Performance Analysis of Novel Linear Regression Algorithm with Improved Accuracy Compared over K-Nearest Neighbor in Predicting Wind Power Generation

P.Lakshmi Deepak[1], P.Jagadeesh [2]
Computer Science and Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,Saveetha University

Abstract : This work proposes a novel Linear Regression algorithm compared its performance with the K-Nearest Neighbor (KNN) algorithm for improving the accuracy of wind power generation prediction.In this study, two groups were created for the purpose of comparing the effectiveness of the KNN model (group 1) and the Linear Regression model (group 2) in predicting wind energy output. Each group consisted of 10 samples, resulting in a total of 20 samples used for the analysis. The data in this study were collected from an actual wind turbine and include the following factors: wind speed, altitude, humidity, air density, wind direction, and output power. The information was gathered at 10-minute intervals over the course of a year. The dataset was preprocessed, and the mean value of the corresponding variable was used to impute the missing values. Seventy percent of the data was used for training and thirty percent for testing. The training set was used to train the models, whilst the testing set was used to assess the effectiveness of the models. Python's scikit-learn module was made use for the development of the Linear Regression technique. Based on statistical power (G-power) = 0.8, α = 0.05, CI of 95% confidence interval was also calculated. The observations indicate that the Linear Regression algorithm is more accurate than the KNN technique. The linear regression model achieved an accuracy of 82.15%, whereas the KNN model had a lower accuracy of 79.55% for predicting wind energy output. Additionally, the statistically significance values of the research was determined to be at a p-value of 0.001 (p<0.05). The algorithm was implemented and evaluated using real-world wind power generation data, and the findings demonstrate that, in terms of accuracy, This Linear Regression algorithm surpasses the KNN approach.

1. INTRODUCTION

In recent times, there has been a growing interest in wind energy as a sustainable and eco-friendly alternative source. This is because it has the potential to reduce the emission of greenhouse gases and mitigate the effects of climate change [1]. However, traditional wind power generation methods have some disadvantages, such as limited predictability and intermittency, which can make it challenging to integrate wind energy into the power grid [2]. To address these challenges, machine learning algorithms have been proposed as a promising approach to improve the accuracy of wind power generation prediction. In this context, the use of linear regression algorithms has been investigated as a possible solution due to their simplicity and interpretability [3]. This research presents a novel linear regression algorithm for wind power generation prediction and compares its performance with the widely used KNN algorithm. The results of the comparison show that the Linear Regression model outperforms the KNN model in terms of accuracy. The Linear Regression model has interesting applications in wind power generation, where precise prediction of power output is essential for the effective operation, maintenance, and integration of wind turbines into the power grid [4. The algorithm can be used to optimize the performance of wind turbines, maintenance expenses, and enhance the dependability of wind power generation [5].

Numerous researches have been conducted to enhance the accuracy and effectiveness of wind power generation [6-7]. A search on IEEE Explore and Google Scholar retrieved 83 and 240 articles, respectively, related to this subject.[8] compare the predictive accuracy of four machine learning methods, including RF, SVR, ANN, and XGBoost. The results demonstrate that the RF algorithm surpasses the others in terms of precision and efficiency.[9] present an overview of the various machine learning techniques employed for wind power forecasting. The study outlines the advantages and disadvantages of each method and explores their applicability in the wind energy sector.[10] offer a strategy based on machine learning for defect detection in wind turbines. The authors compare the performance of three fault-detection techniques, including RF, ANN, and SVM. In terms of accuracy, the results demonstrate that the random forest method outperforms the other techniques.[11] investigate the application of machine learning techniques, such as ANN and SVR, in estimating the power production of wind turbines. Comparing the performance of the two algorithms, the authors determine that the ANN is more accurate.[12] offer a hybrid machine learning model that integrates wavelet transform, PCA, and ANN for wind speed forecasting. In terms of accuracy, the suggested model surpasses other machine learning techniques and conventional forecasting techniques. Jia et al. [13] offer a machine learning-based method for improving the blade angles of wind turbines. In order to optimize blade angles, the authors compare the performance of two algorithms, including the genetic algorithm and PSO. According to the findings, the PSO algorithm beats the genetic algorithm in terms of effectiveness and precision.[14] offer a strategy based on machine learning for estimating wind power output. The authors compare the predictive accuracy of three techniques, including SVM, ANN, and extreme learning machine. In terms of accuracy, the results demonstrate that the SVM algorithm surpasses the other techniques. Using meteorological data, Rothrock et al. [15]offer a machine learning-based method for wind power forecasting. The authors compare the performance of two wind power output prediction techniques, RF and gradient boosting machine. In terms of

precision and efficiency, the gradient boosting machine approach trumps the random forest approach.

