Investigating Black-box Model for Wind Power Forecasting Using Local Interpretable Model-agnostic Explanations Algorithm

Mao Yang, *Senior Member*, *CSEE*, Chuanyu Xu, Yuying Bai, Miaomiao Ma, and Xin Su

***Abstract*—Wind power forecasting (WPF) is important for safe, stable, and reliable integration of new energy technologies into power systems. Machine learning (ML) algorithms have recently attracted increasing attention in the field of WPF. However, opaque decisions and lack of trustworthiness of black-box models for WPF could cause scheduling risks. This study develops a method for identifying risky models in practical applications and avoiding the risks. First, a local interpretable model-agnostic explanations algorithm is introduced and improved for WPF model analysis. On that basis, a novel index is presented to quantify the level at which neural networks or other black-box models can trust features involved in training. Then, by revealing the operational mechanism for local samples, human interpretability of the black-box model is examined under different accuracies, time horizons, and seasons. This interpretability provides a basis for several technical routes for WPF from the viewpoint of the forecasting model. Moreover, further improvements in accuracy of WPF are explored by evaluating possibilities of using interpretable ML models that use multi-horizons global trust modeling and multi-seasons interpretable feature selection methods. Experimental results from a wind farm in China show that error can be robustly reduced.**

***Index Terms*—Black-box model, correlation analysis, feature trust index, local interpretability, local interpretable model-agnostic explanations (LIME), wind power forecasting.**

I. Introduction

ENSURING renewable energy accounts for a high proportion of future energy consumption is the primary goal of future power system development [1]. Wind power forecasting (WPF) and its application to dispatching and operation have promoted implementation of new energy technologies. WPF can help reduce adverse effects of integration of wind power into power systems, reduce operating cost of power grids, improve operational reliability of power systems, and effectively ensure safety of power grids [2].

Research on WPF is of theoretical and practical significance. Numerous studies have been conducted in this field. Existing WPF techniques can be categorized as follows: (a) Based on input for WPF, WPF methods can be classified as time-series extrapolation and regression of numerical weather prediction (NWP) data [3]. (b) Based on the time scale of WPF, WPF methods can be classified as ultra-short-term WPF, short-term WPF, and mid- and long-term WPF [4]. Ultra-short-term WPF typically corresponds to time-series extrapolation-based modeling method [5]. (c) Based on the WPF modeling technique, WPF methods can be classified as physical and statistical methods [6]. Physical methods typically downscale NWP data and then establish mapping relations between weather features and power based on physical models [7]. However, owing to their lack of historical data usage, physical methods are usually only suitable for modeling newly constructed wind farms (WFs) or supplying long-term forecasts of wind power [8]. Statistical methods estimate wind power by determining the statistical pattern of historical data. Early statistical methods include the persistence method, autoregressive moving average method, and parametric/nonparametric regression [9]. With emergence of high-dimensional information and big data, intelligent learning algorithms have attracted the attention of researchers in the field of WPF owing to their advantages in extracting data features [10].

Since the 1970 s, intelligent systems have attracted sporadic attention. Expert systems [11] were first to attract attention, followed by neural networks [12] a decade later, and recommendation systems post 2000 [13]. However, interpretability of intelligent learning models has not received much attention thus far. Machine learning (ML) and deep learning (DL) have mainly been used to investigate forecasting capabilities and models, but their ability to interpret the decision-making process has not been prioritized. Main modeling tools include artificial neural networks [14], [15] and support vector machines [16], [17]. Static modeling problems are solved through extraction and discovery of hidden input–output relations to realize WPF [18]. Compared to conventional power forecasting models, these black-box models are usually capable of achieving higher forecasting accuracies owing to their exceptional ability to discover nonlinear relations; however, their use is becoming increasingly complex and nontransparent. In particular, it is difficult to believe the decision of the model when the dispatcher does not know why the model got the

Manuscript received October 7, 2021; revised January 4, 2022; accepted February 10, 2022. Date of online publication August 18, 2022, date of current version November 24, 2023. This work was supported by the National Key R&D Program of China (Technology and application of wind power/photovoltaic power prediction for promoting renewable energy consumption) under Grant (2018YFB0904200).

M. Yang, C. Y. Xu (corresponding author, email: 2463584212@qq.com), M. M. Ma, and X. Su are with the Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education, Northeast Electric Power University, Jilin 132012, China.

Y. Y. Bai is with Daqing Power Industry Bureau, Daqing 163000, China.

DOI: 10.17775/CSEEJPES.2021.07470

2096-0042 © 2021 CSEE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

result.

The Dutch Artificial Intelligence Manifesto (2018) focuses on explainable artificial intelligence (AI) and requires equal importance to be given to accuracy and interpretability of AI [19]. In a report on responsible AI and national AI strategies published in July 2018, the European Commission described the risks of opaqueness (i.e., black-box risk) and explainability as the two performance-related risks of AI [20]. Technical explainability means that humans can understand and trace a decision made by an AI system. Karatekin *et al*. [21] discussed the tradeoffs between accuracy and interpretability of ML techniques in clinical data. Neonatal experts' intuition was confirmed on several risk factors, such as gender, which were previously considered to be clinically meaningless in predicting retinopathy of prematurity. Wu *et al*. [22] used a regularized decision tree to simulate the prediction of a time series model. The tree model is more self-explanatory than the DL. However, a tree model may not be available for all scenarios. Farhood *et al*. [23] used a local interpretable model-agnostic explanations (LIME) algorithm to analyze key features of an AI system in crime scene decision-making. This study found the recognition results of DL may be destroyed, and interpretable results could find the main factors that cause misrecognition. Kamal *et al*. [24] explained how genes participate in prediction and which genes are particularly responsible for Alzheimer's disease using LIME for combined models of the convolution neural networks support vector classifier and XBoost. Moreover, the explanation provided by LIME was found highly consistent with that provided by doctors in clinical medical judgment [25].

Establishing forecasting models using ML algorithms from the perspective of massive data is a focus area of research in the field of WPF. The primary goal of a forecasting model is to provide wind speed (WS) and power forecasts. Relevant research has focused on discovering the application potential of various types of intelligent learning algorithms for WPF modeling. Without altering the forecasting model, existing technical routes typically involve (a) selection or preprocessing of input features [26]–[28], (b) refined modeling [29]–[31], and (c) secondary modeling for forecasting errors [32]. However, owing to difficulty in understanding and interpreting the internal logical structure of black-box models, reducing forecasting error is usually the sole basis for validating these technical routes. In particular, it is difficult for researchers to explain inferential mechanisms of black-box forecasting models, which limits their application in the power industry. The main contributions in this paper are as follows:

1) LIME algorithm is introduced and improved for WPF model analysis, considering multiple features and fuzzy mapping. Then, a category trust index (CTI) is presented to quantify the level at which a black-box model, that directly participates in forecasting, can trust features involved in training.

2) A risky model for practical applications is defined. The operational principle of black-box models in WPF is explained at multiple scales, and reasonableness and applicability of the technical routes that directly participate in forecasting and preprocessing are evaluated. Route (a), selection or preprocessing of the input features, and Route (b), refined modeling, are analyzed from the perspective of model operation.

3) A multi-horizons global trust modeling (GTM) is proposed, which is based on an interpretable analysis of four modeling methods in different time horizons and is verified in three predictors.

4) An interpretable feature selection mode is proposed. The features are further selected based on the results of Wrapper, Embedded, and Filter methods, which can reduce interference and redundancy of insignificant variables.

5) Prospects and challenges of interpretable analysis in WPF are listed.

The growing importance of interpretable ML is shown by the increasing number of studies and projects on this topic, as illustrated in Fig. 1. This figure shows sporadic research since 1993, and a sharp increase in the number of interpretable machine learning-related works annually since 2017. However, the ML model widely used in WPF does not incorporate interpretable analysis, which can lead to a black-box risk for an actual operational system. To the best of our knowledge, this is the first study to apply the LIME algorithm to WPF.

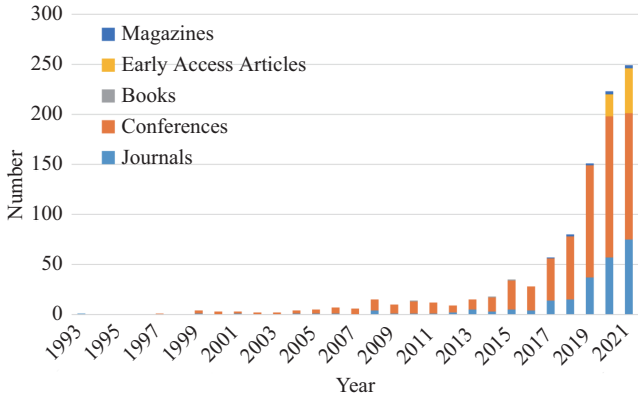

Fig. 1. Number of works that interpretable machine learning-searched word depicted yearly in the IEEE Xplore (up to December 2021).

The remainder of this paper is organized as follows: Section II analyzes time-series and weather-feature correlations in WPF and presents a new trust index based on a local interpretable ML model in conjunction with WPF characteristics. Section III analyzes the operational mechanism of an agnostic model for local sample points at multiple scales and describes how the original WPF model is improved based on interpretable analysis. Finally, Section IV concludes the paper.

II. Interpretable ML for WPF

A. Dataset

A WF in China was selected for investigation. The WF, consisting of 267 wind turbines, each with a capacity of 1.5 MW, has a total installed capacity of 400.5 MW and encompasses an area of approximately 100 km^2. Data used in this study (i.e., the total power generated by the entire WF and corresponding NWP data) were collected at intervals of 15 min between 08:15 on January 1 and 23:45 on December 31, 2018.

B. Analysis of Input Features for WPF

Researchers usually have prior knowledge about the application domain, which they can use to accept (trust) or reject forecasting if they understand the reasoning behind it. A WPF model is usually developed based on stationarity of time-series data or mapping relation between weather information and power. Hence, herein, models with historical power data and NWP data as input were primarily analyzed. Correlation characteristics of each feature were analyzed as prior knowledge using the following correlation measures.

