Team zhangshijin WPFormer: A Spatio-Temporal Graph Transformer with Auto-Correlation for Wind Power Prediction

Xuefeng Liang
Tsinghua University
Haidian Qu, Beijing Shi, China
liangxf20@mails.tsinghua.edu.cn

Qingshui Gu
Tsinghua University
Haidian Qu, Beijing Shi, China
gqs19@tsinghua.org.cn

Su Qiao
Zhejiang Gongshang University
Hangzhou, Zhejiang, China
qiaosu98@outlook.com

Zhuwang Lv
Zhejiang Gongshang University
Hangzhou, Zhejiang, China
zhuwanglv@gmail.com

Xin Song
University of Toronto
Toronto, CA
xsong8486@gmail.com

ABSTRACT

In this paper, we briefly introduce the solutions of our team 'zhangshijin' for the Spatial Dynamic Wind Power Forecasting Challenge in Baidu KDD CUP 2022. Given the spatial distribution of wind turbines and dynamic environmental factors such as time, weather, etc., in the past, the competition aims to predict the wind power generated by wind turbines in the future. The volatility of power generation and the long forecast time are huge challenges for this mission. To address these issues, we propose a series of methods, which includes abnormal data cleaning based on the wind power curve, effective selection and construction of features, wind turbines' coding of spatio-temporal graph neural network, WPFormer model, POPtree, and two model fusion strategies. Both online and offline, our method achieves effective and powerful improvements.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

KEYWORDS

wind power curve, graph neural network, auto-correlation, spatio-temporal prediction

1 INTRODUCTION

The use of wind power, a pollution-free and renewable form of energy, to generate electricity has attracted increasing attention in recent years. However, intermittent electricity generation resulting from the stochastic nature of wind speed poses challenges to the safety and stability of electric power grids. For the efficient use of wind energy, Wind Power Forecasting (WPF) has been widely recognized as one of the most critical issues in wind power integration, and it aims to accurately estimate the wind power supply of wind farms at different time scales.

In recent years, there has been an explosion of research into wind power prediction [14, 17, 18]. However, dealing with the WPF problem is still quite challenging because of the stochastic nature of wind speed and the complexity of turbine properties and status, but also it requires high prediction accuracy to ensure grid stability and supply security.

For WPF, traditional statistical models include autoregressive moving average (ARMA) [8], autoregressive integrated moving average (ARIMA) [27], and so on. These time series models show good performance in short-term WPF, but their long-term forecasting ability is poor. With the rise of machine learning, many new methods have been proposed and applied to WPF, e.g., support vector machine (SVM) [4], multi-layer perceptron (MLP) [9] and LightGBM [10]. With the development of deep learning technology, deep neural networks, including long short-term memory (LSTM) [12], convolutional neural network(CNN) [26], transformer [15], etc. have attracted greater attention in WPF due to their superior capability in dealing with complex nonlinear problems. Furthermore, in many previous studies, deep learning-based models usually showed better forecasting performance than traditional statistical-based models [23].

However, most existing datasets and models treat WPF as a time series prediction problem without knowing wind turbines' locations and context information. The KDD CUP presented a unique Spatial Dynamic Wind Power Forecasting dataset: SDWPF [29], the real-world data from Longyuan Power Group Corp. Ltd. (the largest wind power producer in China and Asia). This dataset has three main features: 1) Spatial distribution: the relative locations of all wind turbines given a wind farm for modeling the spatial correlation among wind turbines. 2) Dynamic context: the weather situations and internal turbine status detected by each wind turbine to facilitate the forecasting task. 3) Abnormal data: since the data comes from real data, there are a large number of outliers. It will also be a big challenge to eliminate the interference of outliers in the model.

Based on the above problems, we propose the following solutions to solve them.

- Firstly, abnormal wind power data may be caused by unstable sensors and equipment damage. The power generation and wind speed align with the S-curve distribution, and the data cleaning is completed for each wind turbine according to this distribution.
- Secondly, different features in the dataset have different effects on the results. Some features have positive effects, and some features may have adverse effects. In response to this phenomenon, we conduct feature correlation analysis, exclude features with poor correlation, and construct some new highly-correlated features.
- Thirdly, we design the WPFormer to use 1D convolutional neural networks and graph neural networks to encode the

temporal and spatial information of the wind turbine, respectively. Sequence-level learning is accomplished through progressive sequence decomposition and auto-correlation mechanisms.
- Finally, to improve model generability, we design point-by-point prediction based on the tree model using auxiliary wind turbine and data downsampling techniques, leveraging newly constructed statistical features. For different models, we used two model fusion methods.

The rest of this paper is arranged as follows. In Section II, we briefly introduce the related work. The structure of the proposed methodology is detailed in Section III. The comparison experiment results are reported in Section IV. Section V summarizes this paper.

