18th IMEKO TC10 Conference
"Measurement for Diagnostics, Optimisation and Control to Support Sustainability and Resilience"
Warsaw, Poland, September 26–27, 2022

Wind Generation Forecast With the Use of AI-Based Regression Methods

Adrian Bilski

Warsaw University of Life Sciences, Poland, adrian_bilski@sggw.edu.pl

***Abstract* – The topic of the paper is the presentation of the methodology and the results of forecasting energy generation by the wind turbine with the utilization of three regression methods: factorization machines, decision trees and random forests. The data coming from the wind farm in Turkey was first preprocessed to facilitate prediction task. The prediction of the eneregy production based on the**

Keywords – Forecasting, prediction, artificial intelligence, factorization machines, decision tress, random forests, wind energy production.

I. INTRODUCTION

The evolution of modern energy is heading towards the highest possible share of renewable energy, represented by wind energy obtained from turbines and solar energy obtained from panels [1], [2]. The advantage of wind energy over solar energy results from its greater environmental resources and more efficient energy generation technologies [3], [4].

According to the World Wind Energy Association, the overall capacity of all wind turbines has in 2021 exceeded 840 Gigawatt (Figure 1) [5]. It is an increase by 107,5 Gigawatt in regards to 2020 and provides more than 7% of the global power demand. This constitutes a growth rate of 13%, compared to 14% growth rate established in 2020 and 10% in 2019. This makes wind energy an indispensable element of the sustainable development of global energy and the process of reducing environmental pollution due to the combustion of fossil raw materials.

Restless operation of wind turbines (resulting with high variability of electricity generation over time) and the degree of their impact on the functioning of the power system are the main reasons the electroenergetic systems operators require windmill owners to forecast the power of the wind power plants. Such forecasts of electric energy generation in turbines, depending on the adopted time horizon, are applicable, inter alia, in balancing electricity supply and demand in the entire power system, load optimization and planning of power reserves [6].

Preparing accurate forecasts (burdened with the lowest possible error), although extremely important from the point of view of ensuring operational reliability of the whole power system and minimization of the financial cost of participating in trading in the electricity market, is a task extremely problematic. The uncertainty of predicting meteorological conditions, the size of power fluctuations (resulting from the technical conditions of this type of power plant and the weather conditions), the dependence of the wind resource characteristics (its force, direction), the geographic location of the turbine (altitude, physical properties of the sybstrate, etc.) are the ket challenges in the context of energy generation forecasting in regards to wind turbines [7].

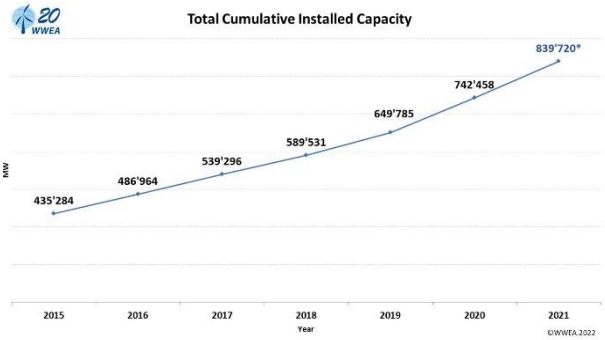

Fig. 1. Global cumulative installed wind capacity established throughout 2015-2021 [5].

A prerequisite for the development of accurate forecast models, besides the features of wind resource, is the comprehensive and complete information on the wind turbines being a part of the wind farm. The most important ones include turbibne's active output power, the angle of the rotor blades, its sweep area, the efficiency of the generator and the height of the tower.

In case of wind farms, only part of this information is recorded and saved in the installed SCADA (Supervisory Control and Data Acquisition) systems. The currently developed forecasting models mostly ignore the impact of the above factors on the energy production volume. This type of data is available only to turbine manufacturers / service technitians and is not made available to the user.

