International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.
© 2025 Scholar AI LLC.
https://ijpsat.org/

Vol. 50 No. 1 April 2025, pp. 442-452

Predictive Optimization of Wind Power: In-Depth Comparison of Decision Tree, Random Forest, and LightGBM Models for Maximum Accuracy

Jean Marc Fabien Sitraka RANDRIANIRINA, Noelinihaja Solofoniaina Lovasoa Feno Fanantenana RAKOTOMALALA, Bernard Andriamparany ANDRIAMAHITASOA, Zely Arivelo RANDRIAMANANTANY

Laboratoire de Thermodynamique, Thermique et Combustion (LTTC),

Université d'Antananarivo, MADAGASCAR

Abstract— This research investigates the prediction of wind turbine active power using machine learning techniques. The aim is to enhance forecast reliability in the face of changing environmental factors. Three models are examined: Decision Tree, Random Forest, and LightGBM. The input features include wind speed, wind direction, and theoretical power, recorded over a full year at ten-minute intervals. The data are divided into 80% for training and 20% for testing. The Decision Tree model produces the highest prediction error. The Random Forest model improves accuracy by reducing variability. LightGBM delivers the most accurate results, with the lowest RMSE =6,39% and strong agreement with actual values. This approach highlights the effectiveness of ensemble models, especially LightGBM, for wind power prediction. These models offer valuable support for operational planning and predictive maintenance in wind energy systems.

Keywords— Wind power forecasting, Machine learning, Random Forest, LightGBM, Renewable energy.

I. INTRODUCTION

Wind power accounted for more than 10% of global electricity production. This share is expected to increase in the coming years. However, managing wind energy production presents challenges due to the variability of weather conditions, which directly affect electricity generation [1].

In the study by Dhungana (2025) [2], Random Forest and Decision Tree models were used for predictive maintenance of wind turbines. The results showed that the Random Forest model, trained on historical production and maintenance data, achieved a root mean square error (RMSE) of 0.15, indicating good accuracy in failure prediction.

In 2023, Liu et al. [3] combined LightGBM with digital filtering techniques to improve short-term forecasting accuracy. Using environmental sensor data collected over a six-month period, the model achieved an RMSE of 0.12, outperforming traditional methods.

Kazmi et al. (2023) [4] developed an architecture combining CNN and RNN to capture both spatial and temporal features of wind data. By using numerical weather forecasts and production data from multiple wind farms, the model outperformed LightGBM and other tree-based regressors, although exact RMSE values were not specified.

Around 2022, Drisya et al. [5] used the Random Forest model to forecast wind speed, a key factor in estimating wind power. By training the model with two weeks of data and forecasting up to three years ahead, they achieved an RMSE below 1.5 m/s for short-term predictions, indicating good accuracy.

This study applies several machine learning models to wind power forecasting. The models examined are Decision Tree, Random Forest, and LightGBM. The goal is to analyze their ability to estimate wind power. The methodological approach includes data collection, model training, and results analysis. These principles aim to optimize the energy management of wind farms and improve renewable energy forecasting tools.

II. DATA USED AND PREPROCESSING

The wind turbine under consideration has a rated power of 3,618.73 kW. It operates optimally at a wind speed of 7.10 m/s and reaches its maximum performance at 13 m/s. Its rotor, with a diameter of approximately 120 meters, sweeps an area of 11,310 m², maximizing wind energy capture. With a mass of 120 tons, it maintains essential structural stability, particularly under extreme conditions. It is equipped with three fiberglass-reinforced composite blades, ensuring both durability and aerodynamic efficiency. The generator, which is slot less, brushless, and uses permanent magnets, provides reliable and continuous electricity production. The electrical system operates at a voltage of 120/240 VAC, with a frequency range of 59.3 to 60.5 Hz.

The data were collected over a one-year period, with a time step of ten minutes, allowing for detailed analysis of parameter variability. The dataset includes five main temporal variables: date and time, active power (in kW), wind speed (in m/s), the theoretical production curve (in kWh), and wind direction (in degrees). For modeling, 80% of the data are used for training, enabling the models to be tuned to the specific characteristics of the system under study. The remaining 20% are used to test the model's robustness by evaluating its performance on unseen data. This approach ensures a reliable assessment of the models under real-world operating conditions.

Figure 1 shows the distribution of active power (LV ActivePower (kW)) after data filtering. This analysis is useful for understanding extreme production behaviors in turbines or identifying data that require preprocessing before applying a predictive model.