2 RELATED WORK

2.1 Statistical-Based Wind Power Forecasting

Most early studies used statistical methods to model weather data and make predictions of wind power. So wind speed is first used to predict wind power as the most direct influencing factor. Brown et al. [3] built an Autoregressive Moving Average (ARMA) model to predict wind speed by exploiting basic features such as autocorrelation and diurnal nonstationarity. Then wind power is formulated as a function of wind speed so that the wind power can be estimated directly by applying transformations to wind speed data. Due to the high correlation between wind speed and wind power, many wind speed prediction methods have been developed. Firat et al. [7] predicted wind speed data by independent component analysis (ICA). Mohandes et al. [16] used support vector machines (SVM) to model wind speed. In addition to wind speed, many other factors (e.g., geographic location, wind direction) can also affect wind power. Alexiadis et al. [2] investigated the spatial correlation of wind speeds and developed a statistical learning model for wind power forecast. Combined with historical power data of the wind farm, Sideratos et al. [19] exploited forecasts of wind speed and direction and proposed an advanced statistical method for wind power forecast. Even though there are so many studies and efforts on the statistical methods to predict wind power, the substantial uncertainties in the weather data, such as wind speed, potentially undermine the linear or nonlinear relationships based on statistical methods, which can only make short-term predictions and is incompetent for long-term predictions.

In order to improve the performance of long-term forecasts, some more sophisticated statistical models are adapted. Ahmadi [1] developed a XGBoost model. Their results show that the tree model can perform well for long-term wind power prediction, especially for modeling wind power in different geographical locations. Thus, we propose a novel point-by-point prediction method based on the tree model.

2.2 Deep Learning-Based Wind Power Forecasting

Since deep learning methods show the superior capability to uncover fairly complex inherent correlations, many researchers working on wind power prediction have shifted their focus to them. Wang et al. [22] developed a novel wind power prediction framework based on convolution neural network (CNN) and wavelet transform. According to the experiment results, their model is robust against noisy data and achieves competitive performance on several real wind farm datasets. Song et al. [20] proposed a combined model based on generalized regression neural network (GRNN). Yu et al. [25] built an improved long short-term memory (LSTM) model and performed spectral clustering to optimize the forecasting effect further.

In recent years, an outstanding deep learning model called transformer [21] has emerged and achieved great success in long-term sequence prediction tasks. Inspired by the transformer, a series of new models and mutants have been studied. Informer [28] improved the self-attention block of the transformer, thus effectively capturing the exact long-range dependency coupling. Autoformer [24] proposed a new decomposition structure that helps the model to asymptotically decompose complex time series, thus significantly outperforming other methods on multiple long-range time series forecasting tasks. Likewise, FEDformer [30] proposed a model based on Transformer and the seasonal-trend decomposition methods, in which the decomposition method captures the global information of time series while Transformers capture detailed structures. However, there has been no extensive study of transformers or their many different variants in wind power prediction so far. The great success of transformers in sequence modeling and prediction has shown their great potential in this field, which has inspired us to develop a wind power forecasting model based on the transformer.

3 METHODLOGY

3.1 Data Engineering Based on Wind Power Curves

Data quality is the foundation of data science and machine learning, especially for data competitions. We first analyze the anomaly of the dataset, such as missing data, unknown data, and other singularities, by performing anomaly detection according to the wind power curve. And then perform the data processing to deal with these anomalies.

Data Analysis According to the data description from the organizing committee, the singular data can be divided into four categories: missing, zero-valued, unknown, and abnormal. By analyzing the 4,727,520 pieces of data in the training set, the proportions of the four kinds of data are counted, as shown in Table 1 below. After analyzing the table, we can draw two important conclusions. First, the proportion of data with problems is as high as 28.64%, which means that there are problems in nearly one-third of the data, which will bring significant challenges to the correct learning of the model. Second, the union of missing data, unknown data, and abnormal data accounts for 22.89% of the total data, which will not be used for score evaluation. Considering the similar distribution between the training dataset and test dataset, there should be an equivalent proportion of data on the test dataset that will not be used for score evaluation. Similar to the mask operation of the score, the position of the dataset on the test set that will not be used for evaluation is relatively random, and this random operation will cause greater volatility between online and offline scores, which has also been verified in this competition.

Table 1: Problem Data Situation

Data Type	Proportion(%)	Contents
Missing values	1.05	NULL
Zero values	26.72	$Patv < 0$ or $Prtv < 0$
Unknown values 1	6.33	$Patv \leq 0$ and $Wspd > 2.5$
Unknown values 2	20.83	$Pab1 > 89°$ or $Pab2 > 89°$ or $Pab3 > 89°$
Abnormal values	22.89	$Ndir > 720°$ or $Ndir < -720°$ or $Wdir > 180°$ or $Wdir < -180°$
Mask values	22.89	The union of missing values, unknown values and Abnormal values.
Total values	28.64	The union of all the above.

