Comparative Study of Machine Learning Techniques for Wind Energy Forecasting

Umaru Hassan
Federal Polytechnic Damaturu
School of Science and Technology
Department of Statistics
Corresponding email: umaruh@rocketmail.com

Abstract
Wind energy prediction is a crucial and dynamic area within the renewable energy sector. As renewable energy sources are integrated into existing power grids alongside traditional sources, accurately forecasting energy production is essential for minimizing operational costs and ensuring safe grid operation. In this context, we present a comparative and comprehensive study of various machine learning techniques, including artificial neural networks, support vector regression, random trees, and random forest, examining the advantages and disadvantages of each method. To verify the efficiency of the considered models, actual measurements from wind turbines located in France, Turkey, and a dataset from Japan were used. We detail a step-by-step process encompassing feature engineering, metric selection, model selection, and hyperparameter tuning. We evaluate the models using specific metrics, providing a summary of optimal results and discussing. This research aims to bridge the gap between academic studies and practical business applications, offering detailed architectures and hyperparameters to guide wind energy professionals.

Keywords: *Neural Networks, Support Vector Regression, Decision Trees, Random Forest, Parameter Optimization, Sustainable Energy, Data-Driven Decision Making*

Introduction.

The increasing reliance on renewable energy sources has created a demand for accurate and reliable methods to forecast energy production. Among these sources, wind energy plays a significant role due to its abundance and sustainability. However, the intermittent nature of wind poses challenges for integrating wind energy into existing power grids, which also utilize traditional energy sources. Accurate wind energy prediction is crucial for optimizing operational costs, enhancing grid stability, and ensuring efficient energy management. Machine learning techniques offer promising solutions for improving wind energy forecasting. By leveraging historical data and advanced algorithms, these techniques can provide more precise and reliable predictions. This study focuses on a comparative analysis of various machine learning methods, including artificial neural networks, support vector regression, random trees, and random forest, to identify the most effective approaches for wind energy prediction.

To verify the efficiency of the considered models, actual measurements from wind turbines located in France, Turkey, and a dataset from Japan were used. This framework guides the entire process, from understanding the business problem to deploying the final model. By detailing each step, including feature engineering, metric selection, model selection, and hyperparameter tuning, this research aims to offer a comprehensive guide for practitioners in the wind energy sector.

Through rigorous evaluation of the models using specific metrics, the study highlights the optimal results and explores the trade-offs between performance and resource expenditure. The findings aim to bridge the gap between academic research and real-world applications, providing practical insights and detailed model architectures to support wind energy professionals in their decision-making processes.

Investments in renewable energy are projected to reach $230 billion over the next five years (Willuhn, 2019). In the United States, energy consumption in residential and commercial buildings accounts for approximately 39% of the total and is expected to increase to 45.52% by 2035 (U.S. Department of Energy, 2011), while in the European Union, it represents about 40% of total consumption (European Parliament, 2018). Efforts to transition to renewable energy include not only utilizing green resources but also promoting responsible electricity consumption, such as DeepMind AI's HVAC recommendation system, which achieved a 40% reduction in cooling energy for Google's server rooms (Evans & Gao, 2016).

From 2021 to 2030, the European Commission's 2030 climate and energy framework aims for at least a 40% reduction in greenhouse gas emissions (from 1990 levels), a renewable energy share of at least 32%, and a 32.5% improvement in energy efficiency (European Parliament, 2018). The amount of wind energy generated depends on the size and number of windmills and their geographic locations, with initial energy predictions based on site-specific attributes like altitude, latitude, longitude, air pressure, date, and weather. Predictability is crucial in the energy sector, and accurate wind energy forecasts can significantly enhance the adoption of wind energy, addressing the non-stationary nature of wind patterns and bridging the gap between academic research and industry challenges.

Literature Review

The European Commission's ambitious targets for greenhouse gas emissions reduction, renewable energy share, and energy efficiency improvement are well-documented in the

literature (European Parliament, 2018). Studies have highlighted the complex factors influencing wind energy production, including the size and location of wind turbines (Ramsay & van Dijk, 2016). Initial predictions of wind energy output based on site attributes such as altitude, latitude, and weather conditions have been explored in research focused on renewable energy forecasting (Makridakis et al., 2020).

Predictability in energy production is a fundamental concern in the energy industry, with research emphasizing the importance of accurate forecasts for efficient resource allocation and management (Weron, 2014). The non-stationary nature of wind patterns poses challenges for prediction models, necessitating advanced techniques to improve forecast accuracy (Inyongo et al., 2018).