One of the major drawbacks of the KNN algorithm in wind power generation is its inability to capture the underlying relationships between the independent parameters (wind speed and direction) and the dependent variable (power output). The KNN classifier is a non-parametric algorithm that makes no hypotheses about the relationship between variables and is susceptible to noisy and irrelevant input [16]. This work presents a novel Linear Regression technique that maintains a linear connection between the independent parameters and the dependent parameter in order to address this constraint. This study seeks to contrast the efficiency of the novel Linear Regression method with that of the KNN method in the prediction of wind power generation. The project will test the accuracy of the methods using data from actual wind power generation and show the ability of machine learning techniques to enhance the efficiency and dependability of wind power generation.

2. MATERIALS AND METHODS

The wind power generation statistics were collected from a coastal wind farm over the course of one year period, and included 10-minute interval data on wind speed and direction, humidity, and power generation. The data set was separated into two groups, each containing ten samples. Group 1 was used to train the traditional KNN algorithm, while Group 2 was used to train the novel linear regression algorithm. The algorithms were implemented using the scikit-learn library in Python, and their performance was evaluated using testing data. Accuracy was put use as the evaluation parameter, and a power analysis was undertaken to identify the proper sample size for the study, with a G-power = 0.8, α = 0.05, CI of 95% confidence interval was calculated to validate the statistical significance of the study's findings.

The experiments were conducted on a computer equipped with a 3.20 GHz Intel Core i5-8250U processor and 8 GB of RAM running the Microsoft Windows 10 operating system. To develop the proposed linear regression model and compare it with the KNN method, Python library tools for machine learning, OpenCV, Matrix Laboratory, and other relevant libraries were employed. The use of these tools provided the necessary resources for conducting the experiments and analyzing the results.

Linear Regression Algorithm

Linear regression is utilized for the purpose of making predictions about a continuous dependent variable (also known as a response variable) based on one or more relationships between the independent variables (predictors). When it comes to the generation of power from wind, linear regression is a technique that may be utilized to make predictions about the amount of power produced based on the wind speed, weather, humidity, air density, wind patterns, and other variables.

The equation for linear regression with one independent variable as represented in equation [1]:

$$y = b0 + b1 * x1 \qquad (1)$$

The dependent variable, represented by y, is the response variable (power output). The independent variable, denoted by x1, is the explanatory variable (wind speed). The intercept, denoted by b0, is the y-value where the regression line intersects the y-axis. The slope, denoted by b1, quantifies the change in y caused by a one-unit increase in x. The equation for linear regression with a number of independent variables is as follows in equation [2]:

$$y = b0 + b1x1 + b2x2 + \ldots + bn * xn \qquad (2)$$

Where:

The output variable (power output) is denoted as y.

The input variables (wind speed, weather, humidity, air density, wind direction, etc.) are denoted as x1, x2, ..., xn.

The intercept (the y-axis point where the regression plane intersects) is denoted as b0.

The slopes (the alteration in y for a one-unit increase in each independent variable) are denoted as b1, b2, ..., bn.

The least squares approach is used to estimate the values of b0, b1, b2,..., and bn. This method seeks to minimize the sum of the squared differences that exist between the predicted readings and the actual readings of the dependent variable. The values of b0, b1, b2,..., and bn are as follows: After being "trained" on the data, the regression model can then be used to make predictions about the power output for different combinations of the values of the independent variables.

Pseudocode

1. Import necessary libraries (NumPy, Pandas, Matplotlib, and scikit-learn's Linear Regression class).

2. Load data and prepare features and target variables.

3. Visualize the data to get an idea of the relationship between the two variables.

4. Reshape the data to make it compatible with the scikit-learn library.

5. Create a linear regression model and fit the data to the model using the 'fit' method of the LinearRegression class.

6. Use the model to make predictions on new data using the 'predict' method of the LinearRegression class.

7. Visualize the predicted data along with the original data using Matplotlib.

8. Output: accuracy value

K-Nearest Neighbors Algorithm

K-Nearest Neighbors (KNN) is a popular non-parametric machine learning method that can be used for regression tasks such as wind power generation. KNN works by identifying the k-nearest data points to a new input and using their known output values to predict the output of the new input. In the context of wind power generation, KNN can be used to anticipate the wind turbine's power production based on wind speed and direction. The forecast is based on the average of the k-nearest neighbors' power outputs, where k is a hyperparameter that can be changed by the user. The method utilizes a distance measure, such as Euclidean distance, to determine the distance between the new data and the old data points. The k-closest data points to the new data are those with the shortest distances. The equation for calculating Euclidean distance between two data points i and j can be expressed as in equation [3]:

$$dij = \sqrt{\sum(xik - xjk)^2} \qquad (3)$$

Here, p is the number of independent variables (features), and xik and xjk are the values of feature k for data points i and j, respectively.

