Clean Technologies and Recycling

https://www.aimspress.com/journal/ctr

Clean Technologies and Recycling, 4(2): 108–124.
DOI: 10.3934/ctr.2024006
Received: 27 September 2024
Revised: 17 November 2024
Accepted: 21 November 2024
Published: 02 December 2024

Research article

Forecasting power generation of wind turbine with real-time data using machine learning algorithms

Asiye Bilgili[1],* and Kerem Gül[2]

[1] Department of Informatics, Institute of Graduate Studies in Sciences, Istanbul University, Istanbul, Turkey
[2] Department of Mechanical Engineering, Faculty of Engineering, Halic University, Istanbul, Turkey

* **Correspondence:** Email: asiyetunar@ogr.iu.edu.tr; Tel: +00902129242444.

Abstract: The escalating concern over the adverse effects of greenhouse gas emissions on the Earth's climate has intensified the need for sustainable and renewable energy sources. Among the alternatives, wind energy has emerged as a key solution for mitigating the impacts of global warming. The significance of wind energy generation lies in its abundance, environmental benefits, cost-effectiveness and contribution to energy security. Accurate forecasting of wind energy generation is crucial for managing its intermittent nature and ensuring effective integration into the electricity grid. We employed machine learning techniques to predict wind power generation by utilizing historical weather data in conjunction with corresponding wind power generation data. The dataset was sourced from real-time SCADA data obtained from wind turbines, allowing for a comprehensive analysis. We differentiated this research by evaluating not only wind conditions but also meteorological factors and physical measurements of turbine components, thus considering their combined influence on overall wind power production. We utilized Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost algorithms to estimate power generation. The performance of these models assessed using evaluation criteria: R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The findings indicated XGBoost algorithm outperformed the other models, achieving high accuracy while demonstrating computational efficiency, making it particularly suitable for real-time applications in energy forecasting.

Keywords: forecasting power production; global warming; renewable energy; sustainable energy

1. Introduction

Energy plays a major role in human life and must be utilized with maximal effectiveness, minimal environmental harm, and the lowest cost. For the last few centuries, energy costs have been the main financial outlay for countries; yet, the growing need for fossil fuels has not been satisfied, and the environmental effects of depending on fossil fuels cannot be ignored [1].

There is greater awareness than ever before that the production of carbon dioxide from burning fossil fuels causes a rise in global temperatures, leading to efforts to reduce the use of these fuels. Many agreements have been signed between countries and organizations to achieve specific targets. The Kyoto Protocol, widely recognized as the most prominent among these agreements, was established with the primary aim of mitigating the release of greenhouse gases (GHGs) caused by human activities. This protocol acknowledges the diverse national disparities in GHG emissions, economic prosperity, and the capability to address these emissions. After a series of contentious conferences, participants of the 2015 COP21 summit in Paris reached a consensus to limit the global average temperature increase to a maximum of 2 °C above preindustrial levels. Additionally, they committed to sustaining the current temperature rise, which stands at 1.5 °C higher. This historic agreement effectively replaced the Kyoto Protocol and received signatures from all 196 UNFCCC signatories. To incentivize developing nations to adopt environmentally friendly technologies, the agreement also stipulated a periodic review of progress every five years and the establishment of a $100 billion annual fund by [2,3].

Global greenhouse gas (GHG) emissions have risen since the turn of the 21st century and up until 2019. This trend has primarily been driven by the rising emissions from China and other emerging economies. The natural greenhouse effect, which may negatively impact Earth's life, is enhanced by significantly higher atmospheric concentrations of greenhouse gases. Global emissions decreased by 3.7% in 2020 compared to 2019 levels due to the COVID-19 pandemic, breaking an upward trend that had been ongoing for more than ten years. However, shortly after the pandemic's peak, global GHG emissions began to increase once more, reaching a level of 53.8 Gt CO_2 eq in 2022, which is 2.3% higher than in 2019 and 1.4% above 2021 [4].

Due to these problems, different energy types and technologies that can reduce emissions have become the focus of studies in the scientific community. For a long time, these investigations have centered on renewable energy technology. Renewable energy sources—biomass, geothermal, solar, tidal, wave, and wind power, among others—are expected to address the world's energy problems without negatively impacting the economy, environment, or the resources of future generations. Clean energy solutions aim to achieve these vital targets for increased sustainability, including higher efficiency, effective resource use, lower costs, a better environment, energy security, and superior design and analysis. Wind energy has rapidly gained acceptance among society, industry, and politics for being a clean, viable, cost-effective, and eco-friendly option [5].

Wind energy power forecasting can be influenced by various factors. One of the most significant factors is wind speed, as higher wind speeds generally result in higher power generation. Additionally, wind direction can impact the efficiency of wind turbines, as turbines are designed to capture wind from specific directions. The design and specifications of the wind turbine itself, such as rotor diameter, blade length, and generator capacity, also influence power estimation. The surrounding terrain and the presence of obstacles like buildings, trees, or hills can cause turbulence and affect wind flow, leading to variations in power estimation. Furthermore, changes in temperature and air density can impact the performance of wind turbines, with cold temperatures and higher air density increasing power output, while warmer temperatures and lower air density decrease it. Finally, over time, wind turbines may experience performance degradation due to wear and tear, mechanical issues, or maintenance requirements, which can affect the accuracy of power estimation.

Machine learning, which has strong capabilities in capturing nonlinear relationships between input and output, is used for short-term wind energy forecasting. Many methods and models have been developed to estimate wind power in past studies. Several studies indicate that machine learning algorithms hold potential for efficiently forecasting wind power generation [6]. Random forest is one of the prominent machine learning methods for short-term wind energy forecasting [7]. The random forest algorithm, which has a strong adaptive ability, produces a more suitable model because it can train on data without normalization, which meets the demands of wind energy forecasting [8].

The gradient boosting machine regression approach outperformed five other optimal machine learning techniques in Singh et al.'s comparison of short-term wind energy generation forecasts [9]. In 2019, Demolli et al. conducted a study showcasing the application of machine learning algorithms in estimating long-term wind power values [10]. By utilizing daily wind speed data, the researchers demonstrated the effectiveness of these algorithms. In a separate investigation conducted by Liu and Fan in 2021, three machine learning algorithms—Decision Trees, and Random Forests—were compared [11]. The findings revealed that the KNN approach surpassed decision trees in accurately predicting wind power. Extreme Gradient Boosting (XGBoost) and ensemble approaches have become popular for forecasting power generation from renewable energy sources, particularly in wind systems and short-term models [12]. The potential of machine learning algorithms, such as gradient boosting regression trees, decision trees, random forests, KNN, XGBoost and multiple linear regression has been highlighted in numerous papers [13,14]. Forecasting wind power methods can also employ various techniques beyond machine learning. Numerous researchers have utilized and developed neural networks, deep learning methodologies, and hybrid approaches that integrate these with machine learning techniques [15–18]. There are also studies on new algorithms and models to reduce the randomness and uncertainty of wind energy and increase forecast accuracy [19–22]. In addition to these studies, many authors continue to publish research that explores the advantages, disadvantages, and future prospects of these methodologies [23–30].