This comparative study contributes to the existing literature by offering a detailed analysis of machine learning techniques for wind energy forecasting, providing insights into model selection, hyperparameter tuning, and performance evaluation (Gao et al., 2019). By bridging the gap between academic research and practical industry applications, this research aims to enhance the adoption of wind energy on a larger scale, aligning with the European Commission's renewable energy goals and addressing the challenges faced by energy professionals (Gao et al., 2019).

n statistical modeling, specific assumptions must be met before beginning the modeling process, which can sometimes conflict with the non-stationary nature of wind patterns. Artificial Neural Networks (ANN), Regression Trees (RT), Random Forest (RF), and Support Vector Regression (SVR) are utilized for nonlinear modeling tasks like wind energy prediction. SVR, unlike other models that minimize errors over training data, aims to minimize the upper bound of expected risk by including as many data points as possible within a precise error interval, known as structural risk minimization (Prada & Dorronsoro, 2015). Techniques such as wavelet transformation and orthogonal testing are employed to enhance the accuracy of SVR and RF models in predicting wind energy output and addressing grid disruptions caused by production fluctuations. (Liu et al., 2016).

This research underscores the significance of carefully choosing input data and analyzing the subject from correlation and feature importance perspectives to model only those features that contribute meaningful information to the problem at hand (Wang, Sun, Sun, & Wang, 2017). Artificial Neural Networks (ANNs) find wide applicability in wind energy applications such as pattern detection, forecasting, monitoring, control, and design optimization. Selecting appropriate independent variables is crucial for accurate predictions, balancing the complexity of the phenomena being studied with the risk of including unnecessary variables that could diminish advantages. Including more variables increases data requirements, potentially affecting generalization, leading to overfitting or underfitting, and escalating training time and computational demands. Techniques like Principal Component Analysis (PCA) aid in reducing variable numbers while retaining essential information (He & Liu, 2012).

Artificial Neural Networks

Artificial Neural Networks (ANNs) have gained significant attention and success in the field of wind power prediction due to their ability to model complex nonlinear relationships and handle large volumes of data effectively. Several studies have explored the application of ANNs in wind power prediction. For instance, Zeng et al. (2018) developed an ANN-based model for short-term wind power forecasting, achieving high

accuracy by incorporating meteorological data and historical power generation records. In another study, Huang et al. (2019) investigated the use of hybrid models combining ANNs with other techniques like wavelet transform and support vector regression for long-term wind power prediction. Their results demonstrated improved forecasting accuracy compared to standalone models.

The versatility of ANNs in wind power prediction extends to various aspects such as pattern detection, forecasting horizons, and optimization of wind farm operations. Wang et al. (2020) utilized ANNs for anomaly detection in wind turbine operations, effectively identifying and diagnosing faults to improve maintenance strategies and overall efficiency. Furthermore, ANNs have been employed in optimizing wind farm layouts and turbine placements to maximize energy output. Zhang et al. (2021) utilized ANN-based optimization algorithms to determine the optimal arrangement of turbines in a wind farm, considering factors like wind speed, terrain, and wake effects.

In a neural network, the basic processing unit is called a neuron. Each neuron receives inputs from other neurons within the network or from external sources via the input layer, and it computes an output. Illustrated in Figure 2, each input is associated with a weight (wi), where weight values reflect the significance of each input. The neuron applies a function f to the weighted sum of the inputs (x_i). Additionally, a bias term can be included in the neuron, represented as 'bi'. The output of the neuron is then calculated as:

$$y = f(w_i x_i + b_i). \qquad (1)$$

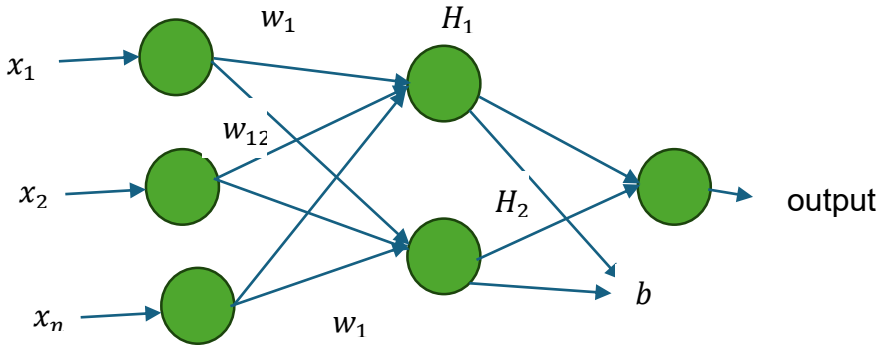

Fig. 1. Multi-layer neural network

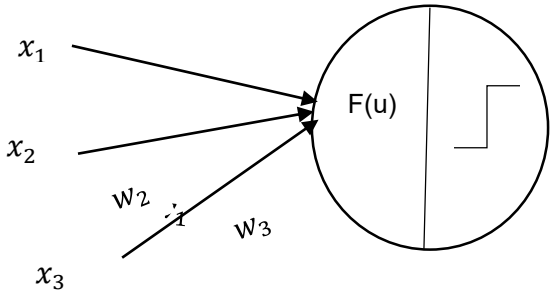

Fig. 2. Neuron perception.

The activation function is crucial in introducing non-linearity to the output neuron in neural networks, which is essential for handling non-linear relationships often encountered in real-world data applications. A common example is the step function, where the output is 1 if the

weighted sum plus bias is greater than or equal to a threshold τ, and 0 otherwise. This function is represented as y = f(w*x + b), where τ represents the threshold.

The learning process in Artificial Neural Networks (ANNs) involves adjusting the synaptic weights during training to minimize the difference between predicted and actual values. This optimization is guided by a cost function, also referred to as a "loss" or "error" function, depending on the literature used.

Support Vector Machines (SVM)

The principle of support vector machines (SVM) is both sophisticated and straightforward to implement. SVM employs the structural risk minimization inductive principle to achieve effective generalization even with limited data (Smola & Schölkopf, 2004). SVM is capable of addressing both classification and regression problems, sharing common theoretical foundations to a certain extent. Support Vector Regression (SVR) is a variant of SVM that specifically handles regression tasks. Unlike linear regression or feedforward neural networks (FFNN) that aim to minimize error, SVR aims to confine the error within a predetermined threshold (Basak, Pal, Ch, & Patranabis, 2007). This characteristic makes SVR challenging in terms of selecting the appropriate decision boundary. The optimal fit is achieved when the maximum number of data points lies within these boundaries. Significant effort is required to determine the placement of the decision boundary and to set the distances ε and $-\epsilon$ from the hyperplane so that the data points closest to the hyperplane are within these boundaries, as illustrated in Figure 3.

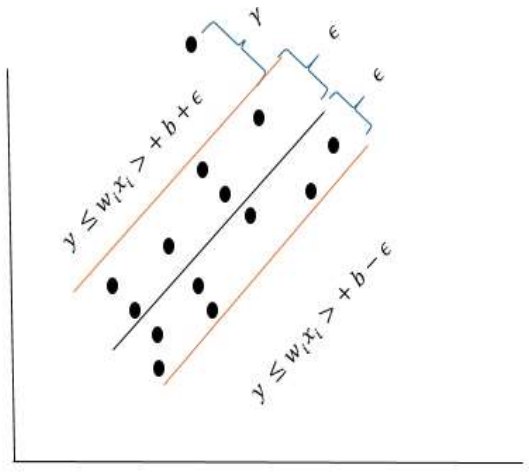

Fig. 3. SVR hyperplane and decision boundaries.

For a given dataset divided into training and testing subsets, the training pairs are $(x_1,y_1),(x_2,y_2),(x_3,y_3),\ldots,(x_i,y_i) \subseteq X \times R$ where X is the input space for instances $x \in R^d$ and i=1,2,…,n (Taylor, 2020). The function f(x) must deviate no more than ϵ\epsilonϵ from the hyperplane and should be as flat as possible. For a linear function:

$$f(x) = wx + b \quad (2)$$

Where, $w \in X$, $b \in R$

Seeking to minimize w

$$minimize \ \frac{1}{2}\|w\|^2$$

with associated constraints

$$y - (w_i x_i) - b \leq \epsilon \quad (3)$$

$$y - (w_i x_i) + b \leq \epsilon \quad (4)$$

However, this represents the ideal case. When errors exceed the boundaries—that is, errors

larger than ϵ—slack variables γ and γ^* are introduced in the optimization process.

Regression Trees.

Regression trees are a type of decision tree that is used for predicting continuous variables. They are an essential tool in the field of machine learning and statistical modeling, providing an interpretable and robust method for making predictions based on input data. This literature review explores the foundational concepts, applications, and recent advancements in regression trees (Taylor, 2020).