The Pearson correlation coefficient (CC) (linear), gray relational analysis (GRA) (nonlinear), maximum mutual information (MI) coefficient (MIC) [33] (nonlinear) are employed to measure magnitude of the correlation between vectors. Formulas are as follows

$$\rho_i = \frac{\operatorname{cov}(X_0, X_i)}{\sqrt{DX_0 \times DX_i}} \tag{1}$$

$$r_i = \frac{1}{m}\sum_{k=1}^{m} \xi_i(k) \tag{2}$$

$$\xi_i(k) = \frac{\min\limits_i \min\limits_k |X_0(k) - X_i(k)| + \lambda \times \max\limits_i \max\limits_k |X_0(k) - X_i(k)|}{|X_0(k) - X_i(k)| + \lambda \times \max\limits_i \max\limits_k |X_0(k) - X_i(k)|} \tag{3}$$

$$\operatorname{mic}(x_i, x_0) = \max_{a \times b < B} \frac{I(x_i, x_0)}{\log_2 \min(a, b)} \tag{4}$$

$$I(X_i, X_0) = H(X_0) - H(X_0|X_i) \tag{5}$$

$$H(X_0) = -\sum_{x_0 \in X_0} P_{X_0}(x_0) \log P_{X_0}(x_0) \tag{6}$$

$$H(X_0|X_i) = -\sum_{x_i \in X_i} P_{X_i}(x_i) \cdot \left(\sum_{x_0 \in X_0} P_{X_0|X_i}(X_0|X_i) \log P_{X_0|X_i}(X_0|X_i) \right) \tag{7}$$

where cov is the covariance, D is the variance, ρ_i is the CC between the i^{th} feature and the reference series, X_0 is the target series, and X_i is the i^{th} feature. GRA of the i^{th} feature for comparison is expressed as r_i. $k = 1, 2, \cdots, m$, $x_0(k)$ is the k^{th} sample in the feature series, and λ is the distinguishing coefficient ($\lambda \in (0, 1)$). Small λ value implies a large difference between the CCs and hence high distinguishability. In this study, λ was set to 0.5. where a and b are the numbers of mesh cells in the x and y directions, respectively. $P(x_0)$ is the probability that $x_0 = X_0$.

1) Importance of feature time series

Here, correlation of the wind power time series is analyzed. Correlation of a seven-day historical power time series is calculated using CC, GRA, and MIC methods. Fig. 2 shows the results. Correlation of a power series decreases as time horizon increases and gradually becomes stable after a certain time horizon is reached. However, based on values of the three correlation measures, a power time series can be considered to exhibit a certain daily pattern. Specifically, the three correlation measures reach their maximum values at intervals of approximately 96 time points (1 d). In addition, the effect of this daily pattern exceeds that of some of the early time points. Moreover, despite exhibiting a stable trend after approximately 200 time points, MIC similarly changes periodically within smaller intervals.

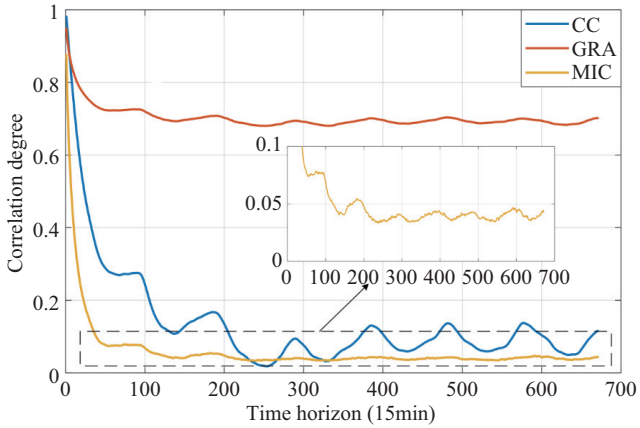

Fig. 2. Characteristic correlation degree of power series.

2) Importance of NWP features

Here, correlations between NWP features and power are calculated. Considering that distribution pattern varies from feature to feature and use of neural networks in WPF generally requires normalization of the mean and variance of the dataset, each feature is normalized based on its mean and variance to maintain compatibility between analysis results, as follows:

$$\lambda^* = \frac{\lambda - \mu(\lambda)}{\sigma(\lambda)} \tag{8}$$

where λ is the vector to be normalized, and $\mu(\lambda)$ and $\sigma(\lambda)$ are the mean and variance of λ, respectively.

Figure 3 shows analysis results obtained using the three correlation measures. Table I summarizes weather features included in Fig. 3 and their corresponding names. We can draw the following conclusions:

TABLE I
NWP NAME AND FEATURE MEANING

Name	Characteristic meaning	Name	Characteristic meaning	Name	Characteristic meaning	Name	Characteristic meaning
T	Temperature	WSS	Sea Level wind speed	SLP	Sea Level Pressure	SP	Surface Pressure
MF	Momentum Flux	WD170	170 m Wind Direction	FC	Fraction of Cloud	TP	Total Precipitation
WS170	170 m Wind Speed	WD100	100 m Wind Direction	LHF	Latent Heat Flux	LSP	Large-Scale Precipitation
WS100	100 m Wind Speed	WD30	30 m Wind Direction	SHF	Sensible Heat Flux	CP	Convective Precipitation
WS30	30 m Wind Speed	WD10	10 m Wind Direction	SWR	Short Wave Radiation	T2	2 m Temperature
WS10	10 m Wind Speed	WDS	Sea Level Wind Direction	LWR	Long Wave Radiation	RH2	2 m Relative Humidity

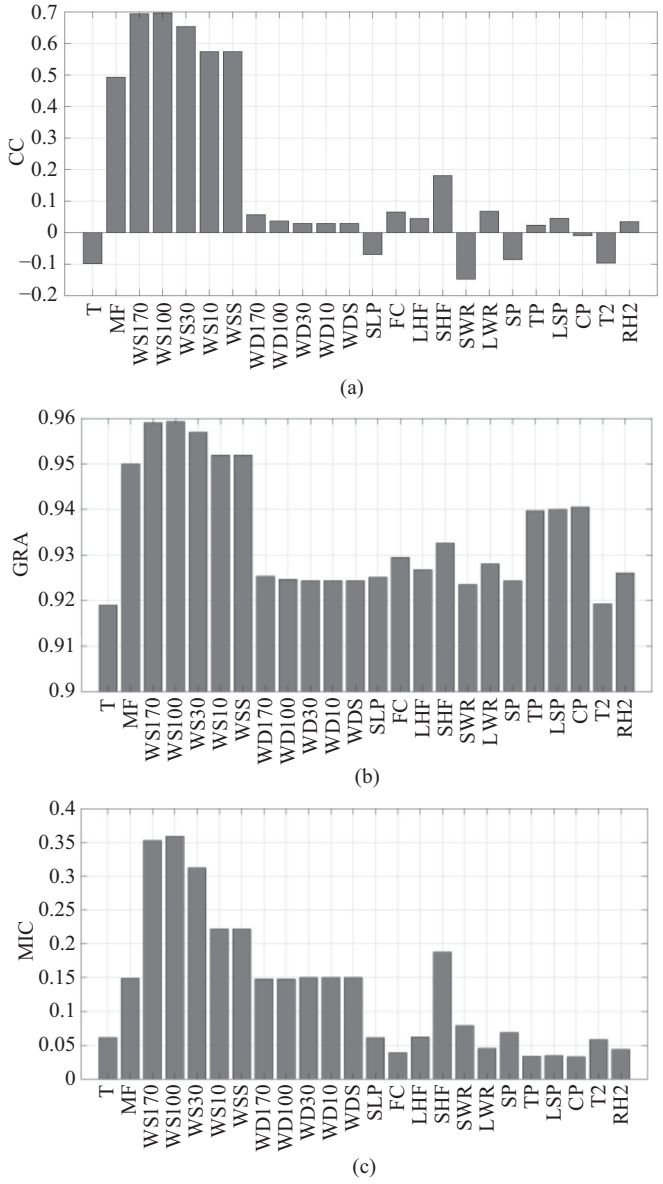

Fig. 3. Correlation degree of each feature of NWP. (a) CC. (b) GRA. (c) MIC.

a) As expected, WS features are the most important features, of which WS100 has highest correlation with power, followed by WS170, WS30, WS10, and WSSL.

b) Rankings based on different measures differ in several aspects, which represent that any pre-analysis index is not comprehensive, of course, we do not deny their ability to extract main factors. For example, based on the GRA values, accuracies for TP, LSP, and CP are relatively high. However, this is attributed to the fact that GRA represents the degree of synchronization of changes (i.e., changes in the same direction). In addition, based on CC values, wind-direction (WD) features at five heights are relatively weakly correlated with power, which contradicts the conclusion derived from GRA and MIC values, suggesting that nonlinear relations exist between WD features and target power, which cannot be extracted by CC. Furthermore, precondition for CC is that a feature is consistent with a Gaussian distribution or shape. However, WD data usually follow multimodal distribution [34].

These measures can be used to preliminarily evaluate correlation between each input feature for WPF and target power series. However, whether each feature is mined according to pre-analysis results during the model learning process cannot be determined from these results. These pre-analysis results are used as the physical basis for black-box model analysis [35], [36]. In the following, we use interpretable ML models to analyze influence of input features of WPF on the model training process and further elucidate which models should be trusted in WPF.

C. Interpretable ML Models

In data mining and ML settings, interpretability is defined as the ability to be explained or presented in a manner that is humanly-understandable [37]. In ML tasks, models are typically established based on a set of statistical rules and assumptions. Hence, interpretability is important, as it aims to enable humans to understand how an ML model learns and why it makes a given decision for each input. Meanwhile, researchers address problems from different perspectives; therefore, they assign different meanings to "interpretability" and formulate interpretability methods that focus on different areas.

For a real-world learning task, it is possible to either select and train a simple-structured easily-interpretable model or train a complex powerful model and subsequently develop interpretability techniques for interpretation. Accordingly, the interpretability of ML models can be generally categorized as ante-hoc interpretability and post-hoc interpretability [38]. Ante-hoc interpretability refers to self-interpretability of a self-explanatory model obtained by training a model with a simple structure and good interpretability, or by incorporating interpretability into a specific model structure. Post-hoc interpretability refers to interpretability trained ML models achieved by developing interpretability techniques.

However, simple models established based on ante-hoc interpretability usually have relatively low performance [39]. Self-interpretation is invalid when a predictor changes. An explainer should be able to explain any model, and thus be model-agnostic, that is, treat the original model as a black box.

Based on the purpose and object of interpretation, post-hoc interpretability can be further classified as global interpretability and local interpretability.

The objective of global interpretability is to enable humans to understand the overall logic underlying a complex model and its internal operational mechanism. Representative techniques used to achieve global interpretability include partial dependence test, individual conditional expectation plot, and sensitivity analysis (SA) [40]. However, there are some limitations, such as inaccuracy, poor data continuity, or inability to explain the relationship between features. In particular, current global interpretability is disadvantageous for WPF because it is difficult to achieve in the case of numerous relevant input variables. This contradicts the trend of high-dimensional input features for WPF [41]–[43]. In addition, global interpretability cannot explain results for individual cases.

The objective of local interpretability is to enable humans to understand the decision-making process and basis of an ML model for each input sample [44]. In contrast to global interpretability, local interpretability of a model is oriented toward the input sample and is typically achieved by analyzing contributions of each feature of the input sample to the final decision of the model. Local interpretability can help explain individual forecasts. LIME [45], a method that need not be adapted to the original model, is highly versatile and can be adopted to quantify the contribution of each feature at local sampling points.

LIME aims to explain the intrinsic logic by which a certain forecast is obtained and identify factors that support and oppose this forecast. LIME is compatible with any supervised learning algorithm and implemented by establishing a simple, highly transparent substitution model near an individual sample to explain local linear relations. In this study, a linear regression model is selected as the surrogate model for local samples. Mathematical explanation of LIME is given below:

$$e(x) = \arg\min_{g \in G} [L(f, g, \pi_x) + \Omega(g)] \tag{9}$$

$$\omega = \exp\left(-\frac{(x - x')^2}{2k^2}\right) \tag{10}$$

where f is a complex model that directly participates in WPF, g is the surrogate model for f, $2k^2$ is set to 1 in this study. The term π_x is the locality of the individual sample x to be explained, i.e., distance between x and any other sample point x' in the neighborhood. $\Omega(g)$ is complexity of the surrogate model, and L-function describes the process of g approximating f in the local definition of x, where $\Omega(g)$ must be sufficiently low to be understandable by humans.