Accurate prediction of wind power is essential for ensuring the reliability of power systems and reducing costs associated with their operation. Additionally, precise predictions benefit governmental bodies, policymakers, and other responsible entities in making informed decisions and taking appropriate measures. We aim to forecast wind energy, considering meteorological factors, wind speed, turbine technology, and physical measurements taken from SCADA data for all turbine components.

The study contributes to the literature in many ways.

1. Real-time data usage: The study provides more up-to-date and accurate estimates based on current meteorological conditions by using real-time SCADA data to estimate wind turbine power generation.

2. Combined dataset: By considering 77 different parameters for wind power generation, we evaluate not only wind speed but also other meteorological factors such as temperature, humidity, pressure, and physical measurements of turbine components.

3. Detailed performance analysis: We comprehensively evaluate the performance of the algorithms according to various criteria such as R^2, MAE, MSE, RMSE, and MAPE, as well as computational cost. This increases the reliability of the results and shows which model is more optimal.

2. Materials and methods

In the study, data analysis is performed using Python, one of the programming languages widely used for statistical calculation and data analysis. In the literature review, it is observed that decision tree, random forest, XGBoost, and KNN algorithms are frequently used [31]. For this reason, the model is obtained using these algorithms during the application phase of the study. Our aim here is to make predictions regarding electricity production from the wind turbine and to determine which algorithm makes predictions with higher performance, taking current weather conditions as input. Within the scope of the analysis, different Python libraries are used for various purposes. Accordingly, pandas is used for data analysis; matplotlib is utilized for visualizing various data, such as the representation of pairwise correlations; and sklearn is employed for machine learning algorithms and model performance evaluations. These are some of the most important libraries used. The Cross-Industry Standard Process for Data Mining, one of the widely used analytical models, consists of six stages: Defining the problem, understanding the data, preparing the data, modeling, model evaluation and selection, and implementing the model [32]. The application part of the study is carried out by following these steps. However, in this study, the "Implementation of the Model" step is not included.

2.1. Defining the problem

Wind power holds a significant position among numerous sustainable and eco-friendly energy sources. Accurate and dependable wind power prediction plays a crucial role in seamlessly integrating wind energy into the power grid [33]. Forecasts pertaining to wind energy production are indispensable for effectively strategizing and optimizing the placement of wind farms. By analyzing and making accurate predictions based on historical wind data, developers and policymakers can identify suitable locations for wind projects, estimate potential energy yield, and optimize the design and capacity of wind turbines. Reliable forecasts also help assess the economic viability of wind energy investments and secure financing for new projects. In recent years, the literature extensively explores the prediction of wind energy production using machine learning algorithms, making it a widely popular application in the renewable energy sector. As a result of the search in the Scopus database according to the TITLE-ABS-KEY ("wind power" and "prediction") criteria, 10,746 studies are obtained. Figure 1 shows the number of publications by year.

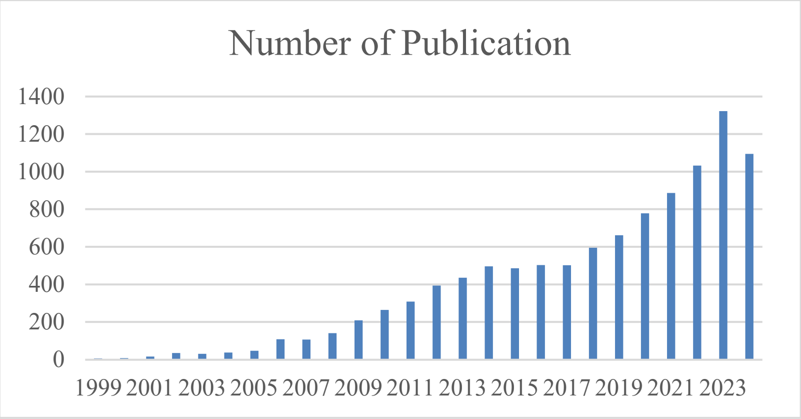


Figure 1. Wind energy estimated number of publications by years.

According to the figure, the first studies date back to 1999 and before. According to the general view, the number of publications has continued to increase since the beginning. After 2017, the number of publications in the field has increased linearly every year, and the subject remains current today.

Significant enhancements in meteorological models, data analytics, and machine learning methodologies have greatly enhanced the precision of wind energy predictions. These progressions encompass the assimilation of up-to-the-minute weather data, high-resolution modeling, ensemble forecasting, and the utilization of artificial intelligence algorithms to scrutinize intricate wind patterns. In the past three years, there has been a significant rise in the examination of wind energy prediction, with a notable progression in these investigations and an anticipation for further advancement in the future. Figure 2 presents a correlation chart, illustrating the interrelation of numerous studies, which our own study compares and incorporates. Additionally, the chart includes the names of researchers, with the size of their representation being determined by the number of citations they have received.

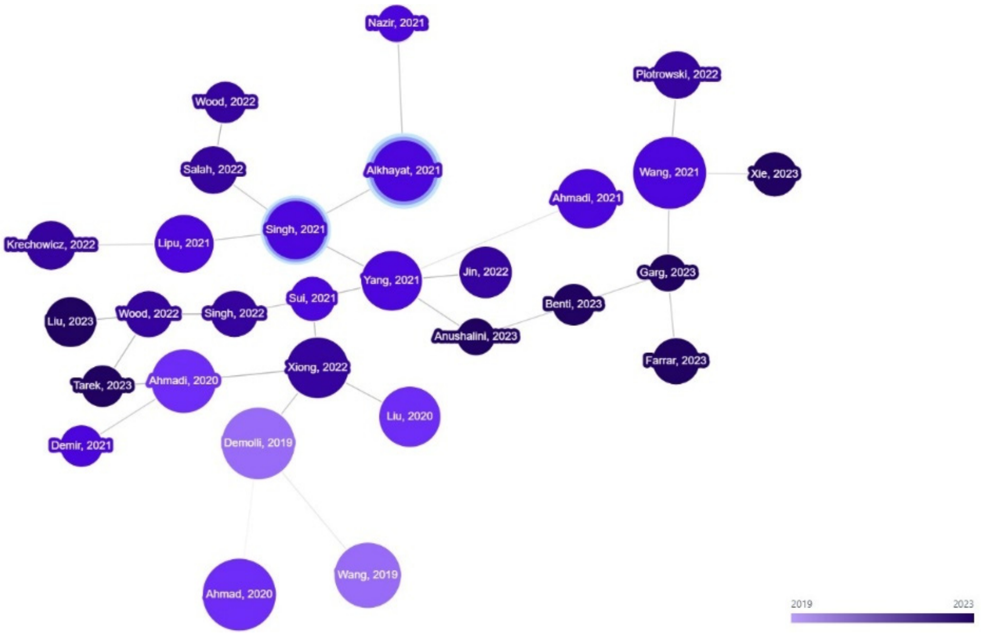


Figure 2. Correlation chart of related studies.