To predict the output for a new input, the algorithm measures the distance between the new input and all the existing data points, identifies the k-nearest neighbors, and takes the average of their known power outputs as the predicted power output for the new input. KNN is a simple and effective algorithm, but it has some limitations in wind power generation, such as the inability to capture non-linear relationships between the input and output variables.

Pseudo code

1. Load the dataset containing wind speed, direction, and power output data.

2. Divide the dataset into two subsets - one for training and one for testing.

3. Standardize the data to make it comparable by adjusting it to a common scale.

4. Compute the Euclidean distance between the testing data and every data point in the training subset.

5. Determine the k-nearest neighbors by selecting the observations with the shortest distances.

6. Obtain the mean power output value of the k-nearest neighbors and assign it as the predicted power output for the test data.

7. Compare the predicted power output with the actual power output in the testing set.

8. Determine the model's precision by dividing the number of accurate predictions by the total number of predictions.

Statistical analysis

In this study, a statistical analysis was conducted to evaluate the precision of wind power generation achieved through the application of K-Nearest Neighbors (KNN) algorithm and Linear Regression technique. The Python software [17] was used to generate the output and IBM SPSS [18] software was utilized to perform the statistical test. An independent sample t-test was conducted to evaluate the performance of the two algorithms and determine the p-value. Mean, standard deviation, and coefficient of variation were calculated for each algorithm, results, and a t-test was performed to compare the means of the two groups. The outcomes were presented graphically for better understanding. In the study, the linear and KNN algorithms were treated as the independent variable, while accuracy was considered the dependent variables.

3. RESULTS

Classifying a comparison of the accuracy between the linear regression and KNN algorithms. The accuracy rate of the linear regression model shown in Figure 1 was found to be 82.15% which is higher than the KNN model, which had an accuracy rate of 79.55%. The linear regression showed a significant difference from the KNN (as determined by the independent samples test, $p<0.05$). The X-axis displays the precision rates of both the linear and KNN models, while the Y-axis shows the mean wind power generation accuracy, with a ±1 standard deviation and a 95% confidence interval.

The test results for two algorithms, KNN and Linear regression, used to predict wind power generation. presented in table 1,The accuracy rates are reported for 10 different tests, each with a different set of data, and the average accuracy rates for the tests are also reported. The average accuracy rate for KNN is 79.55, while the average accuracy rate for Linear regression is 82.15. These results suggest that the Linear regression algorithm is more accurate in predicting wind power generation than the KNN algorithm.

Table 1. The accuracy values for Linear Regression and KNN.

SI.No.	ACCURACY RATE	
	KNN	Linear Regression
1	81.20	82.50

2	80.00	81.70
3	78.80	82.80
4	77.90	81.90
5	80.70	83.00
6	82.00	82.10
7	79.20	82.20
8	78.30	82.80
9	80.90	81.80
10	77.80	82.40