The production for the entire farm, consisting of many wind turbines, is more complex. In this case, new factors such as number of turbines currently in operation and the

topography must be taken into account. Changing the wind speed close to the limit value of turning on the turbines may cause some of them to turn offf. Restarting is time-consuming, which affects the actual production valume and the forecast error.

The utilization of the data in regards to the characteristics of the turbine in the pre-processing stage of prediction process, guarantees that during the model construction only those data will be utilized, which are not randomly drained by power fluctuaction of the turbine [8].

In regards of the wind energy production forecast preparation assessment methods, various (often imcomparable) criteria for the evaluation of such methods and their results are being utilized. Therefore, when the soft computing approach is applied, it is also important to compare different approaches and check their usefulness in this particular scenario.

The established prediction methodologies can be categorized into two categories: physical-based methods and statistical-based methods [9]. The first ones utilize atmospheric motion estimations (prediction of wind speed and its transformation into wind power) to predict long-term trend of wind process. The end result of this process is the acquisition of the wind power curve. Statistical methods describe the relations between historical time series of wind speed or power by recursive methods. The proper prediction model is designed based on training datasets (usually taking a tabelar form, like SCADA) and then used to predict output values for new input data [10]. Short-term forecasting is usually condicted with utilization of statistical methods as they require short operation time and lesser ammount of computational resources than physical-based methods.

The leap in development of computational power of modern computers over the last twenty years has enabled the extensive use of Artificial Intelligence algorithms on methorological Big Data. In [11] the authors utilize Support Vector Regressors for short-term wind power forecasting, while using a modified Dragonfly algorythm (a variant of swarm behavioral methodology) for optimization purposes to enhance prediction accuracy. In [12] the same regressors were used for the purposes of wind speed forecasting, with Jaya gradient-free algorithm (population-based method, which repeatedly modifies a population of individual solutions) as optimizer. The authors of [13] created a hybrid forecasting model based on convolutional neural networks and Informer algorithm (a network structure based on an attention mechanism meant for solving prediction problem of long series data) for short-term wind power prediction. In [14] authors proposed a hybrid system for wind speed prediction. The proposed method is based on backpropagation network implemented to derive the prediction results, which are deconstructed into different intervals according to the classification of different data features via fuzzy clustering. In [15] the authors investigate the performance of enhanced machine learning models to forecast univariate wind power time-series data. The methods employed here are Bayesian optimization for tuning the parameters of Gaussian proces regression, Support Vector Regression and Bagged Trees.

The aim of the analyses carried out and presented for the purpose of this paper was to develop an effective (with a low forecast error) forecast model for wind turbines based on Artificial Intelligence (AI) regression algorithm. The paper presents wind power generation production from the wind speed and wind direction with a one-day horizon.

The content of the paper is as follows. In Section II information about the measurement data is provided. In Section III data preprocessing methods used in the presented methodology are described. Section IV introduces the applied prediction algorithms. In Section V experimental results are discussed. Finally, Section VI contains summary and future prospects.

II. DATA CHARACTERISTICS

This Section contains the description of the dataset, upon which the predictions have been made.

The SCADA data analyzed in this paper comes from the wind turbine located in Esenköy, Çinarcik, Yalova, Marmara Region, Turkey, with geographical coordinates at 40.616118° latitude and 28.955617° longitude (Figure 2).

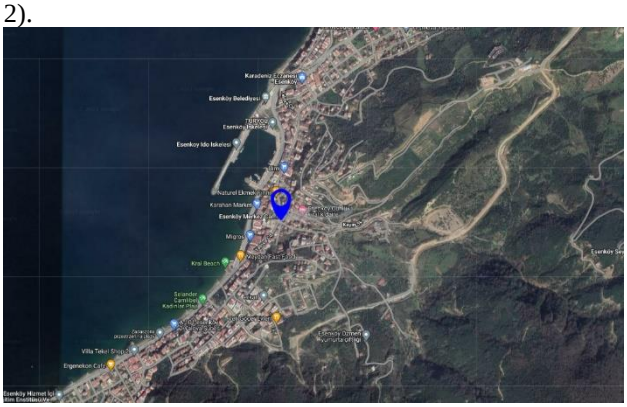

Fig. 2. Geographical location of the Yalova wind farm (Google Maps).