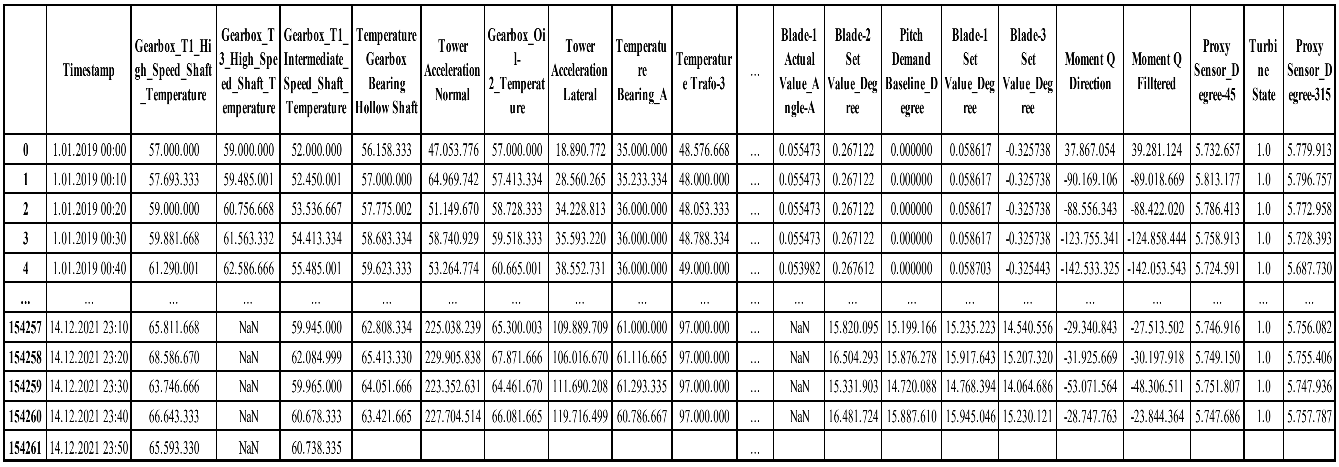
The primary issue addressed in this study is the challenge of accurately forecasting wind power generation due to its inherent variability and dependence on diverse meteorological factors. Wind energy production is subject to fluctuations caused by changes in wind speed, direction, temperature, humidity, and other environmental conditions. Traditional forecasting methods often struggle to integrate these complex variables, leading to unreliable predictions that can hinder the efficient management of wind resources.

We aim to overcome these challenges by employing advanced machine learning algorithms that can analyze large datasets of historical weather and power generation data. By utilizing real-time SCADA data from wind turbines, the research seeks to develop a robust predictive model that not only improves the accuracy of wind power forecasts but also enhances the integration of wind energy into the electricity grid. Thus, the central problem is the need for a more precise and efficient forecasting method that can adapt to the dynamic nature of wind energy production.

2.2. Data understanding

Two distinct datasets, the wind energy production dataset and weather data, including wind speed, wind direction, temperature, humidity, atmospheric pressure, and other pertinent variables, are utilized on Kaggle to assess the algorithms' performance [34]. Dataset contains power (kW) that turbine produces in real time on 10-minute basis between January 1, 2019 and August 14, 2021.

Understanding the data is the stage where the characteristics of the dataset are analized, such as the number of observations, number of attributes, and missing data. Figure 3 shows the raw version of the "Features" dataset, which has not been analized.