Figure 4 shows the construction process of LIME in WPF.

1) Divide dataset into an input feature set X and a target power set P, and input X into the black-box model to train a neural network function f. As shown below, input X contains N features.

$$P = f(X) \tag{11}$$

$$X = \{x_1, x_2, x_3 \cdots x_N\}^{\mathrm{T}} \tag{12}$$

2) Select an individual sample x to be explained and apply a small disturbance to it in high-dimensional space to produce a post-disturbance feature x' set, X' (a three-dimensional space is used as an example in Fig. 4, as the largest number of dimensions that can be currently displayed is three). As shown below, ε is generally a Gaussian distribution.

$$X' = x + \varepsilon \tag{13}$$

3) Calculate the distance ω between each sample in X' and x using (10).

4) Input X' into the black-box model to produce result Y' on the disturbed feature set as follows:

$$Y' = f(X') \tag{14}$$

5) Obtain an interpretable, simple model g via training on new datasets X' and Y' with ω as the weight. Although Y' is the output from the original black-box model, dataset X' is obtained by adding a small disturbance to a single sample, and the mapping relationship, which of the sample set formed by the original sample and its neighborhood, can be captured through a simple model. We use the locally weighted square loss as K as defined in (16).

$$Y' = g(X', \omega) \tag{15}$$

$$\mathrm{K}(f, g, \pi_x) = \arg\min \left[\sum_{x' \in X} \omega(f(x') - g(x'))\right] \tag{16}$$

6) Explain the complex model near a certain point using the simple linear model.

D. Category Trust Index

LIME algorithm is used to discuss local interpretability of a forecasting model. The disturbance added to the feature set is generally small, random, and follows a Gaussian distribution. However, in contrast to interpretability of image recognition or language models [46], weather features and historical power are typically used as input in WPF modeling. Features are distributed in different ways [34]. For example, WD usually follows a multimodal distribution, while WS is generally considered to be consistent with Weibull distribution. Adding a small disturbance consistent with a Gaussian distribution to all the features may add unlikely sample points to the disturbed dataset and cause each feature to follow a Gaussian distribution, which is inconsistent with characteristics of the

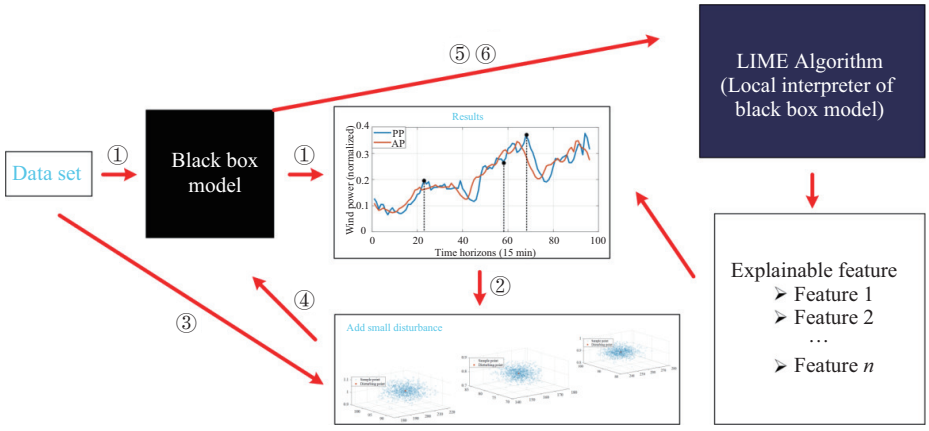

Fig. 4. Correlation degree of each feature of NWP.

original dataset. Hence, we redefine the small disturbance added to the local sample to be explained as follows:

$$x_i' = x_i + \varepsilon_i \tag{17}$$

where x_i or x_i' is the i^{th} input feature ($x_i \in (\delta_{i\min}, \delta_{i\max})$), and ε_i is distributed in a similar way as x_i ($\varepsilon_i \in (\partial_{i\min}, \partial_{i\max})$). $(\partial_{i\min}, \partial_{i\max}) = l \times (\delta_{i\min}, \delta_{i\max})$, where l is set to 0.1 in this study. In this study, for each feature regardless of whether distribution is known or not, a nonparametric probability modeling-sampling method [47]–[50] is used to generate a small similar distribution disturbance sample set. Specifically, an empirical distribution model is used to fit probability density distribution, and samples are obtained by Monte Carlo sampling in the corresponding cumulative empirical distribution probability. Detailed process refers to [47].

Further, in contrast to image recognition or language models, there are no "recognition errors" in WPF. Model performance is quantified based on forecasting error or accuracy. In addition, when months or seasons are considered duration for examination, numerous time points are to be evaluated. Analysis of model interpretability at only a certain time point is inconclusive, as it may be affected by extreme weather events, artificial interference in power output, or inherent data features. Here, we define a CTI η to analyze interpretability of input features of a certain specific type, as shown in (17). Even if an identical model is in the same category, accuracy or fitting ability is different from sample to sample. This is due to mapping fuzziness or data quality. Therefore, a harmonic coefficient σ is introduced as the weight to make results more inclined toward high-performance samples. Classification of categories is explained in detail in the following section. The surrogate model can be any transparent model. To simplify calculation, the ridge model is selected as the surrogate model in this paper. In this study, ridge regression is used to establish an interpretable, simple model as shown in (19). The coefficient λ_i of each variable represents the extent of impact of the corresponding parameter on the forecast.

$$\eta_i = \frac{1}{k}\sum_{j=1}^{k}(\lambda_{ij} \times \sigma_j) \tag{18}$$

$$\sigma_j = 1 - \frac{(y_j - \hat{y}_j)^2}{(y_j - \bar{y})^2} \tag{19}$$

$$Y = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_N x_N \tag{20}$$

where η_i is the level at which the model trusts the i^{th} feature in this category, K is number of samples within the category, σ_j is fitting ability of the model at the j^{th} sample point (to ensure that $\sigma_j \in [0, 1]$, a sufficiently large coefficient must be added to the denominator), y_j is actual power at j^{th} time point in the category, $\hat{y}_j$ is the forecasted power at the j^{th} time point in the category, and $\bar{y}$ is mean actual power at all time points in the category.

III. Interpretability Analysis under Different Category Classifications

A. Equations and Parameter Setting

In this study, the normalized root-mean-square error (N_{RMSE}), qualified rate (Q_{R}), and normalized mean absolute error (N_{MAE}) are used as evaluation metrics for forecasting.

$$N_{\text{RMSE}} = \frac{1}{Cap}\sqrt{\frac{1}{k}\sum_{j=1}^{k}(y_j - \hat{y}_j)^2} \tag{21}$$

$$Q_{\text{R}} = \frac{1}{k}\sum_{j=1}^{k} B_j \times 100\%,$$

$$B_j = \begin{cases} 0, & \text{otherwise} \\ 1, & \text{if } \left(1 - \frac{|y_j - \hat{y}_j|}{Cap}\right) \times 100\% \geq 75\% \end{cases} \tag{22}$$

$$N_{\text{MAE}} = \frac{1}{k}\sum_{j=1}^{k}\frac{|y_j - \hat{y}_j|}{Cap} \tag{23}$$

where the three indexes are expected to approach 1 at higher accuracies.

Gated recurrent unit (GRU) is used as the basic forecasting model. Originally introduced by Cho *et al.* [51], GRU architecture is simpler and reduces computational load and training time while ensuring high forecasting accuracy. Here, GRU is trained using a backward error propagation algorithm and gradient descent method. As this study does not focus on design of neural networks, the number of hidden layer neurons is set to 80 based on experience.

Different disturbed sample sizes will affect stability of results. As an example, a test model was built, and the most trusted features of the model under different disturbance sample sizes are shown in Fig. 5. The result is stable when the number of disturbed samples is greater than 130. In this study, disturbance sample size is set at 200. Under different sizes, running time of the model is 14 to 16 s, which meets the time limit of practical application.

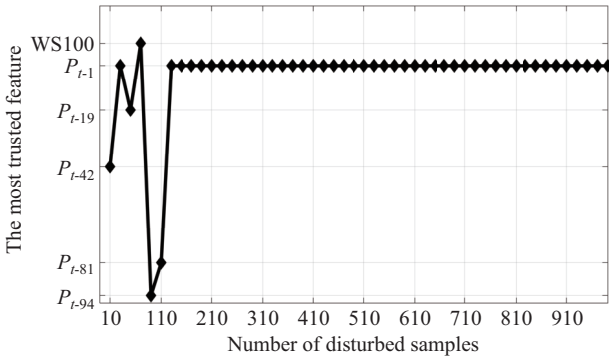

Fig. 5. Most trusted features of the model under different disturbance sample sizes.

In addition, SA, a method used to study feature sensitivity of model, is adopted to verify results. SA examines the importance of features to the forecast by altering their values individually. Altering the value of a certain feature that is relatively important to the forecast increases forecasting error. Feature importance is established through the following procedure:

1) Establish the black-box model given in (11).

2) Assign a random value to a feature x_i, produce a forecast again while keeping model unchanged, and determine change in the forecasted value.

3) Restore the feature to which a random value is assigned in Step 2 and repeat Step 2 with another feature until all the features are traversed.

Although it is impossible to directly analyze how input features compose results by SA in regression problems as verification of linear models approaching local samples of nonlinear relations.

B. Scenario 1: High Forecasting Error-causing Model Interpretability

In this scenario, we present simulated experiments to evaluate utility of interpretability in trust-related tasks. In particular, we address two questions: (1) Is interpretability faithful to the model, and (2) Can interpretability aid users in ascertaining trust in WPF.

Category design: each calendar day (96 time points) is regarded as a category. For example, January and February 2018 are considered here. Data for the first 30 days are used to form the training set, while data for the remaining days are used to form the test set. Training model is updated for each day. The number of samples in the training set remains unchanged (a total of 30 models are established, corresponding to 30 categories). In addition, 24 NWP weather features are used as input to forecast power for the next 24 h.

