC) between estimated feature importance and real feature importance for each sample in test set. The violin and box plots in Fig. 11 and Fig. 12 show probability distributions of four metrics for neural networks and tree models, respectively. Table 3 presents the average metrics.

When interpreting neural networks (i.e., MLP, LSTM, and GRU) in wind power forecasting, the disparity between the real feature importance and the estimated feature importance provided by the SHAP technique is smaller than LIME technique. This results in a smaller MAE and MSE, along with a larger KRC and SRC. For example, after comparing the 2nd and 6th columns of Fig. 11(a), we see that for SHAP, the NMSE is mostly less than 0.2, with a median also less than 0.2. In contrast, for LIME, the NMSE is mostly greater than 0.2, with a median greater than 0.3. This indicates that the explanations for MLP provided by SHAP are closer to reality than those provided by LIME.

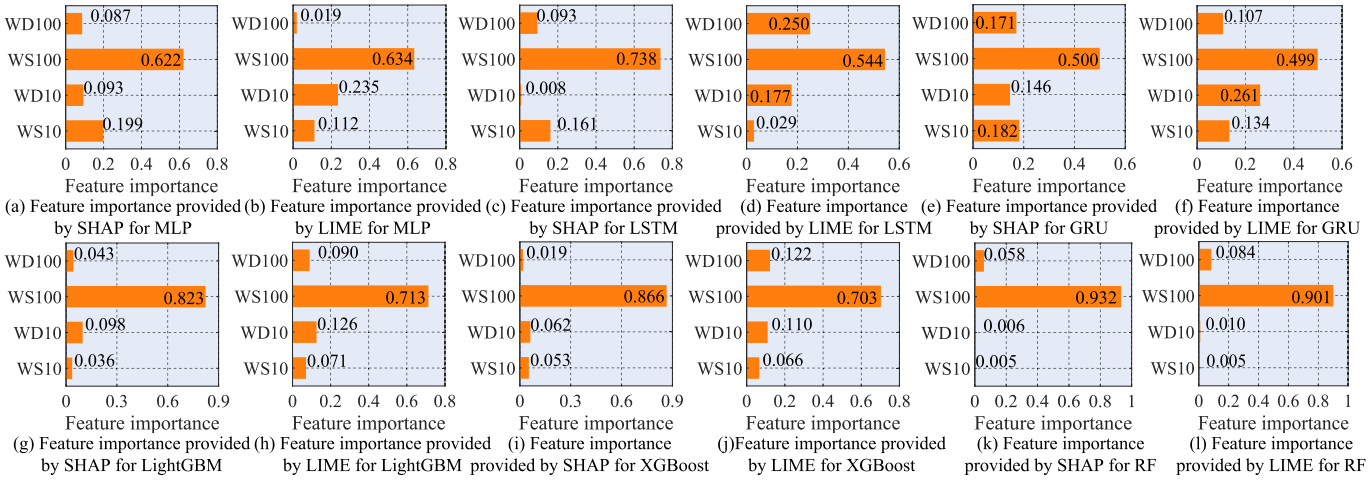


Fig. 10. The instance interpretability of XAI techniques for AI models in wind power forecasting.