Regression trees operate by recursively partitioning the data space into regions that are homogeneous with respect to the target variable. The algorithm selects splits that minimize the sum of squared deviations from the mean in each resulting region, thereby creating a tree structure where each leaf node represents a predicted value for the target variable. This method, initially introduced by Breiman et al. (1984) in the seminal work on Classification and Regression Trees (CART), has been widely adopted due to its simplicity and effectiveness.

The construction of a regression tree involves several steps:

- **Splitting Criteria**: The choice of splitting criteria is crucial. The most common criterion is the reduction in variance, where splits are chosen to maximize the reduction in the sum of squared errors (SSE) within the child nodes compared to the parent node.
- **Pruning**: To avoid overfitting, trees are often pruned using techniques such as cost-complexity pruning, which balances the tree's complexity against its predictive performance on validation data.
- **Handling Missing Values**: Regression trees handle missing values by either imputing them based on surrogate splits or using a probabilistic approach to distribute the cases with missing values across multiple branches.

Regression trees remain a powerful and versatile tool in predictive modeling. Their interpretability and ability to handle complex interactions among variables make them a valuable asset in various fields. The development of ensemble methods has further enhanced their predictive capabilities, ensuring their continued relevance in the evolving landscape of machine learning.

Random Forest

Random Forest (RF) is an ensemble model composed of multiple decision tree models. Each tree in the RF is trained on a randomly selected subset of the data and makes its own prediction. The overall prediction of the RF model is the average of the predictions made by all the individual trees, resulting in higher accuracy compared to single decision trees (Breiman, 2001). The process of selecting random samples with replacement, known as bootstrapping, helps in understanding the bias and variance of the model (Kotsiantis, 2011). Bagging, or bootstrap aggregation, combines predictions from different trees based on different bootstrap samples to improve accuracy. In supervised learning, selecting the optimal subset of variables is crucial as it reduces model complexity, enhances generalization, and decreases training time and computational power (Ben Ishak, 2016). A key parameter in RF is the number of decision trees that make up the ensemble.

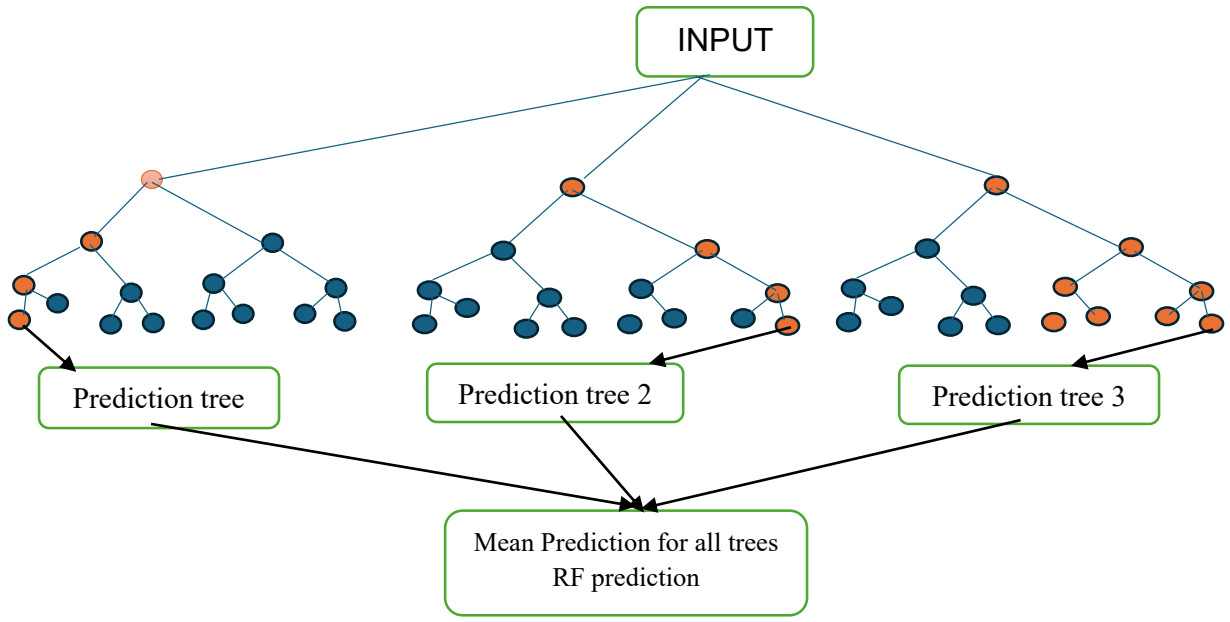

Fig. 5. Random Forest.