The dataset contains 50530 observations, which carry information about:

- Date and time of the registered measurement, given in 10-minute intervals;
- The power generated by the turbine (LV ActivePower) registered at the given moment and denoted in kW;
- wind speed at the hub height of the turbine denoted in m/s;
- theoretical power values provided by the manufacturer of the wind turbine taking into account the specific wind force (TheoreticalPowerCurve) – these values are

denoted in KWh;
- the wind direction in degrees (°) measured at the hub height of the turbine (this is the direction to which the wind turbines turn to automatically).

Examples of values for the above-mentioned characteristics are in Table 1. Such tabelar, numerical data is provided to the AI algorithm as the input information, based upon which the said algorithm produces a set of predictions (also denoted by numerical values) with a day-ahead prediction horizon.

Table 1: First five observation from the analyzed dataset.

Date/Time	LV Active Power [kW]	Wind Speed [m/s]	Theoretical Power Curve [KWh]	Wind Direction [°]
01 01 2018 00:00	380.047790527343	5.3113360449682	416.328907824861	259.994903564453
01 01 2018 00:10	453.769195555664	5.6721668243408 2	519.917511061494	268.641113281 25
01 01 2018 00:20	306.376586914062	5.2160367965698 2	390.900015810951	272.564788818359
01 01 2018 00:30	419.645904541015	5.6596741676330 5	516.127568975674	271.258087158203
01 01 2018 00:40	380.650695800781	5.5779409408569 3	491.702971953588	265.674285888671

Figures 3 and 4 show hourly and monthly average power production, respectively. In Figure 5 there is a polar diagram with wind speed, wind direction and power production from the collected dataset. Based on these figures one can state that the average power production is higher in the months of March, August and November and at the end of each day (after 16:00). The greatest amount of power can be produced if the wind blows from the directions between 0-90 degrees and 180-225 degrees. These details are related with the geographical location of the turbine.

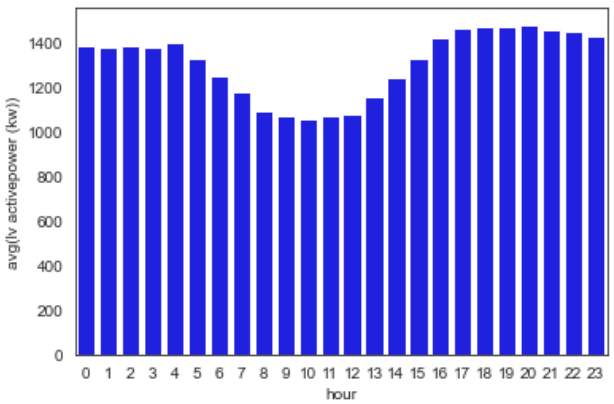

Fig. 3. Hourly average power production.

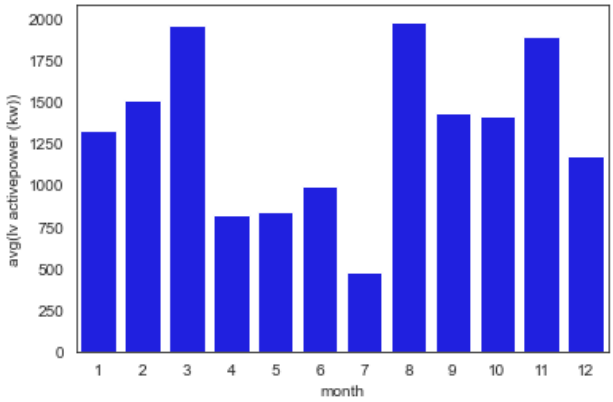

Fig. 4. Monthly average power production.