Anomaly Detection In practical application scenarios, wind power data can be subject to signal interference or equipment failure in the process of collection and transmission. In addition, the grid sometimes has insufficient capacity to accept strong wind power, and excessive power will be abandoned. The above reasons result in some inevitable errors during wind power data collection, which generates anomaly data, as counted in the data analysis section. If we use the anomaly data to train the machine learning model, the model's accuracy will decrease in the inference phase. Thus, we need to detect the anomaly data in pre-processing stage and remove or fix them.

In terms of mathematical modeling, the power generated is positively proportional to the cubic of wind speed measured, as shown in the following equation:

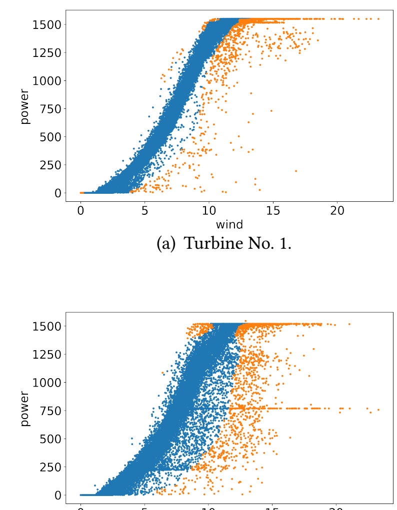
$$P = \frac{1}{2} C_p \rho A v^3, \tag{1}$$

where P is the power generation, C_p and A denote the wind energy utilization coefficient of the wind turbine and the impeller swept area, respectively; ρ is the air density, and v is the actual wind speed.

If drawing data on a scatter plot with wind speed as the x-axis and power generation as the y-axis, the data distribution is roughly S-shape. The data that deviate too much from the S-shaped distribution are considered anomalies, which data mining algorithms can detect according to the distribution. In this competition, we use the isolation forest [13] algorithm to identify these outliers. The isolation forest algorithm randomly selects feature values by dividing data, and sparsely distributed outliers are more likely to be picked out. Wind power curves are drawn by selecting turbines 1 and 2 from 134 wind turbines. Figure 1 shows the result of anomaly detection, where orange and blue dots denote anomaly and normal data, respectively.

In the competition, we need to use the historical power generation series to predict the future power generation series. Thus, the anomaly data in the training data cannot be deleted directly; otherwise, the historical series is incomplete. To fix the anomaly data, we use LightGBM [11] model with wind speed as the input to predict the power generation.

Data Cleaning After the anomaly detection, we continue to deal with the abnormal data. Since the sequencing data are inherently related, the anomalous pieces in the data need to be corrected instead of directly deleted. The anomaly detection can select the normal data out of anomalous data, so we can use normal data of wind turbines with wind speed and power generation to train a



(a) Turbine No. 1.

(b) Turbine No. 2.

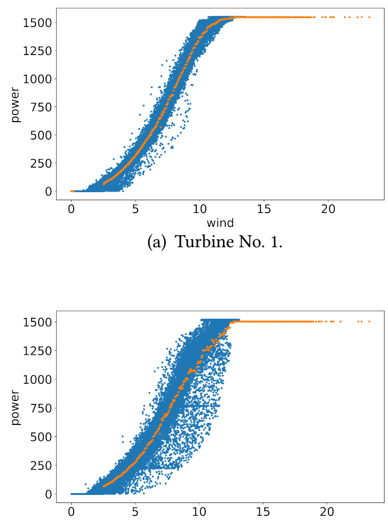
Figure 1: Anomaly detection

LightGBM model and use this model to correct the anomalous part. Eventually, there are 134 wind turbines in total that are corrected. We selected wind turbines 1 and 2 to demonstrate the cleaning effect, as shown in Figure 2, where the blue and orange curves represent normal and corrected data, respectively.

3.2 Feature Engineering

Data is the foundation of modeling, and features are the essential and significant information extracted from data. After the data processing, we analyze the correlation between the features, construct new highly-correlated features and remove some uncorrelated ones.

For more effective feature screening, we performed a correlation analysis on the features. Figure 3 depicts the correlation matrix of all variables, indicating a strong linear relationship between Wspd and Patv and a moderate correlation between angular variables (Pab1, Pab2 and Pab3) and Patv. Furthermore, the relationship between Prtv and Patv is statistically significant [6], so Prtv is also included. Since the remaining variables have no significant correlation with Patv, we exclude them from the modeling. At the same time, the LightGBM model in the data cleaning stage is used to analyze the



(a) Turbine No. 1.

(b) Turbine No. 2.

Figure 2: Distribution correction

feature importance. The features with less importance are Tmstamp, Wdir, Ndir, Etmp, and Itmp.