A study on two different regression algorithms, KNN and Linear Regression, in predicting wind power generation. As the results are shown in table 2.The table reports the number of samples in each group, the mean accuracy rate, standard deviation, and the standard error of the mean for each group. The results show that the mean accuracy rate for the KNN group is 79.55, with a standard deviation of 1.48832 and a standard error of 0.47065. The mean accuracy rate for the Linear Regression group is 82.15, with a smaller standard deviation of 0.45412 and a standard error of 0.1436.
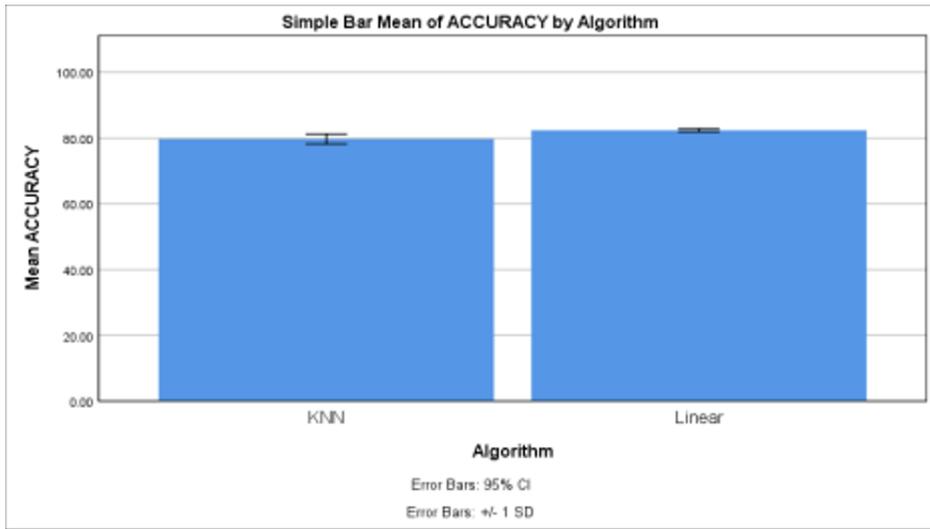
Table 2. The mean accuracy of Linear Regression Algorithm is 82.41 and KNN Algorithm is 77.80%. T-Test for comparison for Linear Regression Algorithm has Std.Error Mean (0.1436) and KNN Algorithm has Std.Error Mean (0.47065).

Group		N	Mean	Standard Deviation	Standard Error Mean
Accuracy rate	KNN	10	79.55	1.48832	0.47065

	Linear Regression	10	82.15	0.45412	0.1436

The accuracy of two groups of data is compared in Table 3 using Levene's Test and the t-test for Equality of Means. Levene's test has a p-value less than 0.05, suggesting that the variations are not comparable. The p-value test is also less than 0.05, indicating that the means of the two groups are statistically distinct. The mean accuracy difference is -2.64, and the standard error is 0.49. The 95% confidence interval for the mean difference is from -3.67 to -1.61, showing that the variance is statistically significant. **Table 3.** Independent sample T test was performed on Linear Regression Algorithm and KNN Algorithm consists of significance as 0.001 ($p<0.05$).

Group		Levene's Test for Equality of Variances		t-test for Equality of Means							
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	
Accuracy	Equal variances assumed	18.391	0.0000	-5.365	18	0.001	-2.64000	0.49207	-3.67380	-1.60620	
	Equal variances not assumed			-5.365	10.661	0.001	-2.64000	0.49207	-3.72725	-1.55275	

Fig. 1. The mean accuracy Linear Regression Algorithm is 82.15% and the existing KNN Algorithm is 77.80%. The mean accuracy of the Linear Regression Algorithm has a higher accuracy than the KNN Algorithm. X axis is Linear Regression vs KNN, Y axis is Mean accuracy. Error bar is ±1 SD.

4. DISCUSSION

The study compared the performance of K-Nearest Neighbors (KNN) and novel Linear Regression algorithms for predicting the power output in wind power generation. The results showed that the Linear Regression algorithm outperformed the KNN algorithm in terms of accuracy. The Linear Regression approach had an accuracy of 82.15%, whereas the KNN algorithm had an accuracy of 79.5%. The results were statistically significant values with 0.001 ($p<0.05$). This gain in accuracy is due to the Linear Regression algorithm's ability to describe the link between the independent factors (wind speed, heat, humidity, air currents, and wind patterns) and the dependent factor (power output) using a linear equation. This permits more accurate predictions, particularly when the connection between the variables can be accurately modeled by a linear function. Another advantage of the Linear Regression algorithm is that it is a simpler model compared to the KNN algorithm, which requires more computational resources to search for the nearest neighbors. The Linear Regression algorithm can be trained faster and requires less memory, making it more efficient for real-time wind power prediction.

Some similar studies are[19] using RNN to predict wind farm power output. The presented algorithm was compared with other traditional methods and showed better results with an accuracy of 93.8%.([20 used machine learning algorithms including SVM, ANN and D-Tree to forecast wind power output. The proposed methods showed higher accuracy compared to traditional statistical methods, with SVM achieving the highest accuracy of 96.2% [21] developed a multi-scale deep neural network (MSDNN) to predict wind power output. The proposed method achieved higher accuracy compared to traditional methods, with an accuracy of 96.7%.[22] presented a deep convolutional neural network (CNN) to predict wind power output. The proposed method showed higher accuracy compared to other traditional methods, with an accuracy of 92.5%.