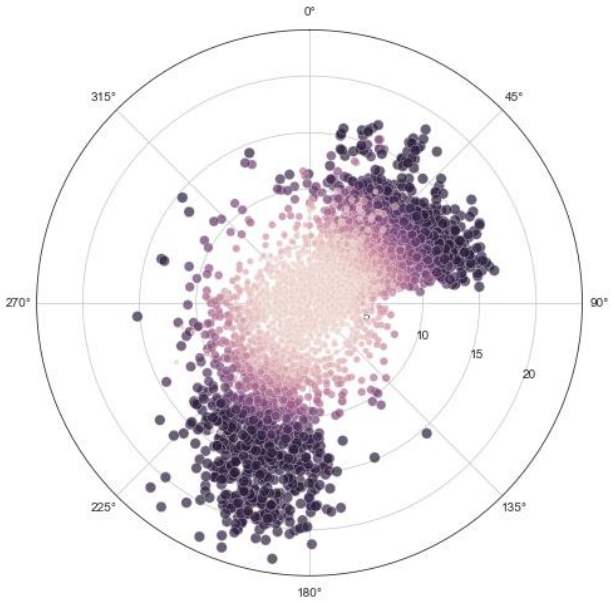

Fig. 5. Polar diagram of power production dependent on the direction and the speed of wind.

III. DATA PREPARATION

This Section provides information on how the data for prediction model were prepared.

In order for the SCADA data to be properly utilized by a soft computing algorithm, it has to be properly cleansed. All values for which the wind force does not exceed 2 m/s were removed from the dataset. This is due to the fact that according to the wind turbine manufacturer's statement, the turbine starts working at wind speed of at least 3 m/s. As its presented in Figure 6, the power production is at it's lowest when reaching 5m/s, while the near-maximum level is reached after the wind speed is about 15 m/s. These parameters strongly depend on the particular model of the farm and wind turbine, which has to be considered for each device individually.

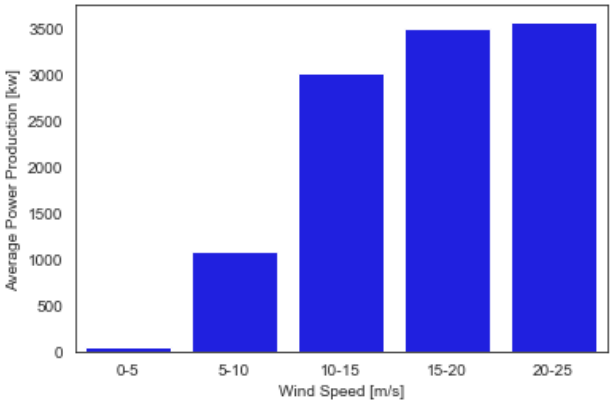

Fig. 6. Average power production for 5 m/s wind speed intervals.

Subsequently during the model estimation, missing data in the SCADA dataset is ignored. Supplementing the gaps with statistical methods turned out to be impossible due to the large size of the data variability. This would burden the final prognostic model.

The data was divided in the ratio 4/1, where 80% (namely 40321 examples) of the readings were used to construct the training data, used to estimate any parameters of a forecasting method, and 20% (i.e. 10209 examples) were used as testing data, evaluating the prediction accuracy.

Extreme outliers were also cleansed from the given dataset. These included production values, which were signifficantly too low for a given wind speed. According to the available data, there are certain months in a year, when the wind turbine doesn't produce any power, even though the average wind speed stays on the level of 4 m/s. This might indicate that during that period the turbine wasn't functioning properly. Therefore all such values (3497 observations total) were excluded from the dataset to facilitate the training process.

Described preprocessing operations allowed for the preparation of data directly for the construction of the forecasting models, which could then be trained and compared regarding the accuracy.

IV. FORECASTING METHODOLOGY

This Section provides details on the prediction methodology. First, the used algorithms are briefly introduced. Next, the prediction accuracy measures are introduced.