Figure 6 shows daily N_{RMSE} values on the test set. N_{RMSE} fluctuates relatively significantly from 6.96% to 48.72%. Forecasting points are arranged in ascending order by N_{RMSE} values. Fig. 7 shows the top 10 key variables that affect 1^{st}, 6^{th}, 11^{th}, 16^{th}, 21^{st}, and 26^{th} days (by ranking of the N_{RMSE} values) and their levels of impact (positive and negative η values of a feature are indicated in red and blue, respectively). Forecasting errors for these six days are 6.96%, 12.71%, 16.04%, 24.27%, 26.75%, and 29.93%, respectively. Results produced by models are subjected to SA. All 24 NWP weather features are input to the model. SA order of features corresponds to ranking in Fig. 7. Fig. 8 shows results of NWP-feature analysis corresponding to categories shown in Fig. 7. Fig. 8(a) and (b) correspond to SA results before and after improvement of LIME, respectively. Although the two subgraphs show an overall downward trend, results obtained by the improved method are more regular than original LIME. In

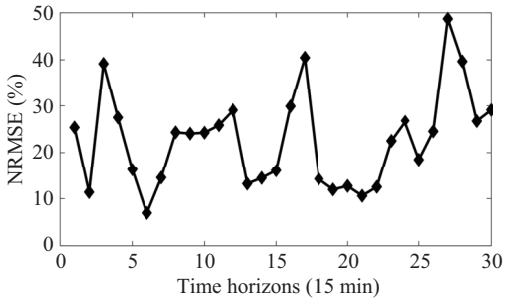

Fig. 6. Daily NRMSE in test set.

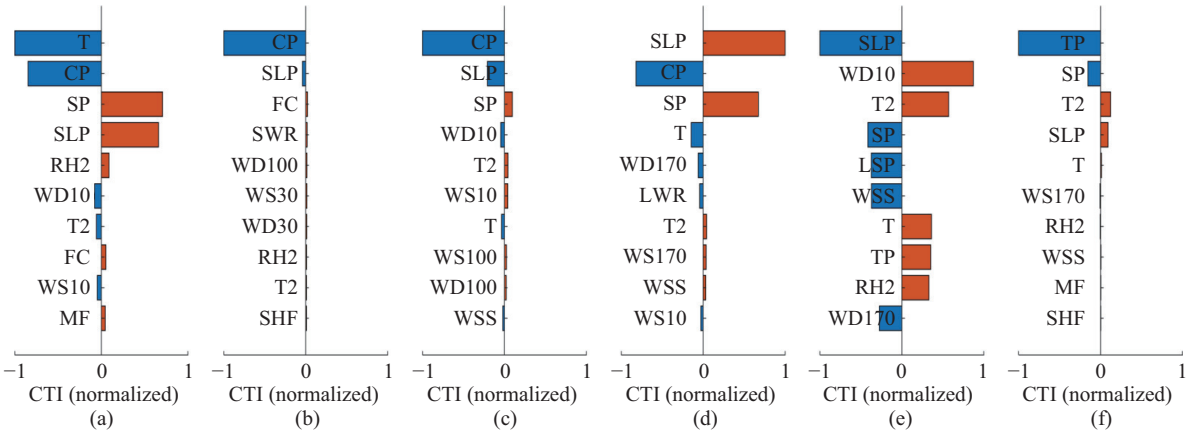

Fig. 7. CTI of the 1^{st}, 6^{th}, 11^{th}, 16^{th}, 21^{st}, and 26^{th} day. (a) 1^{st}. (b) 6^{th}. (c) 11^{th}. (d) 16^{th}. (e) 21^{st}. (f) 26^{th}.

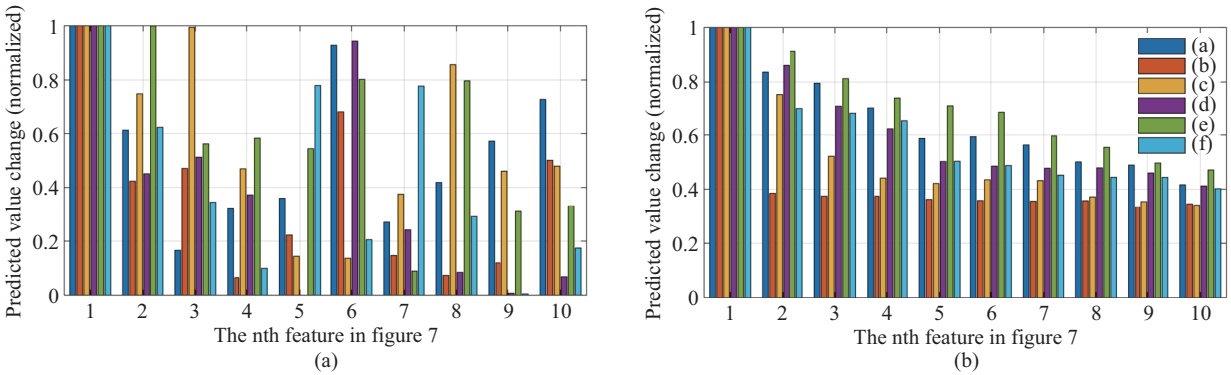

Fig. 8. Results of SA. (The Abscissa of each cluster corresponds to the CTI ranking of each subgraph in Fig. 7, and each color cluster corresponds to one subgraph.). (a) Initial LIME. (b) Improved LIME.

addition, forecasted value for each category shows a downward trend in Fig. 8(b), suggesting the interpretation of the model is correct.

However, CP, SLP, SP, and T2 are the primary meteorological factors affecting categories for which forecasting accuracy is relatively high, as well as those for which are relatively low. This is inconsistent with results of the weather-feature analysis presented in Section II-B.

Moreover, Fig. 9 shows statistics of the top five most trusted features in the 30 categories. Features accounting for the largest proportions of the pie chart are also inconsistent with those derived from previous analysis. Why are these weather features more important than features such as WS from a modeling perspective? Fig. 10 compares the time series of the top six features in Fig. 9 for the period from January 21 to January 30 with the power and WS time series for the same period. Compared to WS time series, the coupled variation between the time series of each of the six features and the power time series shows no significant trend. However, these features are maintained at certain fixed values for a longer period. In other words, because CP remains stable and the gradient descent training method for neural networks determines the forecasted value is unlikely to increase or decrease sharply, stationary data are given greater weights in model training, i.e., the model is more inclined to "believe" relatively stationary data, resulting in a smaller "model penalty". Based on the relatively transparent ML model, we understand why overall forecasting error is relatively large. Part of the error is caused by failure of the model to discover suitable features. Further, the "most important feature" recognized by a machine model is not necessarily most important in terms of physical meaning.

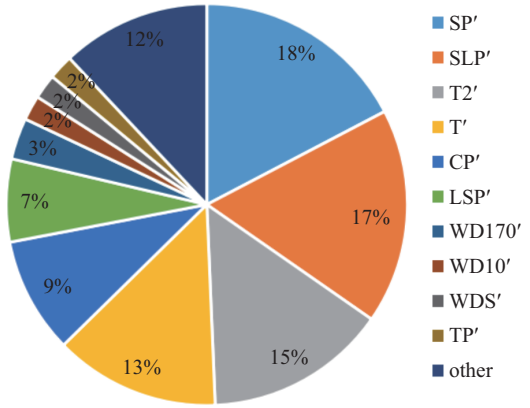

Fig. 9. Statistics of the top five features for CTI.

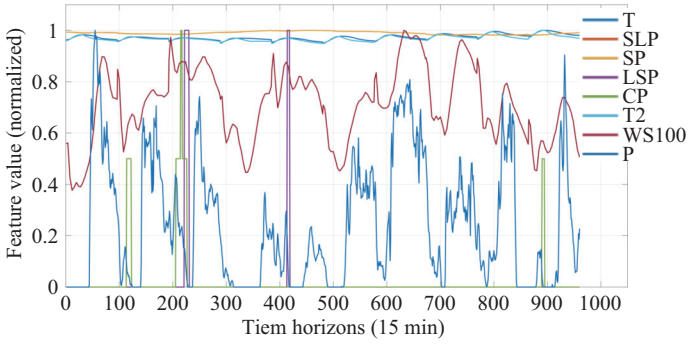

Fig. 10. Time-series curves of several features (normalized by maximum value of each feature).

A model should be considered risky when data mining results completely contradict physical mechanism, regardless of whether error of local samples is high or low. Therefore, it is necessary to analyze interpretability of ML in applications. This provides us with model evaluation information in addition to existing statistical indicators and maybe more meaningful for dispatchers. Meanwhile, such risky models should be improved according to their interpretable results. Two interpretable-based methods are proposed, which are the combined modeling method for multi-time scale (Section III (C)) and interpretable feature selection method (Section III (D)).

C. Scenario 2: Model Interpretability over Different Time Horizons

In the following discussions, we evaluate CTI using the following questions. (1) Can users choose the better model (Analysis in Scenario 2 and 3)? (2) Are researchers able to understand regularities in forecasting by looking at interpretable results (Regularities in Scenario 2 and 3)? (3) Based on interpretability, can some strategy make model performance more powerful (Strategy in Scenario 2 and 3)?

1) Experiment

Category design: To explain models for different time horizons, four modeling methods are used to predict the future 96 time points. In addition, 96-dimensional historical power and NWP WS values at five heights are used as input (each meteorological feature is analyzed in Sections 2.2 and 3.2; in this section, only WS values at five heights are used as input). Time points within the same time horizon are considered to belong to the same category. Training, verification, and test sets are composed of data of 210, 30, 60 days, respectively.

- Model 1: One predictor is trained with the input of future NWP and the past 96-points power. Input dimension of each sample is 576 ($5 \times 96 + 96$), and output dimension is 96.
- Model 2: One predictor is trained with input of future NWP. Input dimension of each sample is 5, and output dimension is 1.
- Model 3: One predictor is trained with input of 96 point-historical power. Input dimension of each sample is 96, and output dimension is 96.
- Model 4: 96 predictors are trained with input of future NWP and the past 96-point power. Input dimension of each sample is 101 ($5 \times 1 + 96$), and output dimension is 1.

2) Analysis

According to pre-knowledge, power correlation decreases with increase in time horizons. Although there is a certain daily cycle law, under the three correlation indexes, P_{t-96} ranks 82^{nd}, 56^{th}, and 54^{th}, respectively (sorting 101 features, including WP lagging 96 points and WS at five heights). WS100 is located on 13^{th}, 18^{th}, and 13^{th}, respectively (five

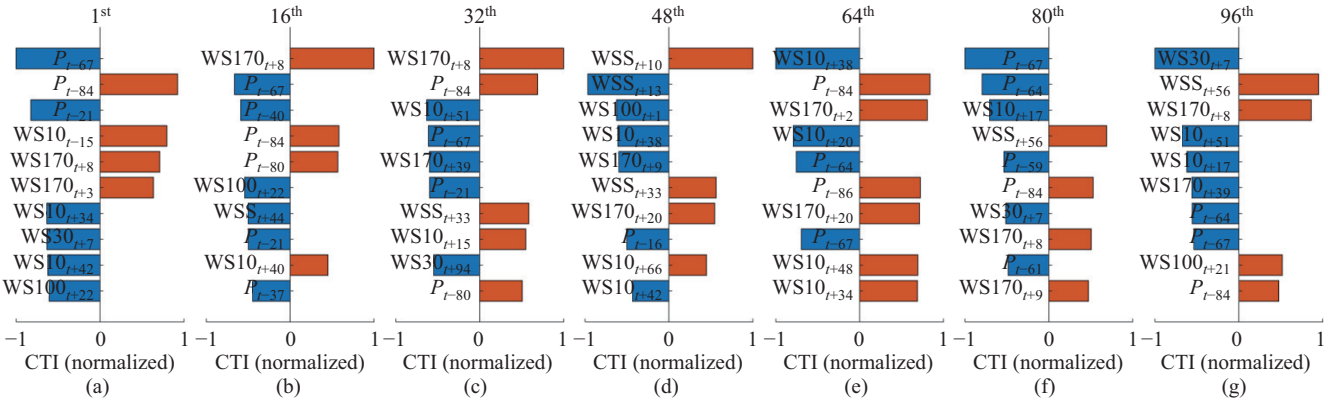

Fig. 11. CTI at different time scales for Model 1. (a) 1st. (b) 16th. (c) 32nd. (d) 48th. (e) 64th. (f) 80th. (g) 96th.

heights-wind speeds are scattered from 13th to 30th). This means that in the 12-step prediction ahead, the model's most trusted feature should be the nearest historical power. Other features may provide useful information, but not the "most trusted". When the forecast exceeds 18 steps, the most trusted feature is wind speed. The most trusted feature in the 18-step advance forecast should be WS100 or P_{t-1}. The most trusted feature is not fixed because we cannot confirm which indicator is more reasonable, all three of which have been used in WPF.