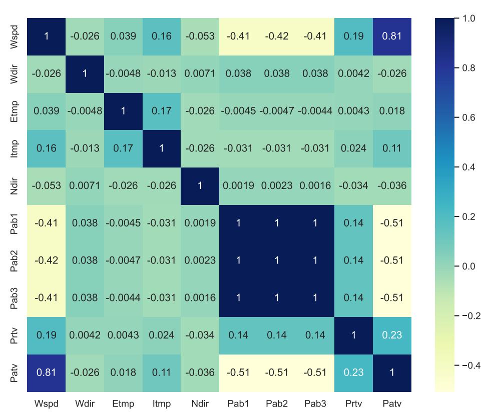


Figure 3: Heat map of feature correlation matrix.

We also construct statistical features to historical power generation data are shown in the A.1. Statistical features (7)-(12) (14)-(21) refer to tsfresh[6].

3.3 WPFormer Structure

Inspired by the innovations from Autoformer [24] and FEDformer [31], WPFormer uses an encoder-decoder structure that includes sequence decomposition using average pooling, auto-correlation mechanism for sequence-level connections, multi-head attention from original transformer [21], graph neural network encoding and feedforward module, as shown in Figure 4.

Encoder The spatial dimension information and temporal dimension information are learned by graph attention network and one-dimensional convolution, respectively. Moreover, the learned spatio-temporal information is taken as the encoder's input so that the sequence can be learned through autocorrelated attention and then decomposed by the average pooling. The entire learning process can be formulated as follows:

$$\begin{aligned}
\mathcal{X}_{\text{en}}^{0} &= \text{Conv1D}\left(\text{GAT}\left(\mathcal{X}, \mathcal{G}\right)\right), \\
\mathcal{S}_{\text{en},-}^{l,1} &= \text{AvgPooling}\left(\text{AutoCorrelation}\left(\mathcal{X}_{\text{en}}^{l-1}\right) + \mathcal{X}_{\text{en}}^{l-1}\right), \\
\mathcal{S}_{\text{en},-}^{l,2} &= \text{AvgPooling}\left(\text{FeedForward}\left(\mathcal{S}_{\text{en}}^{l,1}\right) + \mathcal{S}_{\text{en}}^{l,1}\right),
\end{aligned} \tag{2}$$

where $\mathcal{X}$ represents the initial turbine information, $\mathcal{G}$ represents the graph network code of the turbines, $\mathcal{X}_{\text{en}}^{l} = \text{Encoder}\left(\mathcal{X}_{\text{en}}^{l-1}\right)$ and $\mathcal{S}_{\text{en}}^{l,i}, i \in \{1,2\}$ represents the seasonal component after the i-th AvgPooling in the l-th layer respectively. On the wind power data, increasing the depth of the model will perform better than increasing the width, and our encoder has five layers eventually.

Decoder The decoder contains two parts of the input, one of which is the spatio-temporal embeddings extracted from the seasonal initialization term through graph attention and one-dimensional convolution, and the other one is the direct input of the trend initialization term. The first input part is refined by the auto-correlation mechanism and added to the encoder output for the following decoding. The second input part accumulates the trend items to increase the information utilization and improve the inference ability of the model.

$$\begin{aligned}
\mathcal{X}_{\text{de}}^{0} &= \text{Conv1D}(\text{GAT}\left(\text{ConCat}\left(\text{AvgPooling}\left(\mathcal{X}\right), \mathcal{S}_{\text{init}}\right), \mathcal{G}\right)), \\
\mathcal{T}_{\text{de}}^{0} &= \text{ConCat}\left(\text{AvgPooling}\left(\mathcal{X}\right), \mathcal{T}_{\text{init}}\right), \\
\mathcal{S}_{\text{de}}^{l,1}, \mathcal{T}_{\text{de}}^{l,1} &= \text{AvgPooling}\left(\text{AutoCorrelation}\left(\mathcal{X}_{\text{de}}^{l-1}\right) + \mathcal{X}_{\text{de}}^{l-1}\right), \\
\mathcal{S}_{\text{de}}^{l,2}, \mathcal{T}_{\text{de}}^{l,2} &= \text{AvgPooling}\left(\text{AutoCorrelation}\left(\mathcal{S}_{\text{de}}^{l,1}, \mathcal{X}_{\text{en}}^{N}\right) + \mathcal{S}_{\text{de}}^{l,1}\right), \\
\mathcal{S}_{\text{de}}^{l,3}, \mathcal{T}_{\text{de}}^{l,3} &= \text{AvgPooling}\left(\text{FeedForward}\left(\mathcal{S}_{\text{de}}^{l,2}\right) + \mathcal{S}_{\text{de}}^{l,2}\right), \\
\mathcal{T}_{\text{de}}^{l} &= \mathcal{T}_{\text{de}}^{l-1} + \mathcal{W}_{l,1} \cdot \mathcal{T}_{\text{de}}^{l,1} + \mathcal{W}_{l,2} \cdot \mathcal{T}_{\text{de}}^{l,2} + \mathcal{W}_{l,3} \cdot \mathcal{T}_{\text{de}}^{l,3}
\end{aligned} \tag{3}$$

where $\mathcal{S}_{\text{init}}$ and $\mathcal{T}_{\text{init}}$ represent the initialization of seasonal and trend terms, respectively, as shown in 4. $\mathcal{S}_{\text{de}}^{l,i}, \mathcal{T}_{\text{de}}^{l,i}, i \in \{1,2,3\}$ represent the seasonal and trend component after the i-th decomposition block in the l-th layer respectively. $\mathcal{W}_{l,i}, i \in \{1,2,3\}$ represents the linear projector for the i-th extracted trend $\mathcal{T}_{\text{de}}^{l,i}$.