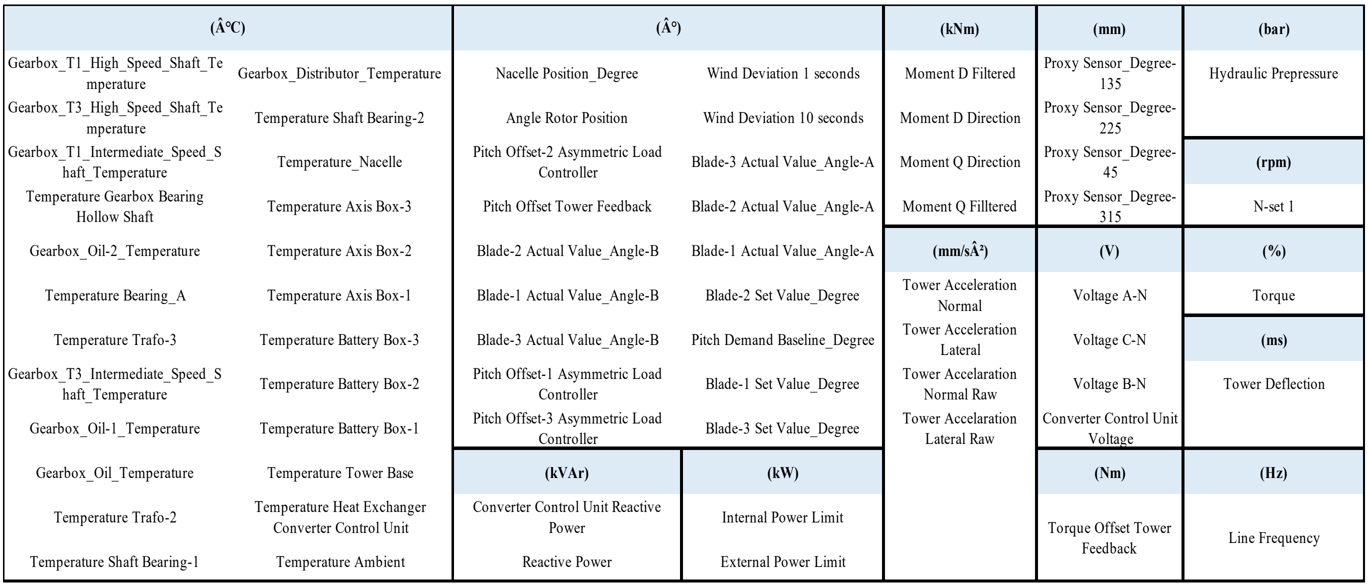


	Timestamp	Gearbox_T1_High_Speed_Shaft_Temperature	Gearbox_T3_High_Speed_Shaft_Temperature	Gearbox_T1_Intermediate_Speed_Shaft_Temperature	Temperature Gearbox Bearing Hollow Shaft	Tower Acceleration Normal	Gearbox_Oil-2_Temperature	Tower Acceleration Lateral	Temperature Bearing_A	Temperature Trafo-3	...	Blade-1 Actual Value_Angle-A	Blade-2 Set Value_Degree	Pitch Demand Baseline_Degree	Blade-1 Set Value_Degree	Blade-3 Set Value_Degree	Moment Q Direction	Moment Q Filltered	Proxy Sensor_Degree-45	Turbine State	Proxy Sensor_Degree-315
0	1.01.2019 00:00	57.000.000	59.000.000	52.000.000	56.158.333	47.053.776	57.000.000	18.890.772	35.000.000	48.576.668	...	0.055473	0.267122	0.000000	0.058617	-0.325738	37.867.054	39.281.124	5.732.657	1.0	5.779.913
1	1.01.2019 00:10	57.693.333	59.485.001	52.450.001	57.000.000	64.969.742	57.413.334	28.560.265	35.233.334	48.000.000	...	0.055473	0.267122	0.000000	0.058617	-0.325738	-90.169.106	-89.018.669	5.813.177	1.0	5.796.757
2	1.01.2019 00:20	59.000.000	60.756.668	53.536.667	57.775.002	51.149.670	58.728.333	34.228.813	36.000.000	48.053.333	...	0.055473	0.267122	0.000000	0.058617	-0.325738	-88.556.343	-88.422.020	5.786.413	1.0	5.772.958
3	1.01.2019 00:30	59.881.668	61.563.332	54.413.334	58.683.334	58.740.929	59.518.333	35.593.220	36.000.000	48.788.334	...	0.055473	0.267122	0.000000	0.058617	-0.325738	-123.755.341	-124.858.444	5.758.913	1.0	5.728.393
4	1.01.2019 00:40	61.290.001	62.586.666	55.485.001	59.623.333	53.264.774	60.665.001	38.552.731	36.000.000	49.000.000	...	0.053982	0.267612	0.000000	0.058703	-0.325443	-142.533.325	-142.053.543	5.724.591	1.0	5.687.730
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
154257	14.12.2021 23:10	65.811.668	NaN	59.945.000	62.808.334	225.038.239	65.300.003	109.889.709	61.000.000	97.000.000	...	NaN	15.820.095	15.199.166	15.235.223	14.540.556	-29.340.843	-27.513.502	5.746.916	1.0	5.756.082
154258	14.12.2021 23:20	68.586.670	NaN	62.084.999	65.413.330	229.905.838	67.871.666	106.016.670	61.116.665	97.000.000	...	NaN	16.504.293	15.876.278	15.917.643	15.207.320	-31.925.669	-30.197.918	5.749.150	1.0	5.755.406
154259	14.12.2021 23:30	63.746.666	NaN	59.965.000	64.051.666	223.352.631	64.461.670	111.690.208	61.293.335	97.000.000	...	NaN	15.331.903	14.720.088	14.768.394	14.064.686	-53.071.564	-48.306.511	5.751.807	1.0	5.747.936
154260	14.12.2021 23:40	66.643.333	NaN	60.678.333	63.421.665	227.704.514	66.081.665	119.716.499	60.786.667	97.000.000	...	NaN	16.481.724	15.887.610	15.945.046	15.230.121	-28.747.763	-23.844.364	5.747.686	1.0	5.757.787
154261	14.12.2021 23:50	65.593.330	NaN	60.738.335							...										

Figure 3. Features dataset.

Through the queries made, dataset reveals 154,262 observations and 77 attributes containing various variables. The timestamp is of object data type, while all other attributes are of float data type.

The unit equivalents of the attributes in the dataset are provided in Figure 4.



(Â°C)		**(Â°)**		**(kNm)**	**(mm)**	**(bar)**
Gearbox_T1_High_Speed_Shaft_Temperature	Gearbox_Distributor_Temperature	Nacelle Position_Degree	Wind Deviation 1 seconds	Moment D Filtered	Proxy Sensor_Degree-135	Hydraulic Prepressure
Gearbox_T3_High_Speed_Shaft_Temperature	Temperature Shaft Bearing-2	Angle Rotor Position	Wind Deviation 10 seconds	Moment D Direction	Proxy Sensor_Degree-225	
Gearbox_T1_Intermediate_Speed_Shaft_Temperature	Temperature_Nacelle	Pitch Offset-2 Asymmetric Load Controller	Blade-3 Actual Value_Angle-A	Moment Q Direction	Proxy Sensor_Degree-45	**(rpm)**
Temperature Gearbox Bearing Hollow Shaft	Temperature Axis Box-3	Pitch Offset Tower Feedback	Blade-2 Actual Value_Angle-A	Moment Q Filltered	Proxy Sensor_Degree-315	N-set 1
Gearbox_Oil-2_Temperature	Temperature Axis Box-2	Blade-2 Actual Value_Angle-B	Blade-1 Actual Value_Angle-A	**(mm/sÂ²)**	**(V)**	**(%)**
Temperature Bearing_A	Temperature Axis Box-1	Blade-1 Actual Value_Angle-B	Blade-2 Set Value_Degree	Tower Acceleration Normal	Voltage A-N	Torque
Temperature Trafo-3	Temperature Battery Box-3	Blade-3 Actual Value_Angle-B	Pitch Demand Baseline_Degree	Tower Acceleration Lateral	Voltage C-N	**(ms)**
Gearbox_T3_Intermediate_Speed_Shaft_Temperature	Temperature Battery Box-2	Pitch Offset-1 Asymmetric Load Controller	Blade-1 Set Value_Degree	Tower Accelaration Normal Raw	Voltage B-N	Tower Deflection
Gearbox_Oil-1_Temperature	Temperature Battery Box-1	Pitch Offset-3 Asymmetric Load Controller	Blade-3 Set Value_Degree	Tower Accelaration Lateral Raw	Converter Control Unit Voltage	
Gearbox_Oil_Temperature	Temperature Tower Base	**(kVAr)**	**(kW)**		**(Nm)**	**(Hz)**
Temperature Trafo-2	Temperature Heat Exchanger Converter Control Unit	Converter Control Unit Reactive Power	Internal Power Limit		Torque Offset Tower Feedback	Line Frequency
Temperature Shaft Bearing-1	Temperature Ambient	Reactive Power	External Power Limit			

Figure 4. Feature units.

The “Power” data set is examined and shown in Table 1.

Table 1. Power dataset.

	Timestamp	Power (kW)
0	1.01.2019 00:00	705.876.648
1	1.01.2019 00:10	884.711.670
2	1.01.2019 00:20	982.875.000
3	1.01.2019 00:30	1.115.943.359
4	1.01.2019 00:40	1.263.841.675
...	...	...
136725	14.08.2021 23:10	2.757.728.271
136726	14.08.2021 23:20	2.758.323.242
136727	14.08.2021 23:30	2.759.243.408
136728	14.08.2021 23:40	2.761.261.719
136729	14.08.2021 23:50	2.758.593.262

The analysis determines that there are two attributes in the power dataset: The timestamp and the amount of power produced in kW, with a total of 136,730 observations. “Power (kW)” in the power dataset is selected as the target attribute. At this stage, the box plot, which allows for a better understanding of the data, is shown in Figure 5 for the “Power (kW)” target attribute.