Figures 11 to 14 shows ranking of the CTI at 1st, 16th, 32nd, 48th, 64th, 80th, and 96th time points for the four models, respectively (t is the forecasting time point, and P_{t-1} is the historical power at the closest time point). We can draw the following conclusions:

- Model 1 should be considered risky model under the current predictor. In Fig. 11(a), there is no timing information close to starting time. Even in other subgraphs, results of feature mining do not accord with general cognizance.
- Model 2 is logically consistent. Feature mining results of Model 3 tend to be confused when the period is large.
- Model 4 is better since it uses more features and conforms to physical mechanisms. All top-ranking features in Fig. 14(a) are historical power values, and their CTI is consistent with time-series correlation. Among features at the 16th point, meteorological factors (i.e., WS30) already have relatively high CTI values. CTI values at subsequent time points suggest that WS values at various heights become the most important features with model perspectives, which is consistent with pre-analysis.

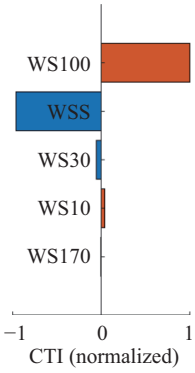

Fig. 12. CTI at different time scales for Model 2.

3) Regularities

To determine the type and importance of model-captured features over different time horizons more accurately, word clouds produced with every 16 time points as one unit and top five features with the highest η values are shown in Figs. 15 and 16 (P_{ti} or WSS_{ti} is used to represent P_{t-i} or WSS_{t-i}). (a), (b), (c), (d), (e), and (f) show explanations for points 1–16, 17–32, 33–48, 49–64, 65–80, and 81–96, respectively. We

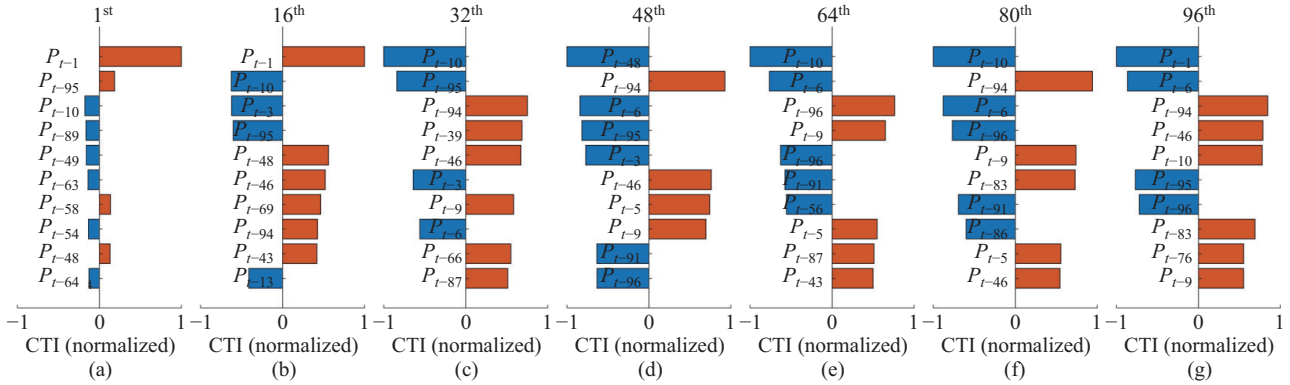

Fig. 13. CTI at different time scales for Model 1. (a) 1st. (b) 16th. (c) 32nd. (d) 48th. (e) 64th. (f) 80th. (g) 96th.

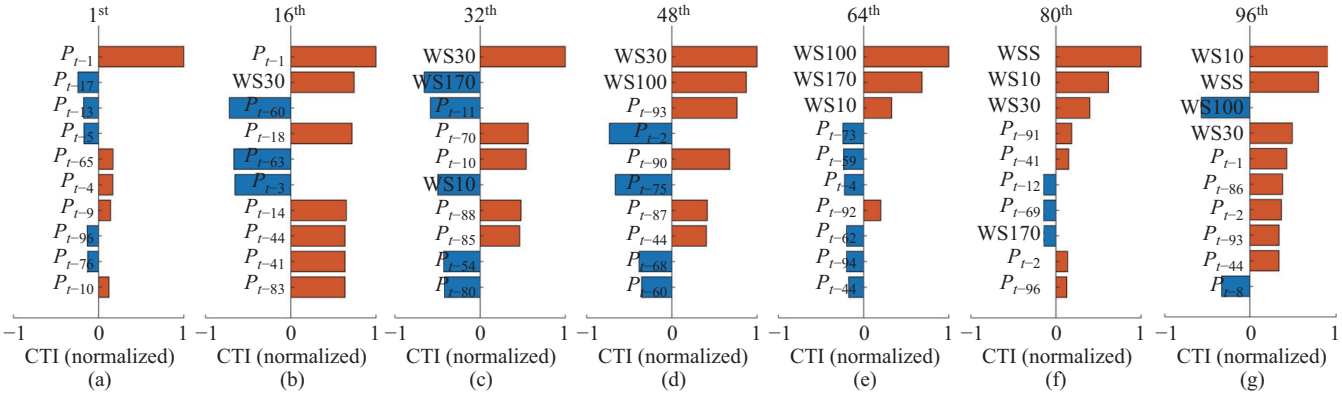

Fig. 14. CTI at different time scales for Model 1. (a) 1^{st}. (b) 16^{th}. (c) 32^{nd}. (d) 48^{th}. (e) 64^{th}. (f) 80^{th}. (g) 96^{th}.

Fig. 15. Word cloud of each unit for Model 1. (a) 1^{st}–16^{th}. (b) 17^{th}–32^{nd}. (c) 33^{rd}–48^{th}. (d) 49^{th}–64^{th}. (e) 65^{th}–80^{th}. (f) 81^{st}–96^{th}.

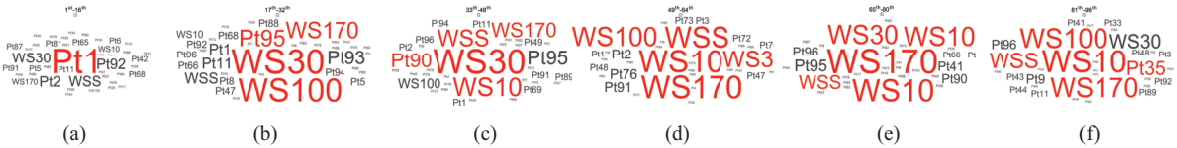

Fig. 16. Word cloud of each unit for Model 4. (a) 1^{st}–16^{th}. (b) 17^{th}–32^{nd}. (c) 33^{rd}–48^{th}. (d) 49^{th}–64^{th}. (e) 65^{th}–80^{th}. (f) 81^{st}–96^{th}.

can get the following regularities:

- The conclusion that Model 1 is risky will not change. Although Fig. 15(a) is mainly time-series information and weather information in a larger period, corresponding relationship of time is vague.
- Even if the pattern of daily cycles is not at the top of correlation rankings, the model may mine information that is excluded from the most recent power timing. This is because in Fig. 13(a) and Fig. 14(a), P_{t-95} and P_{t-96} is in 2^{nd} and 8^{th} place, respectively, which is much higher than their ranking by correlation degree. P_{t-92} accounts for large size of black fonts in Fig. 16.
- Time-series features are basically in accordance with the laws of physics. This also explains why the time-series extrapolation method is usually adopted in ultra-short-term (0–4 h) WPF [52] from the perspective of feature mining ability of the model. However, the model considers WS features to be the most important features, apart from P_{t-1}, at the 4^{th} hour (i.e., the 16^{th} point). This suggests that even in ultra-short-term forecasting, when the time-series extrapolation method is used, ignoring WS information over a relatively long time horizon may still result in a corresponding error. Use of the modeling method based on power time series alone is unreasonable beyond the 4^{th} hour, whether from the perspective of pre-analysis or model.

4) Strategy

However, chaotic feature results still exist under some time horizons in Model 4. To maximize retention of useful information, a combination strategy is proposed, as shown in Fig. 17. Several modeling methods are combined into a GTM according to pre-analysis and interpretable results under all-time horizons. For time horizons of 24 h (96 time points), Model 4 was used for training, and results under all time horizons were tested by interpretable analysis. The decision process is summarized by the following steps.

For $[t, t+12]$, determine whether the most trusted feature of Model 4 is P_{t-1}. If so, output results of Model 4; otherwise, output results of Model 3. For $[t+13, t+17]$, the result of Model 4 is output directly. For larger time horizons, determine whether the most trusted feature of Model 4 is WS100. If so, output result of Model 4; otherwise, output result of Model 2.

Final combination result is shown in Table II. Meanwhile, to verify effectiveness of the proposed method, a classical error-based combination method (EBM) is used as a comparison combination method. Error statistics of the verification set for four models under different time horizons are shown in Fig. 18. The new combination method is compared with EBM under each time horizon, as shown in Table III.

The error of EBM increases unexpectedly in the test set. Compared with EBM, GTM shows stronger robustness. Similar to risky models, trusted models do not necessarily have highest accuracy in every sample but have highest accuracy in total. Although compared with Model 4, accuracy improvement is not obvious; this is because feature mining results of Model 4 have met physical law. Results shown in Figs. 14 and

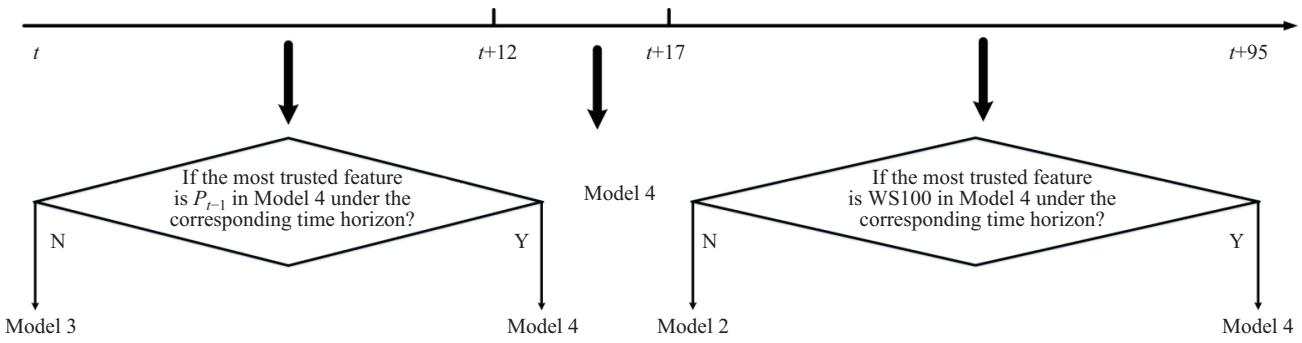

Fig. 17. Combined structures with different modeling methods in multiple time horizons.