Auto-Correlated Mechanism This mechanism was proposed by Autoformer [24] in 2021. It mainly calculates the sequence autocorrelation coefficient, aggregates similar subsequence information, realizes sequence-level connection, and completes better information aggregation. While introducing auto-correlation, we still use multi-head attention to increase model variability, resulting in improved prediction accuracy.

Graph Neural Network Representation Consider a wind farm containing 134 wind turbines, each wind turbine impacts the overall power generation, and there are similarities between different wind turbines. We use a graph structure to encode the relationships between turbines. Each turbine is a graph node. For a particular node, calculate the K nodes with the highest correlation with it, and establish connection edges.

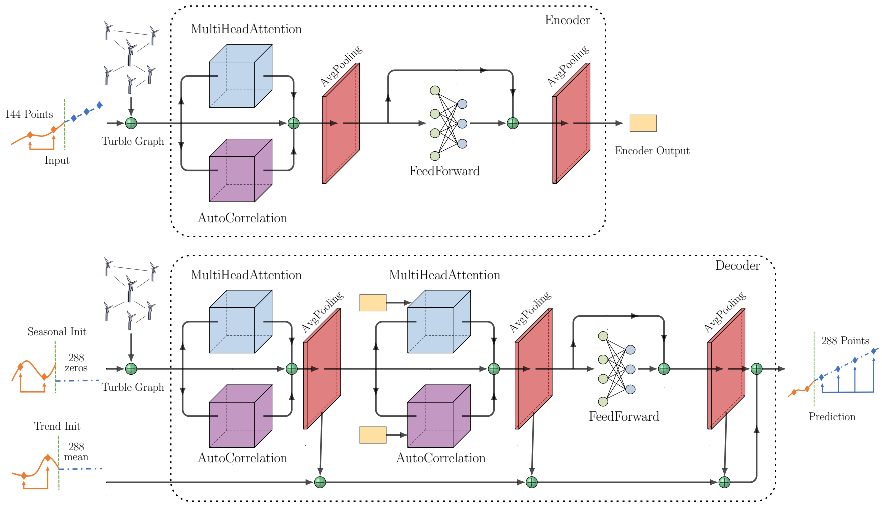


Figure 4: WPFormer Structure

3.4 Point-by-point prediction based on tree model (POPtree)

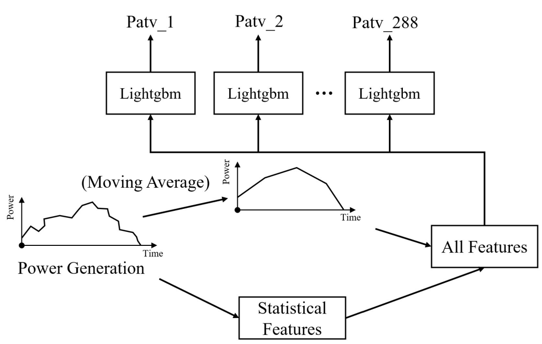


Figure 5: The process of wind power prediction with LightGBM.

Besides building a deep learning model with PaddlePaddle to predict future wind power generation, we also build the LightGBM model to reduce prediction variance and increase generability by ensembling different models. Inspired by the literature [5], the future power generation of each turbine at every time is predicted by an individual model. This prediction approach needs to build 38592(288×134) models in total, which is unacceptable due to the large model size overall. Thus, we reduce the number of models in the following two ways:

- Auxiliary wind turbines: We first use the k-shape algorithm to cluster the wind turbines into 39 classes based on the power generation data and a LightGBM model to predict the class of wind turbines. Then, we can use "auxiliary turbines" data, which is defined as the group of turbines with the closest power generation characteristics to this particular turbine. We can average the power generated by the turbines in the cluster and the power generated by the auxiliary turbines as the model's input to mitigate the effect of data noise.
- Downsampled power generation at time scale: The time resolution of the data is 10min, and the amount of change in power generation data for two adjacent times is insignificant. Therefore, the power generated at N adjacent moments should not differ significantly and can share one prediction value from the same model.

To reduce the dramatic fluctuations in historical power data, focus on data trends, and decrease input data dimensions, historical power data is processed by moving average with the window size of 6 and stride of 3.