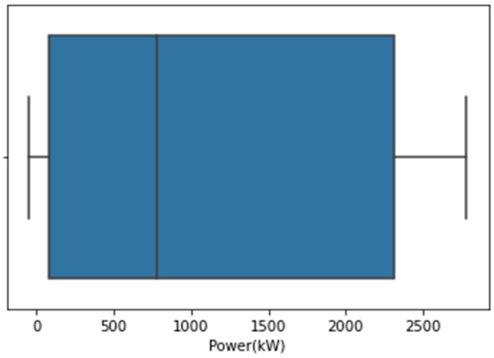
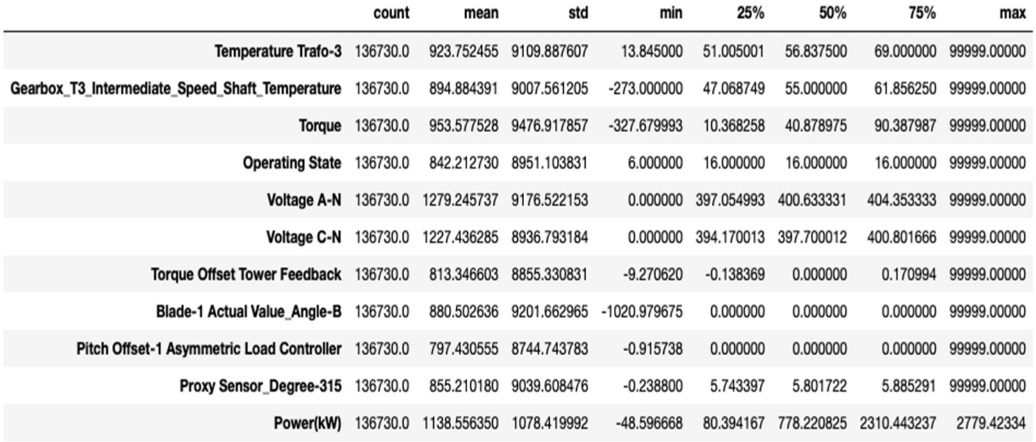


Figure 5. Power(kW) box plot.

The box plot, which basically shows how the values are distributed, also provides information about whether there are extreme values. Accordingly, it is possible to say that the Power(kW) target attribute does not contain extreme values.

2.3. Data preparation

A single data set is created by combining the data sets using the timestamp attributes from both sources. The new data set is checked for missing and repetitive data, and observations with missing data are removed. After this process, the number of observations in the data set is 136,730. Additionally, feature selection is conducted for the 77 attributes affecting the target attribute. Thanks to this method, which allows the selection of the features that contribute the most to the target quality, the top 10 predictors for the target quality were determined in the study. Accordingly, the attributes that are most effective in predicting the target attribute are ‘Temperature Transformer-3’, ‘Gearbox_T1_Intermediate_Speed_Shaft_Temperature’, ‘Torque’, ‘Operating_State’, ‘Voltage A-N’, ‘Voltage C-N’, ‘Torque Offset Tower Feedback’, It has been determined that there are ‘Blade-1 Actual Value_Angle-B’, ‘Pitch Offset-1 Asymmetric Load Controller’, and ‘Proxy Sensor_Degree-315’. The final feature selection version of the data set is summarized in Figure 6.



	count	mean	std	min	25%	50%	75%	max
Temperature Trafo-3	136730.0	923.752455	9109.887607	13.845000	51.005001	56.837500	69.000000	99999.00000
Gearbox_T3_Intermediate_Speed_Shaft_Temperature	136730.0	894.884391	9007.561205	-273.000000	47.068749	55.000000	61.856250	99999.00000
Torque	136730.0	953.577528	9476.917857	-327.679993	10.368258	40.878975	90.387987	99999.00000
Operating State	136730.0	842.212730	8951.103831	6.000000	16.000000	16.000000	16.000000	99999.00000
Voltage A-N	136730.0	1279.245737	9176.522153	0.000000	397.054993	400.633331	404.353333	99999.00000
Voltage C-N	136730.0	1227.436285	8936.793184	0.000000	394.170013	397.700012	400.801666	99999.00000
Torque Offset Tower Feedback	136730.0	813.346603	8855.330831	-9.270620	-0.138369	0.000000	0.170994	99999.00000
Blade-1 Actual Value_Angle-B	136730.0	880.502636	9201.662965	-1020.979675	0.000000	0.000000	0.000000	99999.00000
Pitch Offset-1 Asymmetric Load Controller	136730.0	797.430555	8744.743783	-0.915738	0.000000	0.000000	0.000000	99999.00000
Proxy Sensor_Degree-315	136730.0	855.210180	9039.608476	-0.238800	5.743397	5.801722	5.885291	99999.00000
Power(kW)	136730.0	1138.556350	1078.419992	-48.596668	80.394167	778.220825	2310.443237	2779.42334

Figure 6. Dataset summary.

The mean, standard deviation, the minimum, the first quartile, the middle, and the third quartile values of the attributes of a total of 136,730 observations are shown.

2.4. Modeling

Within the scope of the study, the performance of Decision Tree, Random Forest, KNN, and XGBoost algorithms on the dataset is evaluated. Feature selection is conducted for 77 attributes that affect the target attribute. We consider both the 10 predictors that contribute the most to the target attribute and all attributes for model performance evaluation. The “hold-out” method is used for model performance evaluation, where the performance of the model is assessed by creating an 80% training dataset and a 20% test dataset.

2.5. Modeling and evaluation

At this stage, the performance values of the models that provide solutions to the problem defined at the beginning are evaluated. An application was carried out to make predictions about electricity production from wind turbines with Python programming language using Decision Tree, Random Forest, KNN, and XGBoost algorithms and to determine which algorithm makes predictions with higher performance by taking current weather conditions as input.

Model performance evaluation is based on R², MAE, MSE, RMSE, and MAPE criteria. The performance criteria used are given below.

$$MAE = \frac{1}{n} + \sum_{j=1}^{n} \left| y_j - \hat{y}_j \right| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \left(y_j - \hat{y}_j \right)^2} = \sqrt{MSE} \tag{2}$$

$$MAPE = \frac{1}{n}\sum_{j=1}^{n} \left| y_j - \hat{y}_j \right| \tag{3}$$

$$R^2 = \frac{\sum_{j=1}^{n} \left(y_j - \hat{y}_j \right)^2}{\sum_{j=1}^{n} \left(y_j - \bar{y}_j \right)^2} \tag{4}$$

3. Results

Wind power is a notable type of sustainable energy that attracts global attention owing to its economical nature as an energy generation source. Wind energy is inherently variable, non-linear and weather dependent. This creates a problem in terms of accurate prediction. This article estimates Power (kW) based on the analysis of variables including wind speed, wind direction, temperature, humidity, atmospheric pressure, and other relevant attributes. CRISP-DM steps were followed in the article. However, since the obtained models were not integrated into any system, the last step of CRISP-DM, the implementation of the model, was not carried out. In order to estimate the power to be obtained within the scope of the study, Decision Tree, Random Forest, KNN, and XGBoost. The obtained results were compared using criteria for evaluating the model's performance. In addition to R^2, the MAE, MSE, RMSE, and MAPE criteria were also taken into account when monitoring the models' performance. Model performance evaluations of the algorithms are summarized in Table 2.

Table 2. Performance evaluation criteria.

	ALGORITHMS							
Performance evaluation criteria	Decision tree		Random forest		KNN		XGBoost	
	Feature selection	Normal	Feature selection	Normal	Feature selection	Normal	Feature selection	Normal
R^2	0.99	0.99	0.99	0.99	0.99	0.78	0.99	0.99
MAE	12.48	7.65	9.66	5.39	23.85	285.98	12.01	7.84
MSE	3789.48	1788.94	2209.57	815.99	10572.79	244683.08	1954.09	678.22
RMSE	61.55	42.29	47.00	28.56	102.82	494.65	44.20	26.04
MAPE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

R^2 refers to the rate of change in the dependent variable that can be predicted from the independent variables. In the model obtained by feature selection, all of the algorithms used within the scope of analysis are 99%. This shows that the resulting model can predict the target variable (Power(kW)) in the training data set with 99%. However, when examined without feature selection, it was observed that only the performance of the KNN model decreased. The fact that KNN is a lazy algorithm is thought to affect this result. R^2 alone is not sufficient to explain the model's performance. Other criteria also need to be evaluated for a better interpretation of the model's performance.