TABLE II
FINAL COMBINATION RESULT WITH GRU

Model	Time horizons (i^{th})
Model 2	19, 20, 22, 24–29, 31–48, 50–59, 61–63, 66–68, 71, 73–77, 79–87, 91, 92, 94–96
Model 3	8
Model 4	1–7, 9–18, 21, 23, 30, 49, 60, 64, 65, 69, 70, 72, 78, 88–90, 93

TABLE III
NRMSE OF DIFFERENT METHODS WITH GRU (%)

Time horizons	GTM	EBM	Model 1	Model 2	Model 3	Model 4
1^{st}–16^{th}	11.16	11.38	13.78	14.21	11.86	**11.14**
17^{th}–32^{nd}	**12.19**	13.50	13.70	12.01	19.27	13.50
33^{rd}–48^{th}	**14.30**	17.50	15.45	15.41	20.78	14.80
49^{th}–64^{th}	17.38	18.45	**16.90**	17.37	24.63	17.83
65^{th}–80^{th}	18.87	18.95	**18.05**	19.03	23.42	18.95
81^{st}–96^{th}	**16.14**	16.29	17.39	16.43	20.02	15.90
Total	**14.69**	15.74	15.11	15.02	19.51	14.87

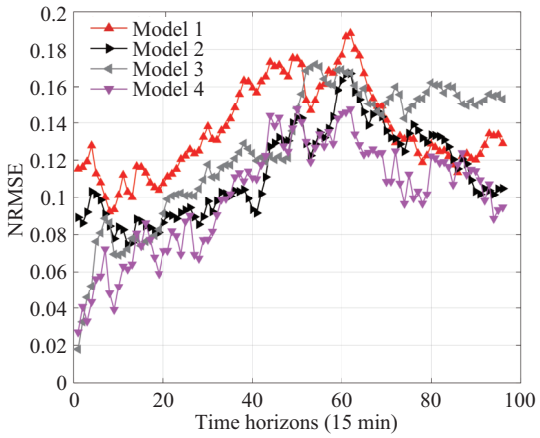

Fig. 18. Errors of four modeling methods in multiple time horizons.

16 are consistent with operation mechanism of the turbines.

To further verify effectiveness of GTM, RBFNN and BPNN are used as predictors to repeat the experiment. Final combination result is shown in Table IV, and accuracy statistics results are shown in Table V. Among the two predictors, overall accuracy of GTM is the highest. However, accuracy of EBM is unstable (in BPNN, error is less than the other four modeling methods, but in RBFNN, error is higher than Model 2 and Model 4). Q_{R} and N_{MAE} statistics are shown in Table VI, and

TABLE IV
FINAL COMBINATION RESULT WITH RBFNN AND BPNN

Model	Time horizons (i^{th})	
	RBFNN	BPNN
Model 2	19, 21, 23, 25, 27, 28, 30–34, 36, 38–40, 43, 46–52, 54–59, 62, 64, 67–73, 75, 76, 78, 79, 82–88, 90–96	19–25, 27–31, 33–39, 41–44, 46–52, 54–64, 66–69, 72, 74–81, 83–85, 87–96
Model 3	1, 2, 12	6, 10
Model 4	3–11, 13–18, 20, 22, 24, 26, 29, 35, 37, 41, 42, 44, 45, 53, 60, 61, 63, 65, 66, 74, 77, 80, 81, 89	1–5, 7–9, 11–18, 26, 32, 40, 45, 53, 65, 70, 71, 73, 82, 86,

GTM shows the best performance. Q_{R} shows the combination model based on interpretable results has fewer extreme error scenarios, which is valuable to the scheduler. The "trustworthy model" in which feature mining results accord with physical law has stronger robustness.

D. Scenario 3: Model Interpretability for Each Season

1) Experiment

Category design: output of a WF typically varies from season to season. In this section, every three calendar months (90 days by default) is regarded as one category. Training, verification, and test sets are composed of data of 60, 10, 20 days, respectively. Only meteorological data are used as input to analyze the model established for each season.

2) Analysis

Scenario 1 proved that all feature inputs may train risky models. Feature selection can theoretically reduce input of redundant information. The mrMR [53] method is used for feature selection in different seasons, and results of the top 12 are shown in Table VII. We use the top 10 weather information as input features. Fig. 19 shows interpretability results for the model established for each season. The model in autumn is more reliable since feature mining results consist of the power generation law of the turbine, even if it is not completely consistent with a priori knowledge. The other three seasons give more weight to the weakly correlated characteristics, which should be regarded as the risky model.

3) Regularities

- The difference between the trust characteristics of data-driven models provides the basis for the technical route of refined WPF modeling. Whether through pre-analysis or trusted by the model, features are varied from season to

TABLE V
NRMSE OF DIFFERENT METHODS WITH RBFNN AND BPNN (%)

Time horizons	RBFNN						BPNN					
	GTM	EBM	Model 1	Model 2	Model 3	Model 4	GTM	EBM	Model 1	Model 2	Model 3	Model 4
1^{st}–16^{th}	10.77	**10.71**	10.74	13.21	11.77	11.05	11.11	**10.66**	10.69	13.54	13.06	10.69
17^{th}–32^{nd}	12.83	13.15	13.55	**12.54**	18.88	13.15	**12.96**	14.08	13.75	12.82	19.23	14.08
33^{rd}–48^{th}	15.39	15.84	15.60	15.45	20.77	**15.36**	15.96	**15.84**	17.40	15.92	21.61	16.00
49^{th}–64^{th}	**17.62**	18.25	17.80	17.17	25.27	18.61	18.14	18.17	18.52	**18.02**	25.79	19.07
65^{th}–80^{th}	19.49	19.93	19.52	**19.10**	24.36	20.20	19.54	19.72	20.49	19.62	24.08	**19.38**
81^{st}–96^{th}	15.80	15.78	17.09	**15.73**	20.01	15.81	16.09	**16.02**	18.56	16.02	20.38	16.37
Total	**14.84**	15.08	15.17	14.91	19.62	15.03	**15.14**	15.19	15.81	15.24	20.08	15.46

TABLE VI
QR AND NMAE OF DIFFERENT METHODS WITH GRU (%)

Predictor	GTM		EBM		Model 1		Model 2		Model 3		Model 4	
	QR	NMAE	QR	NMAE	QR	NMAE	QR	NMAE	QR	NMAE	QR	NMAE
GRU	**88.94**	**11.89**	86.88	12.61	88.07	12.27	88.66	12.16	77.01	16.51	88.47	11.97
RBFNN	**89.50**	**11.80**	88.37	11.98	87.85	12.24	88.47	12.07	77.86	16.52	87.31	12.01
BPNN	**87.83**	**12.11**	87.50	12.20	85.78	12.80	87.81	12.33	77.78	17.00	86.72	12.37

TABLE VII
FEATURE RANKING OF DIFFERENT SEASONS BASED ON MRMR

Ranking	SP	SU	FA	WI
1	WS170	WS100	WS100	WS100
2	WS100	WS30	WS170	WS170
3	SLP	WS170	WS30	WS30
4	SWR	WS10	WS10	SHF
5	RH2	WSS	WSS	WS10
6	WD30	SHF	SP	WSS
7	CP	WD30	CP	SP
8	LWR	CP	MF	CP
9	SP	RH2	SHF	LSP
10	SHF	SP	RH2	RH2
11	WD10	MF	SLP	SWR
12	WS30	WD10	TP	SLP

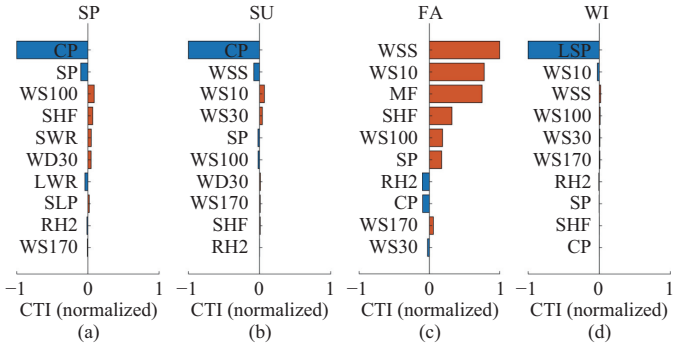

Fig. 19. CTI of each month. (a) SP. (b) SU. (c) FA. (d) WI.

season. Even for WS information, WS170 and WS100 are the most correlated WS features for seasons for SP and SU, respectively. This result is attributed to the difference in mapping characteristics between seasons. In other words, season-based modeling is a simple, refined modeling method.

- Even after pre-feature selection, some features still make the model risky.

4) Strategy

Therefore, without changing the predictor, feature selection can be further refined according to interpretable results, as shown in Fig. 20. Until the most trusted feature of the model meets prior knowledge, we cannot think of it as a trustworthy model. Final input features and error statistics are shown in Table VIII. Improved input, as a trustworthy model, can reduce N_{RMSE} of 0.28%—1.27% based on 11.88–18.43% in each season. Overall accuracy is improved when adjustment of the input feature makes the model trustworthy.

TABLE VIII
FEATURE INPUTS AND RESULTS OF DIFFERENT SEASONS IN VERIFICATION SET BASED WITH MRMR-ISM

Season	Input	Initial NRMSE	Improved NRMSE
SP	WD30, SHF, RH2, SWR, SLP, WS100, LWR, SP, WS170	18.43	**17.16**
SU	RH2, SHF, WS10, WS100	15.15	**14.52**
FA	SHF, RH2, MF, SP, CP, WS170, WS10, WS30, WS100	11.88	**11.60**
WI	SHF, WS170, RH2, SP, WSS, WS10, WS100	16.40	**15.61**

In addition, actual power (AP), original predicted power (PP-initial), improved predicted power (PP-improved), WS100, and actual-WS (AWS) data for three consecutive days with a relatively stationary actual power output are selected from the test set, as shown in Fig. 21. AP is close to 0 before the 200^{th} time point. Trend of PP-actual is inconsistent with trends of NWP WS, AWS, and AP, which suggests the model is affected by redundant features. Trend of PP-improved after removal of interfering features is clearly more consistent with trend of AP. After the 200^{th} time point, there is a ramp-up and a ramp-down in AWS, and even PP-improved is unable to match changes in AP, which is primarily because of the fuzzy mapping relation between NWP WS and AWS or AP for this period. To address this problem, correction of NWP data and further exploitation of their hidden information may be considered. Improving input features for a black-box model based on transparent learning results is fundamentally different from conventional correlation-based feature selection method or trial-and-error method.