Figure 7 summarizes the process of wind power prediction with the utilization of contemporary AI regression methods. It can essentially be divided into three phases: the training and phases of the algorithm and finally quality evaluation of the proposed prediction methodology. During the training phase, the correlation between the input proces variables and the output variable is established based on the observations from the SCADA dataset that are known to the algorithm. During the testing phase, the AI algorithm is tasked to predict output values (wind power) for the unseen input data.

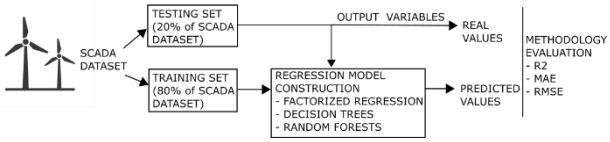

Fig. 7. Data flow chart of the wind energy prediction methodology, based on regression algorithms.

The AI algorithms utilized for the purpose of this study were Factorized Regression, Random Forests Regression and Decision Trees. The quality of each algorithm on this type of data has been evaluated with the utilization of regression evaluators. Each algorithm has been implemented in pyspark, which is an interface for Apache Spark in Python programming language. It utilizes a scalable machine learning library MLlib for the purpose of classification and regression of tabular data like DataFrame in Python.

Factorization machines are an extension of the linear regression model, designed for the purpose of capturing interactions between features within high dimensional sparse datasets. Factorized machines use a factorized parametrization and have a linear complexity and can be optimized in the primal:

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j+1}^{n} \langle v_i, v_j \rangle \, x_i x_j \qquad (1)$$

, where w_0 denotes bias, w denotes the weight of the i-th variable, v_i represents the i-th feature, $\langle \cdot, \cdot \rangle$ is the dot product of two vectors. $\langle v_i, v_j \rangle$ model the interaction between i-th and j-th feature. To achieve good generalization, 2-Way Factorization Machines were used. Even though the equations generalize to higher orders, it is not recommended, because of the loss of numerical stability of the optimization methods.

The decision tree performs a recursive binary partitioning of the feature space, where the tree predicts the same label for each leaf. Each partition is chosen greedily. It is performer by selecting the best split for a set of possible splits. The impurity measure (a measure of the homogeneity of the labels at the node) is variance:

$$\frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2 \qquad (2)$$

, where yi is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N} \sum_{i=1}^{N} y_i$.

The maximum tree depth is set to 5. The Mean Squared Error (MSE) is utilized to evaluate goodness of fit.

Random forests are ensamble of decision trees. They combine many decision trees in order to reduce the risk of overfitting. Random forests train a set of decision trees separately so to enable parallel training. The training process of each tree is randomized. The variance of predictions is at the end combined so to improve performance of the algorithm. Each decision tree predicts a real value which then are all averaged to predict a new label (Figure 8).

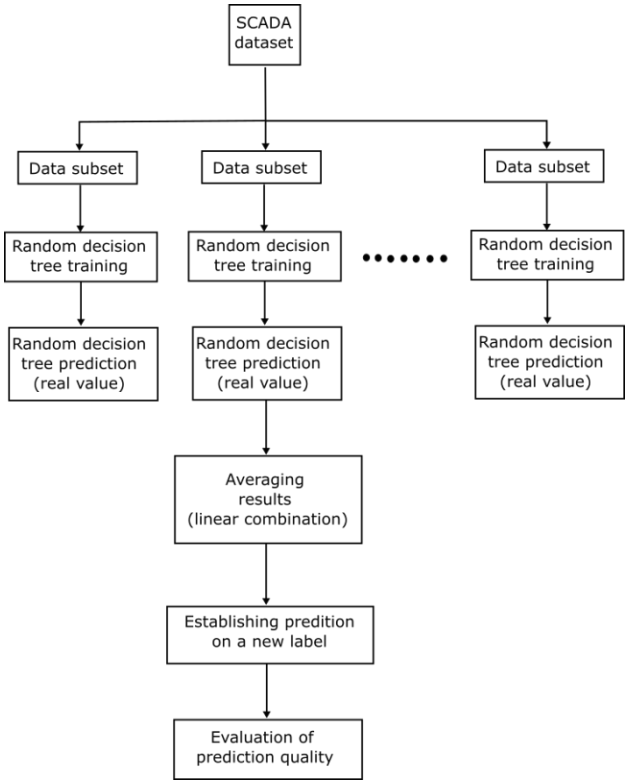

Fig. 8. Random forest model creation.