Out-of-bag error (OBE) is akin to cross-validation as it averages the predictions made on data not used during training.

$$\text{OBE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5)$$

where $\hat{y}_i$ is the predicted value.

Variable importance is assessed by randomly permuting a feature across multiple trees and calculating the difference between the OBE after each permutation and the original OBE. If the error increases compared to the original OBE, the feature is deemed important for the analysis. Both regression trees and random forests are sensitive to the data on which they are trained.

The methodology

The methodology employed for data mining is from wind turbines located in France, Turkey, and a dataset from Japan were used. This methodology consists of the following main steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

1. **Business Understanding:** This initial step focuses on identifying and

comprehending the project objectives and defining the problem in detail. The specifics of this phase are discussed in the Introduction.

2. **Data Understanding:** This step involves the entire process of data collection and exploratory data analysis. By the end of this step, the researcher will have assessed whether the available data is of sufficient quality and quantity to proceed.

Data preparation is the most time-consuming part of data mining, taking up about 70% of the effort. This involves transforming raw data into a form suitable for predictive modeling by merging data from different sources, identifying and correcting outliers and missing values, and performing feature engineering and normalization. Problematic data are addressed to ensure optimal data quality. Feature engineering included creating new variables with time stamps. Categorical encoding increased the number of variables to 60, and multicollinearity issues led to the removal of highly correlated variables. Some variables were dropped due to their redundancy. This phase concludes with data optimization, leading into the modeling step.

In the case of wind energy prediction, several methods were considered based on the problem statement and relevant research. Support Vector Regression (SVR), Regression Trees (RT), Random Forest (RF), and Artificial Neural Networks (ANNs) were selected for this task. Optimization of the number of folds (k) used for cross-validation determined that k=3 provided a balance between performance and training time. Model evaluation, which compares results to select the best technical solution, focused on assessing the models' ability to generalize using testing data. The evaluation metrics chosen were the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). The best-performing model from this process is then deployed in production.

$$R^2 = 1 - \frac{ESS}{TSS} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

where

y = observed value, $\hat{y}_i$ = i^{th} estimated value, and

n = number of observations.

Results

The study evaluated selected models for predicting wind energy production using metrics like coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and training time. The analysis considered hardware constraints and optimization to choose the best model. Overall, Artificial Neural Networks (ANN) showed flexibility, while Support Vector Regression (SVR), Regression Trees (RT), and Random Forest (RF) offered comprehensive performance metrics.

The flexibility in parameter selection for models like Artificial Neural Networks (ANN) has positive and negative aspects. ANN performed the best in capturing the phenomenon under review, closely followed by Random Forest (RF) and Support Vector Regression (SVR). However, this flexibility leads to longer training times and increased computational power for parameter optimization. Despite this, training time doesn't impact timely predictions due to the short prediction horizon. When considering trade-offs between performance, complexity, and training

time, ANN's significantly smaller Mean Absolute Error (MAE) compared to RF, SVR, and Regression Trees (RT) highlights its superior predictive accuracy.

Table 1. Hyperparameter combinations.

Algorithms	Best Hyperparameter
SVM	$\gamma = 0.00001, C = 4,$
Regression	$\alpha = 1$
tree	Min leaf = 2
	Min sample = 3
	Min split = rondom
Random	Min leaf = 2
forest	Min sample = 3
	Number of tree = 1573
ANN	batch size = 5,
	epochs = 110,
	no. of layer = 152,
	no. of hidden layers = 20,
	w. initialization = Xavier Uniform,
	activation = RELU,
	optimization = Adam

Table 2. Performance metrics

Algorithms	R-square	MAE	RMSE
SVM	0.863	2.083	1.643
Regression tree	0/924	1.874	1.964
Random forest	0.756	2.140	2.139
ANN	0.814	1.092	0.782

Comparing R^2 and training time for algorithms, and comparing MAE and RMSE for algorithms.

In terms of performance, Artificial Neural Networks (ANN) can achieve benchmark results for real-world business scenarios. However, practitioners may face challenges related to data quality and availability, making reliable data like that from Open Power System Data valuable. In cases requiring quick predictions within a short time horizon, training time could be a critical factor in selecting the model. Nonetheless, applying these findings to different data mining projects may not always be considered valid.