MAE shows how much the model's predictions deviate from the actual values, on average. Accordingly, in the feature selection model, the Random Forest model gave the least deviation from the real values with 9.66. This model is followed by XGBoost with 12.01, Decision Tree with 12.48 and KNN with 23.85, respectively. When looking at the MAE value of the model obtained without Feature Selection, it was seen that the Random Forest model gave the least deviation. The Decision Tree model came in second and XGBoost came in third. The KNN model has the highest deviation at 285.98.

MSE shows how close the predictions are to the actual values. A lower MSE value means the prediction is more accurate. MSE is obtained by dividing the sum of the squares of the differences between the actual Power (kW) and the predicted Power (kW) by the number of observations. Accordingly, among the models created by feature selection, it was observed that the XGBoost model made a more accurate prediction with 1954.09. The XGBoost algorithm is followed by Random Forest with 2209.57, Decision Tree with 3789.48 and KNN with 10572.79, respectively. When we look at the models obtained without feature selection, it is seen that the XGBoost model comes first with 678.22. It is seen that the KNN model makes the prediction that is furthest from the real values. According to the results obtained, the XGBoost model gave the best result among all models.

RMSE is the square root of MSE. It is used to find the difference between the values predicted from the model and the real values. Except for KNN, all algorithms gave better results without feature selection. The XGBoost model came first with 26.04, followed by the Random Forest and Decision Tree models, respectively. According to the obtained RMSE value, the model has an average error of 26 kW when it predicts Power(kW). Although the model has a bias problem, it is at a low level.

MAPE is a statistical measure used to assess the accuracy of a forecast or prediction model. MAPE expresses the error as a percentage of the actual values, making it easier to interpret the forecast accuracy regardless of the scale of the data. the low MAPE across all models indicates that they perform well in terms of prediction accuracy, making them suitable choices for forecasting tasks in this context.

Looking at the general situation of model performance metrics, the XGBoost model has high R^2 and low MAE, MSE, RMSE and MAPE values. This shows that the model fits the data well and the predictions are generally accurate.

The algorithms are also examined in terms of computational cost. Computational cost means that the resources (time, memory, processing power, etc.) consumed by each model during its implementation and execution are calculated. It emphasizes that a balanced approach should be adopted in terms of both performance and resource usage during model selection and optimization. Therefore, evaluating the computational costs of each model is critical to developing more effective and efficient systems. Within the scope of the study, data analysis was performed in Jupyter Notebook via Anaconda Navigator and Python version 3.11.5 was used. The computational costs of the model were performed on an M1 Mac computer with 8 GB RAM. Table 3 shows the calculation costs.

Table 3. Calculation costs.

Calculation costs				
Algorithms	Time (SEC)	Memory (MB)	CPU usage before (%)	CPU usage after (%)
Decision tree	74.02	406.14	3.4	2.0
Random forest	74.00	310.01	0.7	0.2
KNN	74.72	403.39	2.2	7.2
XGBoost	73.62	372.70	2.6	2.7

XGBoost stands out as the most efficient algorithm in terms of both training time and memory usage. This makes it a suitable option for large datasets. Random Forest and KNN stand out with their high memory and CPU consumption. Therefore, these algorithms should be carefully evaluated before using them on larger and more complex datasets. Decision Trees and XGBoost can provide advantages in data processing with lower resource usage, which makes them more preferable in applications.

According to the results, the XGBoost model stands out in terms of balancing both performance and resource consumption in model selection by offering high performance and low cost.

A correlation matrix was created to determine whether the relationships between the dependent variables (Temperature Trafo-3, Gearbox_T1_Intermediate_Speed_Shaft_Temperature, Torque, etc.) and the independent variable (Power (kW)) obtained as a result of the feature selection made in line with the 10 predictors that have the most impact on the target variable are significant. The correlation matrix is given in Figure 7.

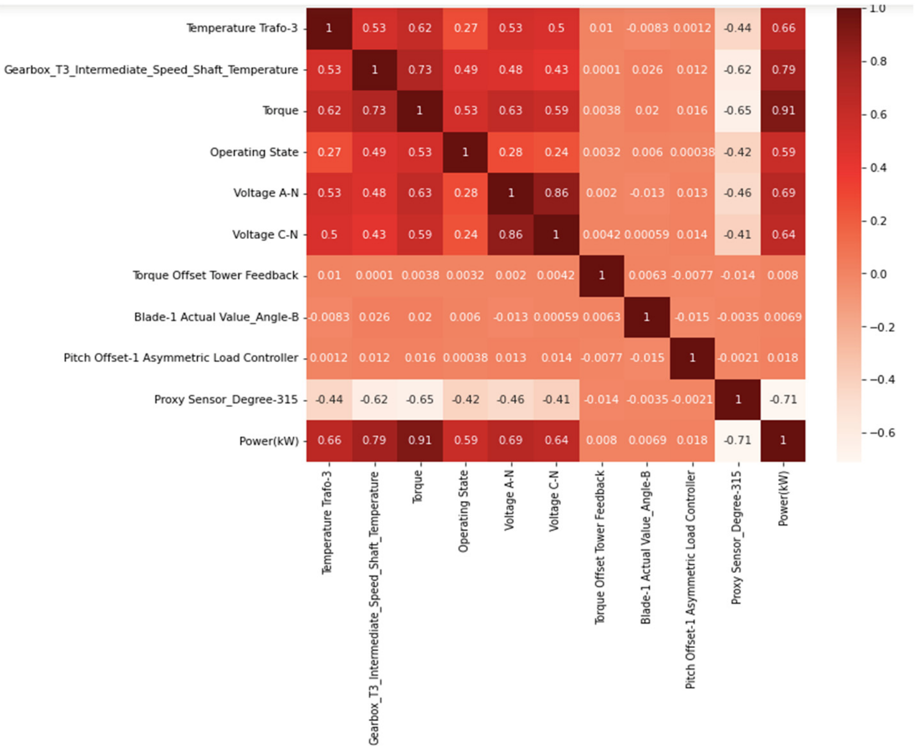


Figure 7. Correlation matrix.

When the correlation matrix was examined, it was seen that the dependent variable that had the most impact on the target attribute was Torque with 91%. That is, there is a strong and significant relationship between Torque and Power(kW), and the most important predictor explaining the target attribute is the Torque attribute. On the other hand, it is seen that there is a moderate negative correlation of 71% between Proxy Sensor_Degree-315 and Power(kW) and there is a statistically significant relationship between these variables.