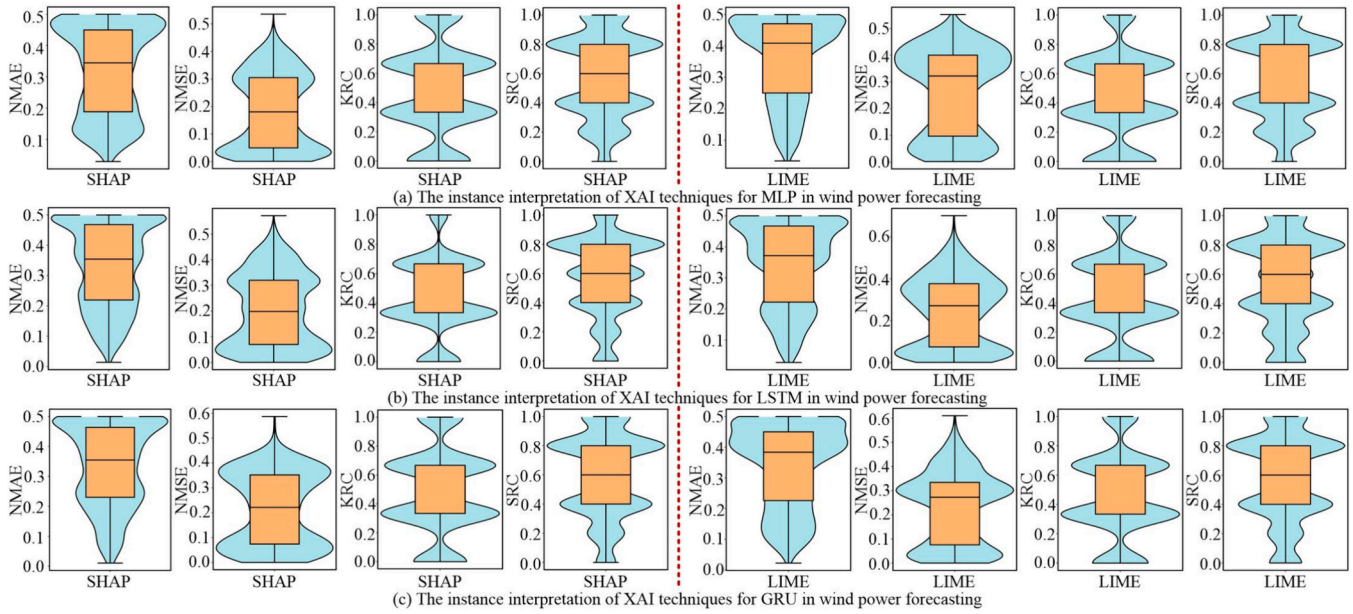


Fig. 11. The distribution of four metrics (i.e., NMAE, NMSE, KRC and SRC) for neural networks in wind power forecasting.