Finally, we concatenate historical power generation(after moving average) with statistical features (from section A.1) as the input of LightGBM. Figure 5 illustrates the process of predicting future power generation with LightGBM.

3.5 Model Fusion Strategy

The model ensemble plays an important role in our scheme. We have tried two different ensemble strategies in this competition:

Hybrid Fusion The first ensemble method uses a hybrid fusion method combining both mean and median. In the competition, we need to predict the wind power in the next two days, a total of 288 timestamps. When analyzing the prediction results of different models, we find some pretty different predictions in some timestamps between them. Because the mean fusion is sensitive to outliers that may reduce the fusion performance, we adopt a hybrid strategy based on both mean and median. Here we introduce a threshold α. For timestamps with larger differences than α in the prediction results, we adopt the mean fusion method; otherwise, the median fusion method is adopted.

Weighted Fusion Another method is to use linear weighting to ensemble the models. However, manually assigning weights not only consumes too much time but also may cause an overfitting problem. Therefore, we use ridgeCV to automatically calculate the models' weights and add appropriate regularizations to avoid overfitting.

3.6 Exploratory Methods

In addition to the main methods above, we also explore and utilize a few other methods, two of which are highlighted below.

Mixed Loss Since a large amount of invalid data in the dataset, these data need to be filtered out, and corresponding loss should be ignored. Based on this, we mainly use MseLoss as follows,

$$L_{mse} = \frac{1}{2k}\sum_{i=0}^{k}(y_i^{'} - y_i)^2. \tag{4}$$

where $y_i^{'}$ is i-th unfiltered predicted value, y_i is i-th unfiltered true value and k is the total number.

In addition, since the competition metric is the average of rmse and mae, we also tried other losses, such as L1loss as follows,

$$L_{l1} = \frac{1}{k}\sum_{i=0}^{k}\left|y_i^{'} - y_i\right|. \tag{5}$$

In order to get closer to the competition metrics, we use a Mse-L1 hybrid loss based on the above two losses, which can be written as,

$$L_{total} = L_{mse} + L_{l1}. \tag{6}$$

Models Based on Mean Learning To mitigate the impacts of random fluctuations and abnormal data on model learning, we attempt to model the deviation relative to mathematical expectations of wind power. We analyze the average power of each turbine at each time interval (Figure 6) and find that the power varies widely from one turbine to another, but the power fluctuation of the same turbine at the same time interval is insignificant. Based on this observation, we train the separate models for each turbine to predict the offset to its mean power over the next two days. Specifically, the model takes the average power $\hat{E}(P)$ and monitoring data of the previous 14 days of a particular turbine as input and then predicts the offset and scale for the next two days. Finally, the wind power P can be estimated by the following equation:

$$P_{i,k} = \alpha_k \hat{E}(P_i) + \beta_k, \tag{7}$$

where $P_{i,k}$ denotes the power generation of the ith turbine on the kth day, α_k and β_k denote the scale factor and offset at the kth day, respectively.

4 EXPERIMENTS

We have achieved steady improvements online using data cleaning, feature engineering, different models, training tricks, and fusion strategies. Compared to Phase 1, Phase 2 and Phase 3 are more representative. The results of Phase 2 are shown in Table 2. Feature screening from 1 to 3 is to delete Prtv, delete Pab1 to Pab3, and delete Tmstamp, Wdir, Ndir, Etmp, and Itmp, respectively. Based on the baseline, we increased the days of the training set, modified the training parameters, expanded the model width, and set the input length to 432, which constituted the model of the basic version of phase 2 (phase 2 basic). On the basic model, adding feature screening and data cleaning has improved the score by 0.26397, then further improved by 0.09275 through model fusion and tuning of the training hyperparameters, resulting in the overall improvement by 0.35672 and top 10 ranking in phase 2.

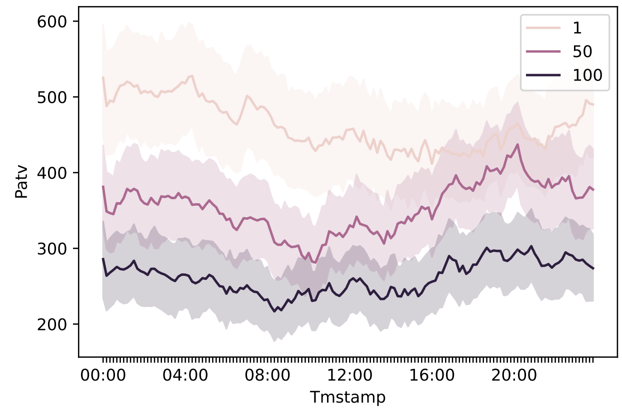


Figure 6: Illustration of the expected value of wind power of turbines 1, 50 and 100. The solid line indicates the expectation of the turbine power and the colored area indicates the 95% confidence interval.