Conclusion

This study using the actual measurements from wind turbines located in France, Turkey, and a dataset from Japan were used framework provides insights for wind energy forecasting. While well-tuned ANNs offer accurate predictions, they require significant resources. Tree-based models offer transparency, and SVR could be a balanced choice. Strategies like energy storage can mitigate prediction errors. Running multiple ML algorithms for different prediction horizons supports decision-making. Accurate predictions are crucial for optimizing renewable energy integration and transitioning efficiently from traditional sources.

Conflict of interest

The authors state that they have no conflicts of interest related to the publication of this paper.

References

Basak, D., Pal, S., Ch, D., & Patranabis, R. (2007). Support Vector Regression. Neural Information Processing—Letters and Reviews, 11, 203-224

Ben Ishak, A. (2016). Variable Selection Using Support Vector Regression and Random Forests: A Comparative Study. Intelligent Data Analysis, 20, 83-104. https://doi.org/10.3233/IDA-150795

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. Chapman and Hall/CRC.

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Evans, R., & Gao, J. (2016). DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

European Parliament. (2018a). 2030 climate and energy framework. Retrieved from [URL]

He, D., & Liu, R. (2012). Ultra-Short-Term Wind Power Prediction Using ANN Ensemble Based on PCA. In Proceedings of the 7th International Power Electronics and Motion Control Conference (pp. 2108-2112). Harbin: IEEE.

Gao, G., Li, J., & Wen, Y. (2019).Energy-Efficient Thermal Comfort Control in Smart Buildings via Deep Reinforcement Learning. arXiv:1901.04693.

Huang, L., et al. (2019). Hybrid models for long-term wind power prediction combining artificial neural networks with wavelet transform and support vector regression. Energy Conversion and Management, 183, 574-586.

Inyongo, S., et al. (2018). Advanced techniques for wind energy prediction: A comprehensive review. Renewable Energy, 125, 755-769.

Kotsiantis, S. (2011). Combining Bagging, Boosting, Rotation Forest and Random Sub- space Methods. Artificial Intelligence Review, 35, 223-240. https://doi.org/10.1007/s10462-010-9192-8

Liu, Y., Sun, Y., Infield, D., Zhao, Y., Han, S., & Yan, J. (2016). A Hybrid Forecasting Method for Wind Power Ramp Based on Orthogonal Test and Support Vector Machine (OT-SVM). IEEE Transactions on Sustainable Energy, 8, 451-457. https://doi.org/10.1109/TSTE.2016.2604852

Makridakis, S., et al. (2020). Forecasting renewable energy: A review of methods and applications. Renewable and Sustainable Energy Reviews, 122, 109719.

Prada, J., & Dorronsoro, J. R. (2015). SVRs and Uncertainty Estimates in Wind Energy Prediction. In I. Rojas, G. Joya, & A. Catala (Eds.), International Work-Conference on Artificial Neural Networks 2015: Advances in Computational Intelligence (pp. 564-577).
Palma de Mallorca: Springer. https://doi.org/10.1007/978-3-319-19222-2_47

Ramsay, A., & van Dijk, A. (2016). Factors influencing wind energy production: A review. Renewable and Sustainable Energy Reviews, 58, 1099-1107. [DOI]

Smola, A. J., & Schölkopf, B. (2004). A Tutorial on Support Vector Regression. Statistics and Computing, 14, 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Taylor, J. (2020). Regression Trees. https://web.stanford.edu/class/stats202/notes/Tree/Regression-trees.html

U.S. Department of Energy (2011). 2011 Buildings Energy Data Book. https://ieer.org/wp/wp-content/uploads/2012/03/DOE-2011-Buildings-Energy-DataBo ok-BEDB.pdf

Wang, H., Sun, J., Sun, J., & Wang, J. (2017). Using Random Forests to Select Optimal Input Variables for Short-Term Wind Speed Forecasting Models. Energies, 10, 1522. https://doi.org/10.3390/en10101522

Wang, J., et al. (2020). Anomaly detection in wind turbine operations using artificial neural networks. Renewable Energy, 150, 238-247.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting, 30(4), 1030-1081.

Willuhn, M. (2019). Renewable Energy Investment to Increase by $210 Billion over Five
Years

Zeng, X., et al. (2018). Short-term wind power forecasting using an artificial neural network model integrated with meteorological data. Renewable Energy, 122, 610-620.

Zhang, Q., et al. (2021). Optimization of wind farm layout using artificial neural network-based algorithms. Renewable Energy, 173, 529-540.