To further judge effectiveness of the feature selection scheme, two other feature selection methods, namely recursive feature elimination (RFE) [54] and random forest-out-of-bag

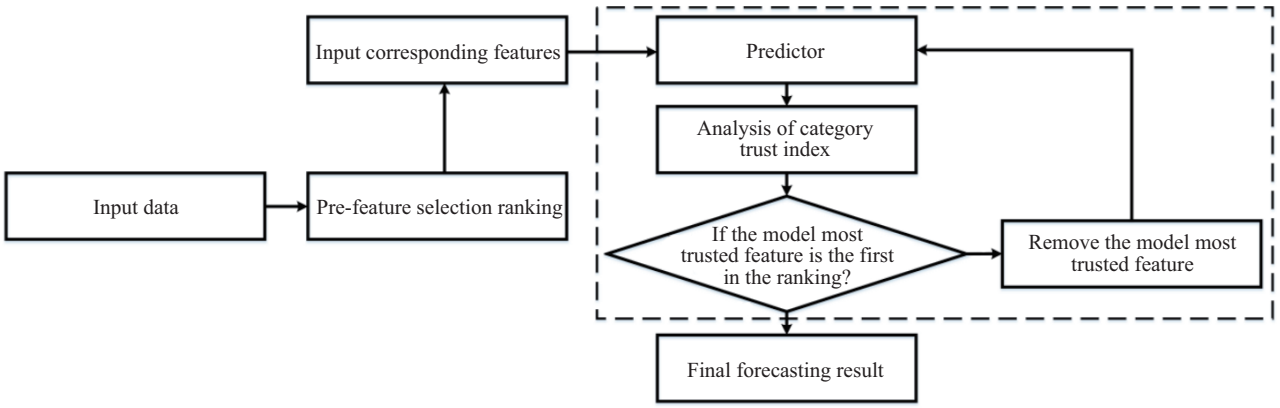

Fig. 20. Process of further features selection.

error (RF-OOB) [55], are adopted, which belong to Wrapper, Embedded, and Filter with mrMR. For all three feature selection methods, input information dimensions are set to 10. Based on the three feature selection methods and interpretable selection method (ISM), result statistics of the test data set are shown in Table IX. ISM can further improve accuracy of the original feature selection method. Range of accuracy improvement is 0.36%–4.69%. Q_{R} and N_{MAE} are shown in Table X, and ISM shows higher performance overall. N_{MAE} statistics are similar in winter before and after adjustment for RFE, but statistical results based on Q_{R} still show that ISM has lower extreme error level generally. Moreover, ISM based on CTI can make the model more robust, such as feature input that improves verification set accuracy, which is also improved in test set.

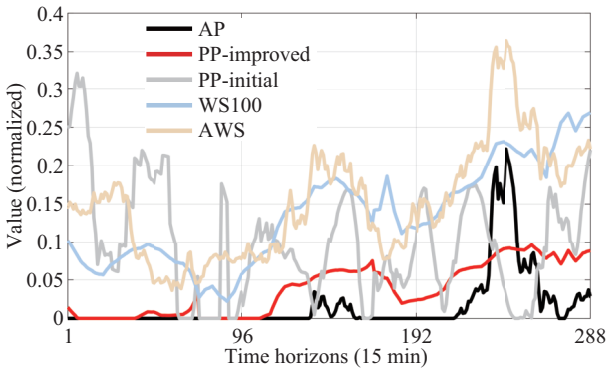

Fig. 21. Time series comparison of wind speed, actual power, and predicted power (normalized by the maximum value of each feature).

TABLE IX
NRMSE OF DIFFERENT SEASONS IN TEST SET (%)

Season	mrMR		RFE		RF-OOB	
	Initial	ISM	Initial	ISM	Initial	ISM
SP	23.09	**22.13**	28.48	**23.85**	23.69	**19.69**
SU	18.25	**15.85**	18.14	**16.67**	20.29	**15.6**
FA	18.79	**16.30**	17.41	**16.19**	20.62	**16.69**
WI	17.23	**16.87**	19.55	**19.13**	18.18	**17.38**

IV. CONCLUSION

The focus of current research or application in the power industry is gradually shifting toward WPF models based on intelligent learning algorithms. However, there is still a lack of qualitative and quantitative analyses of the operational mechanism of such black-box models, which can lead to application risk on the dispatch side. We note that explanations are particularly useful in these scenarios if a method can produce them for any model, so that a variety of models can be compared and provide evaluation information in addition to errors. In this study, based on the LIME algorithm, a new CTI η for wind power was defined to qualitatively and quantitatively explain the internal mechanism of several technical routes for application of ML models to WPF. Interpretable post-feature analysis method is used for modeling adjustment and feature selection, based on the principle that data-driven results should be consistent with physical mechanism. Through the methods, accuracy of the improved model is robustly improved in multiple scenes.

1) When all 24 NWP features are input into the model, the model may assign relatively large weights to features with stationary time series, which leads to relatively large forecasting errors. The model in which the feature mining result is completely contradictory to the physical mechanism

TABLE X
QR AND NMAE OF DIFFERENT SEASONS IN TEST SET (%)

Season	QR						MAE					
	mrMR		RFE		RF-OOB		mrMR		RFE		RF-OOB	
	Initial	ISM	Initial	ISM	Initial	ISM	Initial	ISM	Initial	ISM	Initial	ISM
SP	71.77	**78.23**	59.06	**73.63**	67.66	**81.25**	18.49	**17.41**	23.98	**18.60**	19.91	**15.54**
SU	91.93	**95.94**	91.88	**94.53**	89.11	**96.77**	15.65	**13.15**	15.85	**14.03**	17.54	**13.14**
FA	98.13	**99.27**	98.65	**99.01**	98.13	**99.53**	15.31	**13.28**	14.14	**13.28**	16.53	**13.43**
WI	99.58	**99.79**	99.38	**99.53**	**99.58**	**99.58**	14.23	**13.72**	**16.35**	16.49	14.87	**14.72**

should be regarded as a risky model. CTI can complement these existing systems and allow users to assess trust even when the samples seem "correct" but is made for the wrong reasons. Meanwhile, from the perspective of increasing forecasting accuracy, improving data mining ability of models or preprocessing based on feature selection is necessary.

2) When time-series and NWP WS data are simultaneously input into the model, the model trusts time-series features more for the first 4 h of the period selected for modeling. Applicability of the technical route for the time-series extrapolation method to ultra-short-term forecasting was explained from a modeling perspective. However, CTI or correlation values of the WS features are similar to that of P_{t-1} at the 16^{th} point, suggesting that even in ultra-short-term forecasting, NWP features must be introduced to time points over relatively long time horizons. Meanwhile, the model may mine information from daily cycles that is excluded from the most recent power timing. Moreover, a combined model based on performance in the verification set may be unstable. Based on the rule that the result of feature mining must be consistent with physical mechanism, the combined trust model can improve accuracy robustly in the three predictors with GRU, RBFNN, and BPNN.

3) From a modeling perspective, modeling refinement was shown to be reasonable based on the difference between features captured by data-driven models and analysis results for different seasons. At the same time, features trusted by the models may differ from the physical mechanism even after feature selection. Based on an interpretable ML model, of all three feature selection methods (i.e., mrMR, RFE, and RF-OOB), the proposed feature ranking method can be used to further reduce forecasting error of 17.23%–28.48% by 0.36%–4.69%. ML interpretability provides a bridging relationship between data-driven and physical-mechanism-driven models for WPF.

REFERENCES

[1] B. Yang *et al*., "Classification and Summarization of Solar Irradiance and Power Forecasting Methods: A Thorough Review," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 3, pp. 978–995, May 2023, doi: 10.17775/CSEEJPES.2020.04930.

[2] M. Yang, L. B. Zhang, Y. Cui, Y. Zhou, Y. L. Chen, and G. G. Yan, "Investigating the wind power smoothing effect using set pair analysis," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 3, pp. 1161–1172, Jul. 2020, doi: 10.1109/TSTE.2019.2920255.

[3] M. J. Sanjari, H. B. Gooi, and N. K. C. Nair, "Power generation forecast of hybrid PV–wind system," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 2, pp. 703–712, Apr. 2020, doi: 10.1109/TSTE.2019.2903900.

[4] D. Y. Hong, T. Y. Ji, M. S. Li, and Q. H. Wu, "Ultra-short-term forecast of wind speed and wind power based on morphological high frequency filter and double similarity search algorithm," *International Journal of Electrical Power & Energy Systems*, vol. 104, pp. 868–879, Jan. 2019, doi: 10.1016/j.ijepes.2018.07.061.

[5] Y. N. Zhao, L. Ye, Z. Li, X. R. Song, Y. S. Lang, and J. Su, "A novel bidirectional mechanism based on time series model for wind power forecasting," *Applied Energy*, vol. 177, pp. 793–803, Sep. 2016, doi: 10.1016/j.apenergy.2016.03.096.

[6] M. Yang, C. Y. Shi, and H. Y. Liu, "Day-ahead wind power forecasting based on the clustering of equivalent power curves," *Energy*, vol. 218, pp. 119515, Mar. 2021, doi: 10.1016/j.energy.2020.119515.

[7] Y. Wang, Q. H. Hu, D. Srinivasan, and Z. Wang, "Wind power curve modeling and wind power forecasting with inconsistent data," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 1, pp. 16–25, Jan. 2019, doi: 10.1109/TSTE.2018.2820198.

[8] Q. Y. Xu, D. W. He, N. Zhang, C. Q. Kang, Q. Xia, J. H. Bai, and J. H. Huang, "A short-term wind power forecasting approach with adjustment of numerical weather prediction input by data mining," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1283–1291, Oct. 2015, doi: 10.1109/TSTE.2015.2429586.

[9] N. Korprasertsak and T. Leephakpreeda, "Robust short-term prediction of wind power generation under uncertainty via statistical interpretation of multiple forecasting models," *Energy*, vol. 180, pp. 387–397, Aug. 2019, doi: 10.1016/j.energy.2019.05.101.

[10] M. Yang, Y. Huang, C. Xu, C. Liu, and B. Dai, "Review of several key processes in wind power forecasting: Mathematical formulations, scientific problems, and logical relations," *Applied Energy*, vol. 377, no. PC, p. 124631, 2025, doi: 10.1016/j.apenergy.2024.124631.

[11] W. Swartout, C. Paris, and J. Moore, "Explanations in knowledge systems: design for explainable expert systems," *IEEE Expert*, vol. 6, no. 3, pp. 58–64, Jun. 1991, doi: 10.1109/64.87686.

[12] F. Faggin, "Neural network hardware," in *Proceedings of IJCNN International Joint Conference on Neural Networks*, 1992, pp. 153, doi: 10.1109/IJCNN.1992.287238.

[13] A. A. Patel and J. N. Dharwa, "An integrated hybrid recommendation model using graph database," in *Proceedings of 2016 International Conference on ICT in Business Industry & Government*, 2016, pp. 1–5, doi: 10.1109/ICTBIG.2016.7892680.