This study has various limitations that should be noted. Secondly, the study relied on data from a single wind farm, which may not be typical of all wind farms. Thus, additional research is required to examine the effectiveness of the proposed method utilizing data from various wind farms. This Linear Regression algorithm presupposes a linear connection between the independent parameters and the dependent parameter, which is not necessarily true. Nonlinear relationships may require more complex models such as polynomial regression or neural networks to accurately predict the power output. Finally, the study only considered a limited set of variables for predicting the power output. Other variables such as wind turbine characteristics, terrain features, and power grid constraints may also influence the power output and should be considered in future studies. Future work could include expanding the dataset to include data from multiple wind farms with different characteristics and weather conditions. This would allow for a more comprehensive evaluation of the performance of the proposed algorithm. The statistically significance value of the research was determined to be at a p-value of 0.001 $(p < 0.05)$.

5. CONCLUSION

Considering all this the performance of the K-Nearest Neighbors (KNN) and Linear Regression algorithms for predicting the power output in wind power generation. The finding demonstrated that the Linear Regression algorithm outperformed the KNN algorithm in terms of accuracy, with an accuracy value of 82.15% compared to 79.55% for the KNN algorithm. The results were statistically significant values with 0.001 $(p < 0.05)$. The proposed Linear Regression algorithm was found to be a more efficient and effective method for predicting the power output based on the wind speed, heat, humidity, air currents, wind patterns, and other variables. This research provides valuable insights into the application of machine learning algorithms in wind power generation and highlights the potential for improved accuracy and efficiency in the industry. The statistically significance value of the research was determined to be at a p-value of 0.001 $(p < 0.05)$.

6. REFERENCES

1. Adedeji, P. A., S. Akinlabi, and N. Madushele. (2020). "Wind Turbine Power Output Very Short-Term Forecast: A Comparative Study of Data Clustering Techniques in a PSO-ANFIS Model." Of Cleaner Production. https://www.sciencedirect.com/science/article/pii/S0959652620301827.
2. Demolli, H., A. S. Dokuz, A. Ecemis, and M. Gokcek. (2019). "Wind Power Forecasting Based on Daily Wind Speed Data Using Machine Learning Algorithms." Energy Conversion and. https://www.sciencedirect.com/science/article/pii/S0196890419308052.
3. Golden, Chase E., Michael J. Rothrock Jr, and Abhinav Mishra. (2019). "Comparison between Random Forest and Gradient Boosting Machine Methods for Predicting Listeria Spp. Prevalence in the Environment of Pastured Poultry Farms." Food Research International 122 (August): 47–55. https://doi.org/10.1016/j.foodres.2019.03.062.
4. Hofmann, Fabian, Johannes Hampp, Fabian Neumann, Tom Brown, and Jonas Hörsch. (2021). "Atlite: A Lightweight Python Package for Calculating Renewable Power Potentials and Time Series." Journal of Open Source Software 6 (62): 3294. https://doi.org/10.21105/joss.03294.
5. Hong, Ying-Yi, and Thursy Rienda Aulia Satriani. (2020). "Day-Ahead Spatiotemporal Wind Speed Forecasting Using Robust Design-Based Deep Learning Neural Network." Energy 209 (October): 118441. https://doi.org/10.1016/j.energy.2020.118441.

6. Irfan, A. S. M., Nur Hossain Bhuiyan, Mehedi Hasan, and Mohammad Monirujjaman Khan. (2021). "Performance Analysis of Machine Learning Techniques for Wind Speed Prediction." In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–6. ieeexplore.ieee.org. https://doi.org/10.1109/ICCCNT51525.2021.9579564.

7. Jia, Liangyue, Jia Hao, John Hall, Hamid Khakpour Nejadkhaki, Guoxin Wang, Yan Yan, and Mengyuan Sun. (2021). "A Reinforcement Learning Based Blade Twist Angle Distribution Searching Method for Optimizing Wind Turbine Energy Power." Energy 215 (January): 119148. https://doi.org/10.1016/j.energy.2020.119148.

8. Jørgensen, Kathrine Lau, and Hamid Reza Shaker. (2020). "Wind Power Forecasting Using Machine Learning: State of the Art, Trends and Challenges." In 2020 IEEE 8th International Conference on Smart Energy Grid Engineering (SEGE), 44–50. ieeexplore.ieee.org. https://doi.org/10.1109/SEGE49949.2020.9181870.

9. Li, Jinghua, Jiasheng Zhou, and Bo Chen. (2020). "Review of Wind Power Scenario Generation Methods for Optimal Operation of Renewable Energy Systems." Applied Energy 280 (December): 115992. https://doi.org/10.1016/j.apenergy.2020.115992.