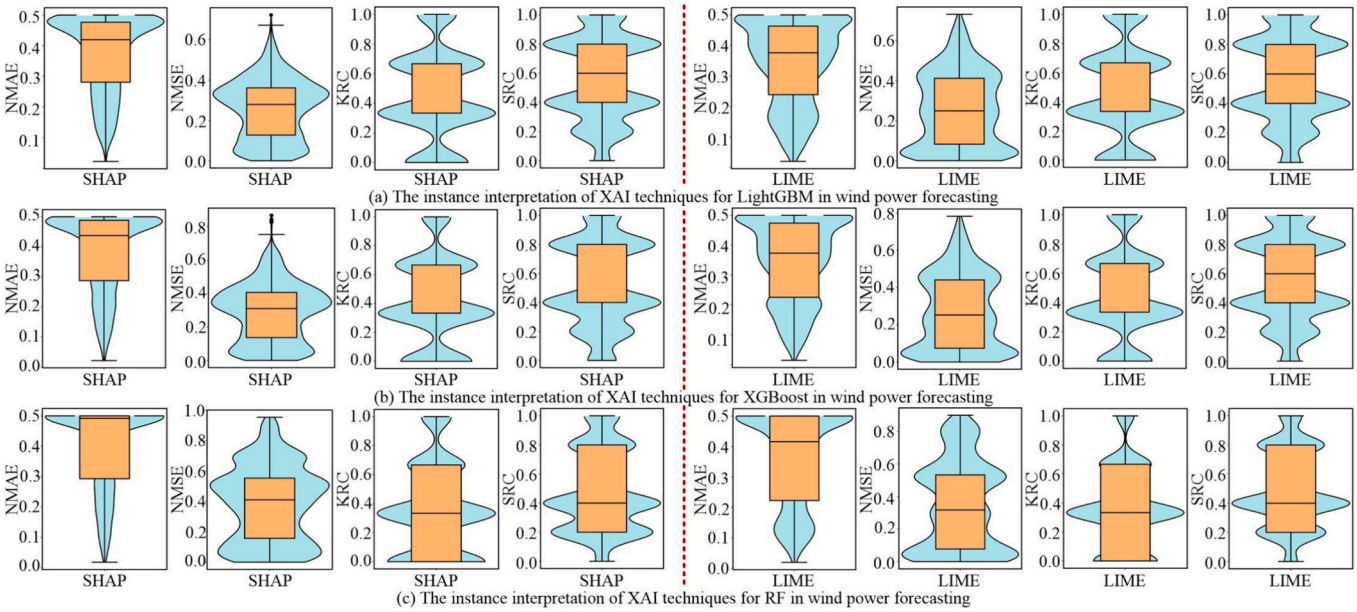


Fig. 12. The distribution of four metrics (i.e., NMAE, NMSE, KRC and SRC) for tree models in wind power forecasting.

Table 3
The evaluation metrics of instance interpretability provided by XAI techniques for AI models in wind power forecasting.

Evaluation metrics	Instance interpretability for MLP		Instance interpretability for LSTM		Instance interpretability for GRU		Instance interpretability for LightGBM		Instance interpretability for XGBoost		Instance interpretability for RF	
	SHAP	LIME	SHAP	LIME	SHAP	LIME	SHAP	LIME	SHAP	LIME	SHAP	LIME
NMAE	0.320	0.359	0.334	0.342	0.335	0.340	0.375	0.344	0.382	0.346	0.396	0.360
NMSE	0.185	0.264	0.203	0.247	0.216	0.230	0.259	0.259	0.287	0.272	0.385	0.346
KRC	0.476	0.435	0.437	0.431	0.485	0.456	0.451	0.459	0.410	0.442	0.355	0.363
SRC	0.585	0.549	0.573	0.546	0.595	0.570	0.563	0.567	0.523	0.558	0.476	0.486

Conversely, in the case of tree-based models (i.e., LightGBM, XGBoost, and RF), the LIME technique closely approximates the real feature importance, followed by SHAP technique. The observed difference in the instance interpretability provided by SHAP and LIME techniques for neural networks and tree-based models is similar to patterns in global interpretability (as seen in section B), which may be

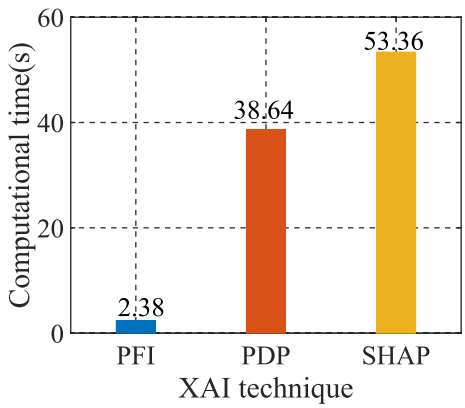


Fig. 13. The computational time of proposed XAI techniques.

due to their inherent differences in model structure and decision processes. For example, after comparing the 1st and 5th columns of Fig. 12 (b), we see that for SHAP, the median of the NMAE is greater than 0.42. In contrast, for LIME, the median of the NMAE is less than 0.38. This indicates that the explanations for XGBoost provided by LIME are closer to reality than those provided by SHAP.

Therefore, when interpreting neural networks, the instance interpretability provided by SHAP technique is more trustworthy, while LIME is better at interpreting tree models for a given sample.

D. *Time Complexity Analysis*

To analyze the computational time of the proposed XAI techniques, we use the RF as a simple example to perform wind power forecasting. Then, we employ the FPI, PDP, and SHAP to estimate the average feature importance. The computational times are shown in Fig. 13.

FPI, PDP and SHAP have computational times of 2.38, 38.64 and 53.36 s respectively. FPI is the fastest and most efficient, PDP strikes a balance between detailed interpretations and computational time, and SHAP provides detailed interpretations at the cost of longer computational time. The choice of technique depends on the specific requirements of the application, whether computational time, detail or depth of interpretation is the priority.

5. Case Study on Datasets From JUVENT

A. *Simulation Settings*

To test whether the XAI technique shows similar patterns in other datasets, simulations are conducted on a real wind power dataset from JUVENT [30]. After data preprocessing, the time resolution is 1 h, and the look-ahead time is 24 h. The features of AI models include the forecast values (i.e., NWP data) of vertical wind shear, wind direction at height level 6, wind speed at height level 6, kinetic energy, Mass fraction of cloud liquid water, and pressure. The dataset spans from January 2017 to December 2020.