For the purpose of regression methodology evaluation, three regression quality measures were utilized: R2 Score (coefficient of determination), MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) [16].

R2 score is the proportion of the variation in the dependent variable that is predictable from the independent variable. It provides a measure of how well obcserved outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (1)$$

where SS_{res} is the residua sum of squares, while SS_{tot} is the total sum of squares.

MAE is calculated as the sum of absolute errors divided by the sample size. Here different errors are not weighted, but the scores increase linearly with the increase in errors. The MAE score is measured as the average of the absolute error values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2 \quad (2)$$

where y_i is the prediction and $\hat{y}_i$ is the true value, while n denotes the total amount of observations.

RMSE (Root Mean Square Error) denotes the square root of mean squared difference between the true values and the predicted ones.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (3)$$

where y_i is the prediction and $\hat{y}_i$ is the true value, while n denotes the total amount of observations.

V. EXPERIMENTAL RESULTS

This section presents the results of data analysis and utilized methodology efficiency. In order to test various algorithms methods of regression of the dataset, it was important to draw a wind farm power curve in regards to the wind speed based on the data contained in the dataset and the predicted values.

Figure 9 presents curves of the amount of produced wind energy based on the wind speed. The black color represents the theoretical production, the red color represents the actual production while the blue color represents the predicted production. Both theoretical and predicted curves fit with the real energy production curve.

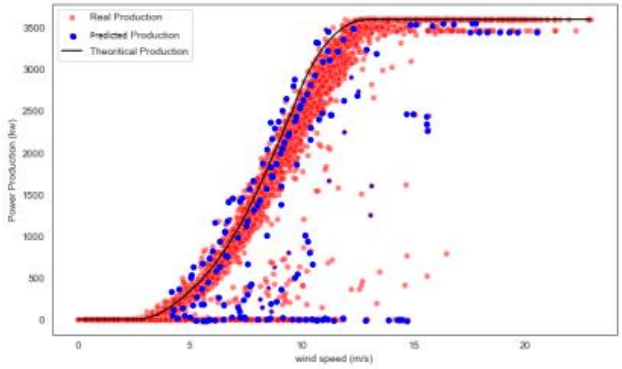

Fig. 9.Wind turbine power production prediction base on wind speed.

Furthermore it can be read from the graphic that the energy production reaches its full potential of 3500 kW at the 13 m/s of wind speed. The maximum value of the wind speed is 19m/s, but these values are very rare in the dataset (407 observations). The average power production at this wind speed is 3566.5 kW. The uneven spread of points above and below the characteristic may be related to the uneven distribution of wind speed inside the wind turbine.

Table 2 presents the values of the utilized error measurement methods (R2 Score, MAE and RMSE).

Table 2: Results of the utilized error evaluation methods.

Error Measuring Method	Value
Factorized Regression	
R2	0.6603720145184516
MAE	572.3217767482964
RMSE	763.2129017697233
Random Forest Regression	
R2	0.9211277183049995
MAE	209.51988580915867
RMSE	369.2358225661089
Decision Tree Regression	
R2	0.9331849691596146
MAE	144.51627124894884
RMSE	340.69963678142705

Though all accuracy evaluation criteria are of different ranges, it is possible to compare subsequent methods

regarding the same factor.

VI. CONCLUSIONS

Based on the experiments conducted on the Yalova wind turbine dataset with the utilization of various regression methods for prediction of wind energy production based on the wind speed it can be stated, that the best method (the one which produces the smallest efficiency evaluation errors) is the one based on the decision tree regression method. Out of the three algorithms utilized, decision tree has the smallest amount of MAE and RMSE value (144.5 and 340 respectively).