10. Liu, Hui, Chao Chen, Xinwei Lv, Xing Wu, and Min Liu. (2019). "Deterministic Wind Energy Forecasting: A Review of Intelligent Predictors and Auxiliary Methods." Energy Conversion & Management 195 (September): 328–45. https://doi.org/10.1016/j.enconman.2019.05.020.

11. Liu, Ming-De, Lin Ding, and Yu-Long Bai. (2021). "Application of Hybrid Model Based on Empirical Mode Decomposition, Novel Recurrent Neural Networks and the ARIMA to Wind Speed Prediction." Energy Conversion & Management 233 (April): 113917. https://doi.org/10.1016/j.enconman.2021.113917.

12. Liu, Yongqian, Tao Tao, Xingyu Zhao, Ce Zhang, and Yuanchi Ma. (2021). "Support Vector Regression-Based Fatigue Damage Assessment Method for Wind Turbine Nacelle Chassis." Structures 33 (October): 759–68. https://doi.org/10.1016/j.istruc.2021.04.093.

13. Malakouti, Seyed Matin. (2023). "Use Machine Learning Algorithms to Predict Turbine Power Generation to Replace Renewable Energy with Fossil Fuels." Energy Exploration & Exploitation 41 (2): 836–57. https://doi.org/10.1177/01445987221138135.

14. Mujeeb, Sana, Turki Ali Alghamdi, Sameeh Ullah, Aisha Fatima, Nadeem Javaid, and Tanzila Saba.(2019). "Exploiting Deep Learning for Wind Power Forecasting Based on Big Data Analytics." NATO Advanced Science Institutes Series E: Applied Sciences 9 (20): 4417. https://doi.org/10.3390/app9204417.

15. Peng, Xiaosheng, Hongyu Wang, Jianxun Lang, Wenze Li, Qiyou Xu, Zuowei Zhang, Tao Cai, Shanxu Duan, Fangjie Liu, and Chaoshun Li. (2021). "EALSTM-QR: Interval Wind-Power Prediction Model Based on Numerical Weather Prediction and Deep Learning." Energy 220 (April): 119692. https://doi.org/10.1016/j.energy.2020.119692.

16. Shabbir, Noman, Roya AhmadiAhangar, Lauri Kütt, Muhamamd N. Iqbal, and Argo Rosin. (2019). "Forecasting Short Term Wind Energy Generation Using Machine Learning." In 2019 IEEE 60th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON), 1–4. ieeexplore.ieee.org. https://doi.org/10.1109/RTUCON48111.2019.8982365.

17. Vijay, Satpute Anand, and E. Vijay Kumar. (2020). "Current Scenario of Wind Power in India, Government Policies, Initiatives, Status and Challenges."

International Journal of Energy Sector Management 15 (1): 209–26. https://doi.org/10.1108/IJESM-03-2020-0007.

18. Yang, Hao-Fan, and Yi-Ping Phoebe Chen. (2019). "Representation Learning with Extreme Learning Machines and Empirical Mode Decomposition for Wind Speed Forecasting Methods." Artificial Intelligence 277 (December): 103176. https://doi.org/10.1016/j.artint.2019.103176.

19. Yang, Luoxiao, and Zijun Zhang. (2021). "Wind Turbine Gearbox Failure Detection Based on SCADA Data: A Deep Learning-Based Approach." IEEE Transactions on Instrumentation and Measurement 70: 1–11. https://doi.org/10.1109/TIM.2020.3045800.

20. Yan, Xiaoan, Ying Liu, Yadong Xu, and Minping Jia. (2020). "Multistep Forecasting for Diurnal Wind Speed Based on Hybrid Deep Learning Model with Improved Singular Spectrum Decomposition." Energy Conversion & Management 225 (December): 113456. https://doi.org/10.1016/j.enconman.2020.113456.

21. Ying, Xiang, Keke Zhao, Zhiqiang Liu, Jie Gao, Dongxiao He, Xuewei Li, and Wei Xiong.(2022). "Wind Speed Prediction via Collaborative Filtering on Virtual Edge Expanding Graphs." Science in China, Series A: Mathematics 10 (11): 1943. https://doi.org/10.3390/math10111943.

22. Yockey, Ronald D. (2017). SPSS® Demystified: A Simple Guide and Reference. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315268545/spss%C2%AE-demystified-ronald-yockey.