The first 80 % of the data is used as the training set, the following 10 % is used as the validation set, and the last 10 % is designated as the test set. In addition, we will test the interpretability of the AI models, including neural networks (e.g., MLP in [10], LSTM in [13], and GRU in [14]) and tree models (e.g., LightGBM in [11], XGBoost in [12], and RF in [27]). Their suitable parameters are determined by using the Bayesian optimization in [12].

6. Results and discussions

To quantitatively evaluate the trustworthiness of global and instance interpretability provided by the XAI techniques, we calculate the metrics (i.e., NMAE, NMSE, KRC and SRC) between estimated feature importance and real feature importance, as presented in Tables 4–6.

In the JUVENT dataset, we can find similar patterns as demonstrated in the wind power dataset from GEFCOM 2014. This indicates that the trustworthiness rankings of the XAI techniques are applicable to other wind power datasets as well. Specifically, for neural networks in wind power forecasting, the descending ranking of trustworthiness for global interpretability is PFI, SHAP, and PDP, while the ranking of instance interpretability is SHAP and LIME. As for tree models in wind power forecasting, the descending ranking of trustworthiness for global interpretability is SHAP, PDP, and PFI, while the ranking of instance interpretability is LIME and SHAP.

7. Conclusion

Most advanced AI models in wind power forecasting are black boxes, which have difficulty in understanding their decision-making mechanisms. To this end, four model-agnostic XAI techniques are reformulated to be tailored to provide interpretability for such models, and several indicators are proposed to evaluate the trustworthiness of interpretations. Simulations on real wind power datasets lead to the following conclusions:

The SHAP, PFI, and PDP techniques can provide global interpretability for AI models, which helps users identify important features in wind power forecasting. In addition, the SHAP and LIME techniques allow users to know the contribution of features to the forecast values for a specific sample. These XAI techniques are applicable to a variety of AI

Table 4
The evaluation metrics of global interpretability provided by XAI techniques for neural networks in the JUVENT dataset.

Evaluation metrics	Global interpretability for MLP			Global interpretability for LSTM			Global interpretability for GRU		
	SHAP	PFI	PDP	SHAP	PFI	PDP	SHAP	PFI	PDP
NMAE	0.091	0.088	0.101	0.080	0.079	0.124	0.139	0.100	0.182
NMSE	0.012	0.011	0.013	0.009	0.009	0.024	0.024	0.012	0.038
KRC	0.600	0.600	0.400	0.200	0.200	0.000	0.400	0.600	0.200
SRC	0.700	0.700	0.600	0.000	0.200	0.100	0.600	0.700	0.200

Table 5
The evaluation metrics of global interpretability provided by XAI techniques for tree models in the JUVENT dataset.

Evaluation metrics	Global interpretability for LightGBM			Global interpretability for XGBoost			Global interpretability for RF		
	SHAP	PFI	PDP	SHAP	PFI	PDP	SHAP	PFI	PDP
NMAE	0.138	0.185	0.147	0.107	0.194	0.127	0.066	0.088	0.082
NMSE	0.027	0.057	0.032	0.022	0.049	0.025	0.009	0.017	0.012
KRC	0.200	0.200	0.200	0.200	0.000	0.200	0.600	0.400	0.600
SRC	0.200	0.200	0.200	0.300	0.100	0.300	0.700	0.600	0.700

Table 6
The evaluation metrics of instance interpretability provided by XAI techniques for AI models in the JUVENT dataset.

Evaluation metrics	Instance interpretability for MLP		Instance interpretability for LSTM		Instance interpretability for GRU		Instance interpretability for LightGBM		Instance interpretability for XGBoost		Instance interpretability for RF	
	SHAP	LIME	SHAP	LIME	SHAP	LIME	SHAP	LIME	SHAP	LIME	SHAP	LIME
NMAE	0.208	0.216	0.200	0.261	0.187	0.268	0.264	0.229	0.264	0.251	0.252	0.245
NMSE	0.094	0.121	0.069	0.112	0.079	0.130	0.130	0.107	0.155	0.142	0.183	0.176
KRC	0.341	0.238	0.277	0.167	0.412	0.343	0.165	0.208	0.166	0.182	0.214	0.232
SRC	0.425	0.311	0.332	0.201	0.503	0.427	0.214	0.262	0.217	0.231	0.274	0.290

models in wind power forecasting, as they are model-agnostic.