Table 3 shows the results of Phase 3, where the base version model is the optimal model combination of Phase 2. Following the model structure in Section 3.3, we add a variety of new models with different hyperparameters and features. From Table 3, data cleaning, feature filtering, auto-correlation mechanism, and model fusion can consecutively improve the score about 0.2, 0.032, 0.05, and 0.07. With the help of all these techniques, the eventual online score is improved to -45.13867.

The steady improvement of online experiments proves the effectiveness of our modeling techniques, including data cleaning, feature screening, model design, and fusion strategies. In addition, due to the limited submission time, our best score model could not be submitted online at the last moment.

5 CONCLUSIONS

This paper proposes an excellent framework for the problems and challenges of wind power forecasting. We first design data cleaning and correction based on the wind power curve and complete feature screening. Then we develop two integrated, highly accurate models specifically for wind power scenarios. One is the WPFormer which uses an auto-correlation mechanism to learn the information of sequence-level relationships, and a spatio-temporal graph neural network encodes multi-turbine information. The other one is the POPtree which uses a bunch of LightGBM models to predict the wind power of the individual turbine. Furthermore, we propose two effective model fusion methods to ensemble these two models. Eventually, the steady improvement of online scores shows the excellent performance of our solution.

ACKNOWLEDGMENTS

This competition has been successfully completed, thanks to Baidu platform for its strong support during the competition and thanks to Longyuan Power Company for the precious data.

Table 2: Results of Phase 2.

Serial No.	Model Components	Online Score	Offline Score
1	baseline	-	-46.83
2	phase 2 basic	-44.72740	-43.78076
3	+ feature screening1	-44.60840	-43.59530
4	+ model fusion	-44.58287	-43.63643
5	+ feature screening2 + data cleaning	-44.46343	-43.36928
6	+ Change the training parameters	-44.40641	- 43.26067
7	+ feature screening3	-44.39223	-43.28473
8	+ model fusion + Change the training parameters	-44.37068	-43.31144
9	+ POPtree	**-44.32841**	-43.28473

Table 3: Results of Phase 3.

Serial No.	Model Components	Online Score	Offline Score
1	phase 3 basic	-45.48641	-43.33292
2	+ data cleaning	-45.28938	-43.31111
3	+ feature screening	-45.25772	-43.40196
4	+ autocorrelation mechanism	-45.20941	-43.10517
5	+ model fusion	**-45.13867**	-43.10917
6	Add the POPtree based on model No.2	-45.22724	43.35802

REFERENCES

[1] Amirhossein Ahmadi, Mojtaba Nabipour, Behnam Mohammadi-Ivatloo, Ali Moradi Amani, Seungmin Rho, and Md. Jalil Piran. 2020. Long-Term Wind Power Forecasting Using Tree-Based Learning Algorithms. *IEEE Access* 8 (2020), 151511–151522. https://doi.org/10.1109/ACCESS.2020.3017442

[2] MC Alexiadis, PS Dokopoulos, HS Sahsamanoglou, and IM Manousaridis. 1998. Short-term forecasting of wind speed and related electrical power. *Solar Energy* 63, 1 (1998), 61–68.

[3] Barbara G Brown, Richard W Katz, and Allan H Murphy. 1984. Time series models to simulate and forecast wind speed and wind power. *Journal of Applied Meteorology and Climatology* 23, 8 (1984), 1184–1195.

[4] Maria Grazia De Giorgi, Stefano Campilongo, Antonio Ficarella, and Paolo Maria Congedo. 2014. Comparison between wind power prediction models based on wavelet decomposition with least-squares support vector machine (LS-SVM) and artificial neural network (ANN). *Energies* 7, 8 (2014), 5251–5272.

[5] Shereen Elsayed, Daniela Thyssens, Ahmed Rashed, Lars Schmidt-Thieme, and Samer Hadi Jomaa. 2021. Do We Really Need Deep Learning Models for Time Series Forecasting? (2021).

[6] James D Evans. 1996. *Straightforward statistics for the behavioral sciences.* Thomson Brooks/Cole Publishing Co.

[7] Umut Firat, Seref Naci Engin, Murat Saraclar, and Aysin Baytan Ertuzun. 2010. Wind Speed Forecasting Based on Second Order Blind Identification and Autoregressive Model. In *2010 Ninth International Conference on Machine Learning and Applications*. 686–691. https://doi.org/10.1109/ICMLA.2010.106

[8] Qinkai Han, Fanman Meng, Tao Hu, and Fulei Chu. 2017. Non-parametric hybrid models for wind speed forecasting. *Energy Conversion and Management* 148 (2017), 554–568.

[9] Xiaohui He, Ying Nie, Hengliang Guo, and Jianzhou Wang. 2020. Research on a Novel Combination System on the Basis of Deep Learning and Swarm Intelligence Optimization Algorithm for Wind Speed Forecasting. *IEEE Access* 8 (2020), 51482–51499.