[14] M. Yang, Y. Guo, T. Huang, and W. Zhang, "Power prediction considering NWP wind speed error tolerability: A strategy to improve the accuracy of short-term wind power prediction under wind speed offset scenarios," *Applied Energy*, vol. 377, p. 124720, 2025, doi: https://doi.org/10.1016/j.apenergy.2024.124720.

[15] M. Yang, Y. Guo, B. Wang, Z. Wang, and R. Chai, "A day-ahead wind speed correction method: Enhancing wind speed forecasting accuracy using a strategy combining dynamic feature weighting with multi-source information and dynamic matching with improved similarity function," *Expert Systems with Applications*, vol. 263, p. 125724, 2025, doi: https://doi.org/10.1016/j.eswa.2024.125724.

[16] C. Y. Liu, X. M. Zhang, S. W. Mei, and F. Liu, "Local-pattern-aware forecast of regional wind power: adaptive partition and long-short-term matching," *Energy Conversion and Management*, vol. 231, pp. 113799, Mar. 2021, doi: 10.1016/j.enconman.2020.113799.

[17] H. S. Dhiman, D. Deb, and J. M. Guerrero, "Hybrid machine intelligent SVR variants for wind forecasting and ramp events," *Renewable and Sustainable Energy Reviews*, vol. 108, pp. 369–379, Jul. 2019, doi: 10.1016/j.rser.2019.04.002.

[18] Q. M. Zhu, H. Y. Li, Z. Q. Wang, and B. Wang, "Short-term wind power forecasting based on LSTM," *Power System Technology*, vol. 41, no.12, pp. 3797–3802, Dec. 2019, doi: 10.13335/j.1000-3673.pst.2017.1657.

[19] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, "XAI—Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, pp. eaay7120, Dec. 2019, doi: 10.1126/scirobotics.aay7120.

[20] A. S. Rao, (2019, Jan. 22). Responsible AI & national AI strategies [Online]. Available: https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/4%20International%20initiatives%20v3_0.pdf.

[21] T. Karatekin, S. Sancak, G. Celik, S. Topcuoglu, G. Karatekin, P. Kirci, and A. Okatan, "Interpretable machine learning in healthcare through generalized additive model with pairwise interactions (GA2M): predicting severe retinopathy of prematurity," in *Proceedings of 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications*, 2019, pp. 61–66, doi: 10.1109/Deep-ML.2019.00020.

[22] M. K. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: tree regularization of deep models for interpretability," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 1670–1678.

[23] H. Farhood, M. Saberi, and M. Najafi, "Improving object recognition in crime scenes via local interpretable model-agnostic explanations," in *Proceedings of the 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop*, 2021, pp. 90–94, doi: 10.1109/edocw52865.2021.00037.

[24] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 2513107, Aug. 2021, doi: 10.1109/TIM.2021.3107056.

[25] N. B. Kumarakulasinghe, T. Blomberg, J. T. Liu, A. S. Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," in *Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems*, 2020, pp. 7–12, doi: 10.1109/CBMS49503.2020.00009.

[26] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1399–1414, Mar. 2003, doi: 10.5555/944919.944980.

[27] Z. Lin and X. L. Liu, "Wind power forecasting of an offshore wind turbine based on high-frequency SCADA data and deep learning neural network," *Energy*, vol. 201, pp. 117693, Jun. 2020, doi: 10.1016/j.energy.2020.117693.

[28] W. C. Yeh, Y. M. Yeh, P. C. Chang, Y. C. Ke, and V. Chung, "Forecasting wind power in the Mai Liao Wind Farm based on the multi-layer perceptron artificial neural network model with improved simplified swarm optimization," *International Journal of Electrical Power & Energy Systems*, vol. 55, pp. 741–748, Feb. 2014, doi: 10.1016/j.ijepes.2013.10.001.

[29] A. Tascikaraoglu and M. Uzunoglu, "A review of combined approaches for prediction of short-term wind speed and power," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 243–254, Jun. 2014, doi: 10.1016/j.rser.2014.03.033.

[30] O. Abedinia, M. Lotfi, M. Bagheri, B. Sobhani, M. Shafie-Khah, and J. P. S. Catalão, "Improved EMD-based complex prediction model for wind power forecasting," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2790–2802, Oct. 2020, doi: 10.1109/TSTE.2020.2976038.

[31] Y. Zhang and J. J. Dong, "Least squares-based optimal reconciliation method for hierarchical forecasts of wind power generation," *IEEE Transactions on Power Systems*, to be published, doi: 10.1109/tpwrs.2018.2868175.

[32] Y. Hao and C. S. Tian, "A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting," *Applied Energy*, vol. 238, pp. 368–383, Mar. 2019, doi: 10.1016/j.apenergy.2019.01.063.

[33] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. Mcvean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.

[34] M. Optis and J. Perr-Sauer, "The importance of atmospheric turbulence and stability in machine-learning models of wind farm power production," *Renewable and Sustainable Energy Reviews*, vol. 112, pp. 27–41, Sep. 2019, doi: 10.1016/j.rser.2019.05.031.

[35] Q. W. Li, J. Z. Wang, and H. P. Zhang, "A wind speed interval forecasting system based on constrained lower upper bound estimation and parallel feature selection," *Knowledge-Based Systems*, vol. 231, pp. 107435, Nov. 2021, doi: 10.1016/j.knosys.2021.107435.

[36] X. J. Liu, H. Zhang, X. B. Kong, and K. Y. Lee, "Wind speed forecasting using deep neural network with feature selection," *Neurocomputing*, vol. 397, pp. 393–403, Jul. 2020, doi: 10.1016/j.neucom.2019.08.108.

[37] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv: 1702.08608, 2017.

[38] S. L. Ji, J. F. Li, T. Y. Du, and B. Li, "Survey on techniques, applications and security of machine learning interpretability," *Journal of Computer Research and Development*, vol. 56, no. 10, pp. 2071–2096, Oct. 2019, doi: 10.7544/issn1000-1239.2019.20190540.

[39] K. R. Chen and X. F. Meng, "Interpretation and understanding in machine learning," *Journal of Computer Research and Development*, vol. 57, no. 9, pp. 1971–1986, Sep. 2020, doi: 10.7544/issn1000-1239.2020.20190456.

[40] C. Molnar, *Interpretable Machine Learning*: *A Guide for Making Black Box Models Explainable*. Independently Published, 2022.

[41] Y. Wang, N. Gao, and G. Hug, "Personalized Federated Learning for Individual Consumer Load Forecasting," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 1, pp. 326–330, Jan. 2023, doi: 10.17775/CSEEJPES.2021.07350.

[42] K. Li, Y. Wang, N. Zhang, F. Wang, and C. Huang, "Spatio-temporal Granularity Co-optimization Based Monthly Electricity Consumption Forecasting," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 5, pp. 1980–1984, Sep. 2023, doi: 10.17775/CSEEJPES.2022.01040.

[43] Y. N. Zhao, L. Ye, P. Pinson, Y. Tang, and P. Lu, "Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5029–5040, Sep. 2018, doi: 10.1109/tpwrs.2018.2794450.

[44] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K. R. Müller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, Aug. 2010.

[45] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[46] H. Ishibuchi and T. Yamamoto, "Interpretability issues in fuzzy genetics-based machine learning for linguistic modelling," in *Modelling with Words*, J. Lawry, J. Shanahan, and A. Ralescu, Eds. Berlin, Heidelberg: Springer, 2003, pp 209–228, doi: 10.1007/978-3-540-39906-3_11.

[47] M. Yang and H. Dong, "Short-term wind power interval prediction based on wind speed of numerical weather prediction and Monte Carlo method," *Automation of Electric Power Systems*, vol. 45, no. 5, pp. 79–85, Mar. 2021, doi: 10.7500/AEPS20200426001.

[48] R. Errouissi, J. Cardenas-Barrera, J. L. Meng, E. Castillo-Guerra, X. Gong, and L. C. Chang, "Bootstrap prediction interval estimation for wind speed forecasting," in *Proceedings of 2015 IEEE Energy Conversion Congress and Exposition*, 2015, pp. 1919–1924, doi: 10.1109/ECCE.2015.7309931.

[49] Z. Q. Xie, T. Y. Ji, M. S. Li, and Q. H. Wu, "Quasi-Monte Carlo based probabilistic optimal power flow considering the correlation of wind speeds using copula function," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2239–2247, Mar. 2018, doi: 10.1109/TPWRS.2017.2737580.

[50] Y. X. Wen, D. AlHakeem, P. Mandal, S. Chakraborty, Y. K. Wu, T. Senjyu, S. Paudyal, and T. L. Tseng. "Performance evaluation of probabilistic methods based on bootstrap and quantile regression to quantify PV power point forecast uncertainty," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 4, pp. 1134–1144, Apr. 2020, doi: 10.1109/TNNLS.2019.2918795.

[51] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

[52] B. Ernst, B. Oakleaf, M. L. Ahlstrom, M. Lange, C. Moehrlen, B. Lange, U. Focken, and K. Rohrig, "Predicting the wind," *IEEE Power and Energy Magazine*, vol. 5, no. 6, pp. 78–89, Nov./-Dec. 2007, doi: 10.1109/MPE.2007.906306.

[53] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[54] H. Liu and C. Chen, "Data processing strategies in wind energy forecasting models and applications: a comprehensive review," *Applied Energy*, vol. 249, pp. 392–408, Sep. 2019, doi: 10.1016/j.apenergy.2019.04.188.

[55] M. Yang and Y. Y. Bai, "Ultra-short-term prediction of wind power based on multi-location numerical weather prediction and gated recurrent unit," *Automation of Electric Power Systems*, vol. 45, no. 1, pp. 177–183, Jan. 2021, doi: 10.7500/AEPS20200521007.

Mao Yang received the Ph.D. degree in Control Theory and Control Engineering from Jilin University, Changchun, China, in 2010. Since 2018, he has been a Professor with the School of Electrical Engineering, Northeast Electric Power University, Jilin, China. His research interests are in the areas of load forecasting and micro-grid operation management with an emphasis on wind power/photovoltaic power prediction.

Chuanyu Xu received the B.Sc. degree in New Energy Science and Engineering from Northeast Electric Power University, Jilin, China, in 2019. He is currently pursuing his M.S. degree in Electrical Engineering at Northeast Electric Power University, Jilin, China. His research interest is in the area of wind power forecasting.

Yuying Bai received the M.S. degree from Northeast Electric Power University, Jilin, China, in 2019. She is currently working in the company of State Grid in Daqing. Her research interests span the areas of power system analytics and machine learning and data mining with applications in wind power forecast.

Miaomiao Ma received the Ph.D. degrees in Control Theory and Control Engineering from Jilin University, Changchun, China, in 2009. She is currently an Associate Professor with the North China Electric Power University, Beijing, China. Her research interests include model predictive control, optimal and robust control, and applications in renewable power systems.

Xin Su received the M.S. degree from Jilin University, Changchun, China, in 2011. She is currently working in the School of Science, Northeast Electric Power University, Jilin, China. Her research interest is in the area of data mining.