The proposed evaluation metrics (i.e., NMSE, NMAE, KRC and SRC) can help users to comprehensively evaluate the trustworthiness of XAI techniques in wind power forecasting.

In terms of global interpretability, the PFI technique is more suitable for neural networks in wind power forecasting compared to both SHAP and PDP techniques, while SHAP stands out as the optimal choice for producing global interpretability of tree models. In terms of instance interpretability, when interpreting neural networks, the instance interpretability provided by the SHAP technique is more trustworthy, while LIME is better at interpreting tree models for a specific sample.

Although the proposed XAI techniques (SHAP, PFI, PDP and LIME) have been shown to be effective in providing interpretability for AI models in wind power forecasting, their computational complexity can be a major drawback. Some XAI techniques (e.g. SHAP) are computationally intensive, especially when applied to large datasets or complex models such as neural networks. Future research should focus on optimizing these techniques to reduce their computational burden without compromising interpretability.

Funding

This work is funded by the Swiss Federal Office of Energy (Grant No. SI/502135–01). Also, this work is carried out in the frame of the "UrbanTwin: An urban digital twin for climate action: Assessing policies and solutions for energy, water and infrastructure" project with the financial support of the ETH-Domain Joint Initiative program in the Strategic Area Energy, Climate and Sustainable Environment.

CRediT authorship contribution statement

Wenlong Liao: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Jiannong Fang:** Conceptualization. **Lin Ye:** Writing – review & editing. **Birgitte Bak-Jensen:** Writing – review & editing. **Zhe Yang:** Writing – review & editing. **Fernando Porte-Agel:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgment

The authors would like to thank Sophie Bosse and BKW FMB Energie AG for providing the Juvent dataset.