[10] Yun Ju, Guangyu Sun, Quanhe Chen, Min Zhang, Huixian Zhu, and Mujeeb Ur Rehman. 2019. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting. *Ieee Access* 7 (2019), 28309–28318.

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017)* (2017), 3149–3157.

[12] Hui Liu, Xiwei Mi, and Yanfei Li. 2018. Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM. *Energy Conversion and Management* 159 (2018), 54–64.

[13] Tony Fei Liu, Ming Kai Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. *ICDM* (2008), 413–422.

[14] Zhengjing Ma and Gang Mei. 2022. A hybrid attention-based deep learning approach for wind power prediction. *Applied Energy* 323 (2022), 119608.

[15] Dr M Madhiarasan, Prof Roy, and Partha Pratim. 2022. Hybrid Transformer Network for Different Horizons-based Enriched Wind Speed Forecasting. *arXiv preprint arXiv:2204.09019* (2022).

[16] Mohammad A Mohandes, Talal O Halawani, Shafiqur Rehman, and Ahmed A Hussain. 2004. Support vector machines for wind speed prediction. *Renewable energy* 29, 6 (2004), 939–947.

[17] Thi Hoai Thu Nguyen and Quoc Bao Phan. 2022. Hourly day ahead wind speed forecasting based on a hybrid model of EEMD, CNN-Bi-LSTM embedded with GA optimization. *Energy Reports* 8 (2022), 53–60.

[18] Dongxiao Niu, Lijie Sun, Min Yu, and Keke Wang. 2022. Point and interval forecasting of ultra-short-term wind power based on a data-driven method and hybrid deep learning model. *Energy* 254 (2022), 124384.

[19] George Sideratos and Nikos D Hatziargyriou. 2007. An advanced statistical method for wind power forecasting. *IEEE Transactions on power systems* 22, 1 (2007), 258–265.

[20] Jingjing Song, Jianzhou Wang, and Haiyan Lu. 2018. A novel combined model based on advanced optimization algorithm for short-term wind speed forecasting. *Applied Energy* 215 (2018), 643–658.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[22] Huai-zhi Wang, Gang-qiang Li, Gui-bin Wang, Jian-chun Peng, Hui Jiang, and Yi-tao Liu. 2017. Deep learning based ensemble approach for probabilistic wind power forecasting. *Applied energy* 188 (2017), 56–70.

[23] Yun Wang, Runmin Zou, Fang Liu, Lingjun Zhang, and Qianyi Liu. 2021. A review of wind speed and wind power forecasting with deep neural networks. *Applied Energy* 304 (2021), 117766.

[24] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* 34 (2021), 22419–22430.

[25] Ruiguo Yu, Jie Gao, Mei Yu, Wenhuan Lu, Tianyi Xu, Mankun Zhao, Jie Zhang, Ruixuan Zhang, and Zhuo Zhang. 2019. LSTM-EFG for wind power forecasting based on sequential correlation features. *Future Generation Computer Systems* 93 (2019), 33–42.

[26] Yixiao Yu, Xueshan Han, Ming Yang, and Jiajun Yang. 2020. Probabilistic Prediction of Regional Wind Power Based on Spatiotemporal Quantile Regression. *IEEE Transactions on Industry Applications* 56, 6 (2020), 6117–6127.

[27] Kalid Yunus, Torbjörn Thiringer, and Peiyuan Chen. 2015. ARIMA-based frequency-decomposed modeling of wind speed time series. *IEEE Transactions on Power Systems* 31, 4 (2015), 2546–2556.
[28] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11106–11115.
[29] Jingbo Zhou, Xinjiang Lu, Yixiong Xiao, Jiantao Su, Junfu Lyu, Yanjun Ma, and Dejing Dou. 2022. SDWPF: A Dataset for Spatial Dynamic Wind Power Forecasting Challenge at KDD Cup 2022. *Techincal Report* (2022).
[30] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. 27268–27286.
[31] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. *International Conference on Machine Learning* (2022), 27268–27286.

A APPENDIX

A.1 Statistical Features

Table 4: Statistical Features Part 1.

No.	Implication	No.	Implication
1	mean	2	standard deviation
3	maximum	4	skewness
5	minimum	6	kurtosis
7	absolute energy	8	autoregressive coefficient
9	autocorrelation	10	approximate entropy
11	c3 statistics	12	complex of time series

Table 5: Statistical Features Part 2.

No.	Implication
13	quantile (0.1, 0.25, 0.5, 0.75, 0.9)
14	the sum over the absolute value of consecutive changes.
15	augmented Dickey-Fuller test
16	the number of values in the time series that are higher than the mean of the time series.
17	the length of the longest consecutive subsequence in the series that is bigger than the mean of the time series.
18	the length of the longest consecutive subsequence in the series that is smaller than the mean of the time series.
19	the mean over the absolute differences between subsequent series values.
20	the number of peaks in the time series.
21	the sum of all data points that are present more than once.