Upon reviewing the literature, we find that Support Vector Regression, Random Forest, XGBoost, Decision Tree, Gradient Boosting, and KNN algorithms are frequently employed in machine learning studies. From January to December 2018, experimental data was gathered at a 10-minute sampling rate using a SCADA system as part of a study to estimate wind power based on wind speed and wind direction data. Using random forest, KNN, gradient boosting, decision trees, extra trees, and regression algorithms, the model was built for the study. It was found that the gradient boosting algorithm produced the best results [9]. In a different study, the Random Forest, Least Absolute Shrinkage Selector Operator, Support Vector Regression, KNN and XGBoost, algorithms were used to estimate wind power based on wind speed data from four distinct regions. Modeling was done using. The best model performance was shown by XGBoost and Support Vector Regression [10]. Additionally, our research findings are similar to another study comparing the performance of traditional, deep learning and ensemble learning models for short-term wind speed prediction, and the XGBoost model appears to outperform other models [35].

The energy produced by a turbine is influenced by various factors, including wind speed and location. However, these factors can have varying physical impacts on different components of the turbine. While researchers have focused solely on wind speed as a parameter for wind energy production, we consider 77 separate parameters, including meteorological effects and SCADA data. We identified the most significant parameters affecting wind energy production and incorporated the effect of wind speed on multiple turbine components as separate data in our algorithm, allowing for more accurate predictions.

4. Discussion

Forecasting wind energy production is of great importance in the renewable energy industry. Wind power forecasting enables energy providers to anticipate and strategize the potential electricity generation from wind turbines. Precise prediction aids in optimizing the seamless integration of wind power into the grid, thereby ensuring a consistent and dependable energy supply. It also enables better management of energy resources, reduces costs, and improves overall efficiency. By predicting wind energy production, operators can make informed decisions regarding energy distribution, storage, and backup power generation. Furthermore, the utilization of forecasting techniques plays a crucial role in effectively addressing energy demand, diminishing the dependence on non-renewable energy sources, and alleviating the adverse environmental consequences associated with energy generation. Machine learning techniques have exhibited significant potential in predicting wind power generation, which is vital for the effective management of renewable energy resources. A variety of methodologies have been investigated, encompassing both individual algorithms and ensemble techniques.

Our findings of this study contribute significantly to the existing body of literature on wind power forecasting by employing machine learning techniques to enhance prediction accuracy. Our results indicate that the XGBoost model outperformed other algorithms, achieving a high R^2 value and low MAE, MSE, RMSE, and MAPE metrics. This reinforces the findings of previous studies, such as those by Singh et al. and Demolli et al., which also identified XGBoost as a leading algorithm for wind power prediction [9,10].

Unlike prior research that primarily focused on wind speed as a solitary predictor, our study integrated 77 distinct parameters, including meteorological conditions and SCADA data from turbine components. By incorporating these diverse factors, we were able to achieve more accurate predictions, demonstrating that the performance of wind energy forecasting can be significantly improved through a multi-faceted data approach.

Furthermore, the utilization of real-time SCADA data contributes to the novelty of our research. This real-time aspect is crucial, as highlighted by Anushalini and Revathi, who noted that timely data integration can lead to better decision-making in energy management [15]. Our findings support this assertion, as the use of real-time data allowed for more dynamic and responsive forecasting, ultimately aiding in the integration of wind energy into the grid.

The comparison of model performances in our study also reveals crucial insights. While studies have shown that Random Forest and KNN algorithms can deliver reliable results, our analysis indicates that these methods may not perform as well under certain conditions, particularly when feature selection is not applied. This underscores the importance of feature selection, as shown by our findings where the Random Forest and Decision Tree models demonstrated improved accuracy with selected features.

Overall, we not only confirm the efficacy of machine learning techniques in forecasting wind power generation but also enhance the literature by providing a detailed analysis of various algorithms and the significance of incorporating a broader range of influencing factors. The implications of our findings are substantial for energy providers and policymakers, highlighting the necessity for advanced forecasting methods that can adapt to the complexities of renewable energy generation.

5. Conclusions

We employ machine learning methods to predict wind power generation by utilizing historical weather data alongside corresponding wind power generation data. The dataset, sourced from real-time SCADA data of wind turbines, associates each weather data point with a specific time period, paired with the respective power generation output. A model was developed using Decision Tree, Random Forest, KNN, and XGBoost algorithms.

The performance of these models was rigorously evaluated using metrics such as R^2, MAE, MSE, RMSE, and MAPE. Among the models assessed, the XGBoost algorithm demonstrated the highest R^2 value alongside the lowest MAE, MSE, RMSE, and MAPE values, indicating that it fits the data exceptionally well and provides accurate power generation predictions.

In addition to performance metrics, the computational costs of each model were analyzed. The XGBoost algorithm stood out for its efficiency, exhibiting a balance between lower training time and memory usage compared to other models. This efficiency makes it particularly suitable for real-time applications in wind power forecasting.

Overall, our findings affirm the effectiveness of machine learning techniques, particularly XGBoost, in accurately predicting wind power generation while highlighting the importance of computational efficiency in practical implementations. These results contribute valuable insights to the literature on renewable energy forecasting and suggest pathways for future research in optimizing predictive models across different energy types.

In this study, similar results were obtained with the literature. For future studies, it is recommended to explore the integration of advanced algorithms, such as deep learning models, to further enhance prediction accuracy. Additionally, the continued use of real-time SCADA data is essential, and future research should focus on developing frameworks for dynamic model updates as new data becomes available. Investigating multi-factor interactions among the 77 parameters identified in this study could yield further insights into forecasting accuracy. Regional case studies could also be valuable, as they would assess the impact of local geographic and climatic factors on wind power generation. Furthermore, incorporating economic feasibility analyses could provide insights into the cost-effectiveness of implementing these forecasting models. Last, research findings should be used to inform policymakers about the importance of predictive technologies in renewable energy, and comparative studies with other renewable sources could help develop a more comprehensive understanding of energy forecasting.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

There is no conflicts of interest between the authors.

Author contributions

Author Asiye Bilgili performed data analysis and wrote the materials and methodology section. Author Kerem Gül made the planning of the study. Wrote the introduction and results section.

References

1. Michaelowa A, Dransfeld B, Blodgett C, et al. (2012) IRENA handbook on renewable energy nationally appropriate mitigation actions (NAMAs) for policy makers and project developers. Abu Dhabi, United Arab Emirates: IRENA (International Renewable Energy Agency).
2. Breidenich C, Magraw D, Rowley A, et al. (1998) The Kyoto Protocol to the United Nations framework convention on climate change. *American J Int Law* 92: 315–331. https://doi.org/10.2307/2998044
3. Kumar S, Madlener R (2018) Energy systems and COP21 Paris climate agreement targets in Germany: An integrated modeling approach. *2018 7th International Energy and Sustainability Conference (IESC)*, 1–6. https://doi.org/10.1109/IESC.2018.8440004
4. Crippa M, Guizzardi D, Schaaf E, et al. (2023) GHG emissions of all world countries: 2023. Publications Office of the European Union. European Commission, Joint Research Centre. Available from: https://data.europa.eu/doi/10.2760/953322.
5. Breitschopf B, Herbst A (2023) Supply chain risks in the EU's energy technologies: Terms of reference. Publications Office of the European Union. European Commission, Directorate-General for Energy. Available from: https://data.europa.eu/doi/10.2833/818557.