Future research will include more factors determining the energy produced (including the internal state of the turbine itself). Also, additional regression methods for prediction should be used to compare their efficiency. Finally, the forecasting time horizon should be evaluated (i.e. how far in the future the predictions can be made).

REFERENCES

[1] **Yang, B., Wang, J., Zhangp, X., Yu, T., Yao, W., Shu, H.,** Comprehensive Overview of Meta-Heuristic Algorithm Applications on Pv Cell Parameter Identification. *Energy Conversion and Management*. Vol. 208, 2020.

[2] **Vennila C., Tits A., Sudha T.Sri, Sreenivasulu U., Pandu Ranga Reddy N., Jamal K., Lakshmaiah Dayadi, Jagadeesh P., Belay A.**, Forecasting Solar Energy Production Using Machine Learning, *International Journal of Photoenergy*, vol. 2022, 2022.

[3] **de Falani, S.Y.A., Gonzalez, M.O.A., Barreto, F.M., de Toledo, J.C., Torkomain, A.L.V.,** Trends in the technological development of wind energy generation, *International Journal of Technology Management & Sustainable Development*, Vol. 19, No. 1, 2020, pp. 43-68(26).

[4] **Zhou Wu, Gan Luo, Zhile Yang, Yuanjun Guo, Kang Li, Yusheng Xue,** A comprehensive review on deep learning approaches in wind forecasting applications, CAAI Transactions on Intelligence Technology, vol 7, issue 2, 2022, pp. 129-143.

[5] https://wwindea.org/world-market-for-wind-power-saw-another-record-year-in-2021-973-gigawatt-of-new-capacity-added/

[6] **Hearps, P., i McConnell, D.**, Renewable energy technology cost review, *Melbourne Energy Institute*, 2011, p.57.

[7] **Kazi, S.**, Adaptation of energy production to forecast values using external storage, *Acta Universitatis Sapientiae Electrical and Mechanical Engineering*, vol. 3, 2011, pp. 51-60.

[8] **Pinson, P.**, Wind energy: Forecasting challenges for its operational management, *Statistical Science, Special Issue on Mathematics of Planet Earth*, vol. 28, no. 4, 2013, pp. 564-585.

[9] **Tascikaraoglu A., Uzunoglu M.**, A review of combined approaches for prediction of short-term wind speed and power, *Renewable and Sustainable Energy Reviews*, vol. 34, 2014, pp. 243–254.

[10] **Renani E. T., Elias M. F. M., Rahim N. A.**, Using data-driven approach for wind power prediction: A comparative study, *Energy Conversion and Management*, vol. 118, 2016, pp. 193–203.

[11] **Li L-L, Zhao X., Tseng M-L, Tan R.R.**, Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm, *Journal of Cleaner Production*, vol. 242, 2020.

[12] **Mingshuai L., Zheming C., Jing Z., Long W., Chao H., Xiong L.**, Short-term wind speed forecasting based on the Jaya-SVM model, *International Journal of Electrical Power & Energy Systems*, vol. 121, 2020.

[13] **Wang H-K, Song K., Cheng Y.**, A hybrid forecasting model based on cnn and informer for short-term wind power, *Frontiers in Energy Research*, vol. 9, 2022.

[14] **Ne Y., Bo H., Zhang W., Zhang H.**, Research on Hybrid Wind Speed Prediction System Based on Artificial Intelligence and Double Prediction Scheme, *Complexity*, vol. 2020, 2020.

[15] **Alkesaiberi A., Harrou F., Sun Y.**, Efficient Wind Power Prediction Using Machine Learning Methods: A Comparative Study, *Energies*, no. 15, 2022.

[16] **Hyndman, R.**J.**, Koehler, A.B**, Another look at measures of forecast accuracy, *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688.