References

[1] Krannichfeldt LV, Wang Y, Zufferey T, Hug G. Online ensemble approach for probabilistic wind power forecasting. IEEE Trans Sustain Energy Apr. 2022;13(2):1221–33.
[2] Liao W, Yang Z, Chen X, Li Y. WindGMMN: scenario forecasting for wind power using generative moment matching networks. IEEE Trans Artif Intell Oct. 2022;3(5):843–50.
[3] Wen H, Pinson P, Ma J, Gu J, Jin Z. Continuous and distribution-free probabilistic wind power forecasting: a conditional normalizing flow approach. IEEE Trans Sustain Energy Oct. 2022;13(4):2250–63.
[4] Guo N, Shi K, Li B, Qi L, Wu H, Zhang Z, et al. A physics-inspired neural network model for short-term wind power prediction considering wake effects. Energy Dec. 2022;261:1–10.
[5] Zhang W, Lin Z, Liu X. Short-term offshore wind power forecasting - a hybrid model based on discrete wavelet transform (DWT), seasonal autoregressive integrated moving average (SARIMA), and deep-learning-based long short-term memory (LSTM). Renew Energy Feb. 2022;185:611–28.
[6] Li Y, Bai X, Liu B. Forecasting clean energy generation volume in China with a novel fractional time-delay polynomial discrete grey model. Energ Build Sept. 2022;271:1–13.
[7] Messner JW, Pinson P. Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. Int J Forecast Dec. 2019;35(4):1485–98.
[8] Liao W, Porte-Agel F, Fang J, Bak-Jensen B, Yang Z, Zhang G. Improving the accuracy and interpretability of neural networks for wind power forecasting. arXiv: 2312.15741. Dec. 2023. p. 1–10.
[9] Liao W, Wang S, Bak-Jensen B, Pillai JR, Yang Z, Liu K. Ultra-short-term interval prediction of wind power based on graph neural network and improved bootstrap technique. J Mod Power Syst Clean Energy Jul. 2023;11(4):1100–14.
[10] Inac T, Dokur E, Yuzgec U. A multi-strategy random weighted gray wolf optimizer-based multi-layer perceptron model for short-term wind speed forecasting. Neural Comput & Applic May. 2022;34:14627–57.
[11] Li Y, Wu Z, Su Y. Adaptive short-term wind power forecasting with concept drifts. Renew Energy Nov. 2023;217:1–14.
[12] Xiong X, Guo X, Zeng P, Zou R, Wang X. A short-term wind power forecast method via XGBoost hyper-parameters optimization. Front Energy Res May. 2022;10:1–9.
[13] Liao W, Bak-Jensen B, Pillai J, Yang Z, Liu K. Short-term power prediction for renewable energy using hybrid graph convolutional network and long short-term memory approach. Electr Power Syst Res Oct. 2022;211:1–7.
[14] Quan H, Zhang W, Zhang W, Li Z, Zhou T. An interval prediction approach of wind power based on skip-GRU and block-bootstrap techniques. IEEE Trans Ind Appl Jul. 2023;59(4):4710–9.
[15] Liao W, Porte-Agel F, Fang J, Bak-Jensen B, Ruan G, Yang Z. Explainable modeling for wind power forecasting: A glass-box approach with exceptional accuracy. arXiv: 2310.18629. Oct. 2023. p. 1–8.
[16] Rawal A, McCoy J, Rawat DB, Sadler BM, Amant RS. Recent advances in trustworthy explainable artificial intelligence: status, challenges, and perspectives. IEEE Trans Artif Intell Dec. 2022;3(6):852–66.
[17] Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller K-R. Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE Mar. 2021;109(3):247–78.
[18] Teneggi J, Luster A, Sulam J. Fast hierarchical games for image explanations. IEEE Trans Pattern Anal Mach Intell Apr. 2023;45(4):4494–503.
[19] Huang Q, Yamada M, Tian Y, Singh D, Chang Y. GraphLIME: local interpretable model explanations for graph neural networks. IEEE Trans Knowl Data Eng Jul. 2023;35(7):6968–72.
[20] Kadir MA, Mosavi A, Sonntag D. Evaluation metrics for XAI: A review, taxonomy, and practical applications. In: 27th international conference on intelligent engineering systems (INES), Nairobi, Kenya; Jun. 2023. p. 111–24.
[21] Stodt J, Reich C, Clarke N. A novel metric for XAI evaluation incorporating pixel analysis and distance measurement. In: 35th international conference on tools with artificial intelligence (ICTAI), Atlanta, GA, USA; Nov. 2023. p. 1–9.
[22] Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, et al. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Comput Surv Jan. 2023;55(9):1–33.
[23] Au Q, Herbinger J, Stachl C, Bischl B, Casalicchio G. Grouped feature importance and combined features effect plot. Data Min Knowl Disc Jun. 2022;36:1401–50.
[24] Angelini M, Blasilli G, Lenti S, Santucci G. A visual analytics conceptual framework for Explorable and steerable partial dependence analysis. In: IEEE Trans Vis Comput Graphics, early access; Apr. 2023. p. 1–16.

[25] Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman R. Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond. Int J Forecast Jul. 2016;32(3):896–913.
[26] Li D, Jiang P, Hu C, Yan T. Comparison of local and global sensitivity analysis methods and application to thermal hydraulic phenomena. Prog Nucl Energy Apr. 2023;158:1–11.
[27] Liu D, Sun K. Random forest solar power forecast based on classification optimization. Energy Nov. 2019;187:1–11.
[28] Machlev R, Heistrene L, Perl M, Levy KY, Belikov J, Mannor S, et al. Explainable artificial intelligence (XAI) techniques for energy and power systems: review, challenges and opportunities. Energy and AI Aug. 2022;9:1–13.
[29] Huang H, Jia R, Shi X, Liang J, Dang J. Feature selection and hyper parameters optimization for short-term wind power forecast. Appl Intell Feb. 2021;51:6752–70.
[30] Tabas D, Fang J, Porte-Agel F. Wind energy prediction in highly complex terrain by computational fluid dynamics. Energies Apr. 2019;12(7):1–12.