6. Saini VK, Kumar R, Al-Sumaiti AS, et al. (2023) Learning based short term wind speed forecasting models for smart grid applications: An extensive review and case study. *Electric Power Syst Res* 222: 109502. https://doi.org/10.1016/j.epsr.2023.109502
7. Jørgensen KL, Shaker HR (2020) Wind power forecasting using machine learning: state of the art, trends and challenges. *Proceedings of the 2020 the 8th IEEE International Conference on Smart Energy Grid Engineering (SEGE)*, 44–50. https://doi.org/10.1109/SEGE49949.2020.9181870
8. Ho CY, Cheng KS, Ang CH (2023) Utilizing the random forest method for short-term wind speed forecasting in the coastal area of central Taiwan. *Energies* 16: 1374. https://doi.org/10.3390/en16031374
9. Singh U, Rizwan M, Alaraj M, et al. (2021) A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments. *Energies* 14: 5196. https://doi.org/10.3390/en14165196
10. Demolli H, Dokuz AS, Ecemis A, et al. (2019) Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Convers Manage* 198: 111823. https://doi.org/10.1016/j.enconman.2019.111823
11. Liu T, Fan L (2021) Wind power prediction based on three machine-learning algorithms: Decision tree, k-nearest neighbors and random forest. In: Xu, J., Márquez, F.P.G., Hassan, M.H.A, Duca, G., Hajiyev, A., Altiparmak, F. Author, *Proceedings of the Fifteenth International Conference on Management Science and Engineering Management,* New York: Springer, Cham. 78: 490–499. https://doi.org/10.1007/978-3-030-79203-9_38
12. Krechowicz A, Krechowicz M, Poczeta K (2022) Machine learning approaches to predict electricity production from renewable energy sources. *Energies* 15: 9146. https://doi.org/10.3390/en15239146
13. Demir F, Tasci B (2021) Predicting the power of a wind turbine with machine learning-based approaches from wind direction and speed data. *2021 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP)*, 37–40. https://doi.org/10.1109/ICT-PEP53949.2021.9600959
14. Sui A, Qian W (2021) Forecasting the wind power generation in China by seasonal grey forecasting model based on collaborative optimization. *RAIRO-Oper Res* 55: 3049–3072. https://doi.org/10.1051/ro/2021136
15. Anushalini T, Sri Revathi B (2023) Role of machine learning algorithms for wind power generation prediction in renewable energy management. *IETE J Res* 70: 4319–4332. https://doi.org/10.1080/03772063.2023.2205838
16. Jin H, Li Y, Wang B, et al. (2022) Adaptive forecasting of wind power based on selective ensemble of offline global and online local learning. *Energy Conver Manage* 271: 116296. https://doi.org/10.1016/j.enconman.2022.116296
17. Wood DA (2022) Trend decomposition aids short-term countrywide wind capacity factor forecasting with machine and deep learning methods. *Energy Conver Manage* 253: 115189. https://doi.org/10.1016/j.enconman.2021.115189
18. Saini VK, Kumar R, Mathur A, et al. (2020) Short term forecasting based on hourly wind speed data using deep learning algorithms. *3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE*), Jaipur, India: 1–6. https://doi.org/10.1109/ICETCE48199.2020.9091757

19. Ayene SM, Yibre AM (2024) Wind power prediction based on deep learning models: The case of Adama wind farm. *Heliyon* 10: e39579. https://doi.org/10.1016/j.heliyon.2024.e39579
20. Wang X, Hao Y, Yang W (2024) Novel wind power ensemble forecasting system based on mixed-frequency modeling and interpretable base model selection strategy. *Energy* 297: 131142. https://doi.org/10.1016/j.energy.2024.131142
21. Shinde SK, Tirlangi S, Kumar NK, et al. (2024) Enhancing wind power generation forecasting with advanced deep learning technique using wavelet-enhanced recurrent neural network and gated linear units. *International J Renewable Energy Res* 14: 324–338. https://doi.org/10.20508/ijrer.v14i2.14577.g8893
22. Wang Y, Zhao K, Hao Y, et al. (2024) Short-term wind power prediction using a novel model based on butterfly optimization algorithm-variational mode decomposition-long short-term memory. *Appl Energy* 366: 123313. https://doi.org/10.1016/j.apenergy.2024.123313
23. Piotrowski P, Rutyna I, Baczyński D, et al. (2022) Evaluation metrics for wind power forecasts: A comprehensive review and statistical analysis of errors. *Energies* 15: 9657. https://doi.org/10.3390/en15249657
24. Farrar NO, Ali MH, Dasgupta D (2023) Artificial Intelligence and machine learning in grid connected wind turbine control systems: A comprehensive review. *Energies* 16: 1530. https://doi.org/10.3390/en16031530
25. Xie Y, Li C, Li M, et al. (2023) An overview of deterministic and probabilistic forecasting methods of wind energy. *IScience* 26: 105804. https://doi.org/10.1016/j.isci.2022.105804
26. Wood DA (2023) Feature averaging of historical meteorological data with machine and deep learning assist wind farm power performance analysis and forecasts. *Energy Syst* 14: 1023–1049 https://doi.org/10.1007/s12667-022-00502-x
27. Benti NE, Chaka MD, Semie AG (2023) Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects. *Sustainability* 15: 7087. https://doi.org/10.3390/su15097087
28. Alkhayat G, Mehmood R (2021) A review and taxonomy of wind and solar energy forecasting methods based on deep learning. *Energy AI* 4: 100060. https://doi.org/10.1016/j.egyai.2021.100060
29. Garg S, Krishnamurthi R (2023) A survey of long short term memory and its associated models in sustainable wind energy predictive analytics. *Artif Intell Rev* 56: 1149–1198 https://doi.org/10.1007/s10462-023-10554-9
30. Nazir MS, Wang Y, Bilal M, et al. (2022) Wind energy, its application, challenges, and potential environmental impact. *Handbook of Climate Change Mitigation and Adaptation* New York: Springer, 1–38. https://doi.org/10.1007/978-1-4614-6431-0_108-2
31. Malakouti SM (2023) Improving the prediction of wind speed and power production of SCADA system with ensemble method and 10-fold cross-validation. *Case Stud Chem Environ Eng* 8: 100351. https://doi.org/10.1016/j.cscee.2023.100351
32. Wirth R, Hipp J (2000) Crisp-dm: Towards a standard process modell for data mining. *Computer Sci*
33. Mansoury I, El Bourakadi D, Yahyaouy A, et al. (2023) A novel wind power prediction model using graph attention networks and bi-directional deep learning long and short term memory. *Int J Electr Comput Eng*, 6847–6854. http://doi.org/10.11591/ijece.v13i6.pp6847-6854
34. Kaggle, Wind Turbine Power (kW) Generation data, 2023. Available from: https://www.kaggle.com/datasets/psycon/wind-turbine-energy-kw-generation-data.

35. Saini VK, Mathur F, Gupta V, et al. (2020) Predictive analysis of traditional, deep learning and ensemble learning approach for short-term wind speed forecasting. *International Conference on Computing, Power and Communication Technologies (GUCON)*, Greater Noida: IEEE, 783–788. https://doi.org/10.1109/GUCON48875.2020.9231081

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)