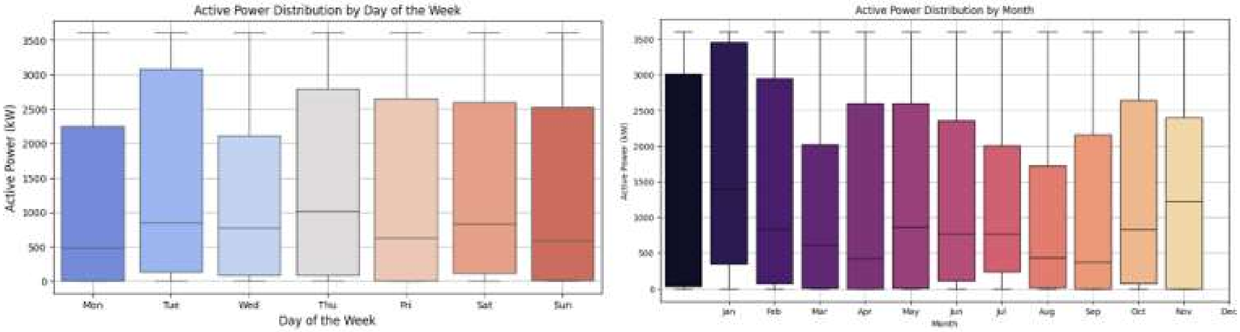

Figure 1: Distribution of Active Power

III. METHODOLOGY OF MACHINE LEARNING MODELS

III.1 Analysis of Correlation Between Numerical Variables

Figure 2 shows a correlation matrix between the variables. This visualization illustrates the strength and direction of the linear relationships between these variables.

The correlation between active power and wind speed is 0.94: This is a strong positive correlation, confirming that an increase in wind speed is directly related to an increase in power produced by the wind turbine, up to a certain threshold.

The correlation between active power and wind direction is -0.07: The relationship is almost nonexistent (value close to 0), indicating that wind direction has no significant impact on the power produced, likely due to the wind turbine's mechanism that automatically aligns itself with the wind.

The correlation between wind speed and wind direction is -0.07: This relationship is also weak, reflecting the relative independence of these two variables in the analyzed data.

This analysis helps identify the most relevant variables for modeling. Wind speed is a key variable for predicting active power, while wind direction is used for secondary analyses.

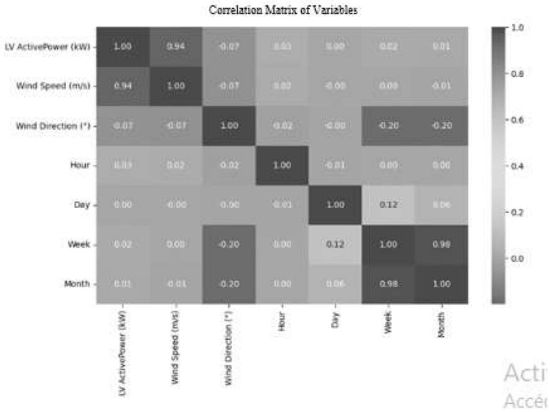

Figure 2 : Correlation Analysis

III.2 Handling of Temporal Variables

For the extraction of temporal features from the "Date/Time" column, information such as month, day, hour, minute, and season is retrieved. To create temporal categories, the season (winter, spring, summer, autumn) and time of day (morning, afternoon, evening, night) are derived from the hour. These new categories can be used for machine learning model analysis. The probability density is determined based on the Weibull distribution relationship [6]:

$$y = \frac{k}{\lambda}\left(\frac{V}{\lambda}\right)^{k-1} e^{-\left(\frac{V}{\lambda}\right)^{k}} \qquad (1)$$

Where,

y : probability density for a standardized wind speed V (unitless),

V : standardized wind speed (unitless),

k : shape parameter that determines the form of the curve,

λ : standardized scale parameter.

If $k > 2$: the curve is more symmetric with a well-defined peak.

If $k < 2$: the curve is more spread out and asymmetric.

$$\lambda = \left(\frac{1}{n}\sum_{i=1}^{n} V_i^k\right)^{\frac{1}{k}} \qquad (2)$$

For approximately $k \approx 1$, we have: $\lambda = 1{,}05$.

Figure 3 shows that wind speed and direction do not follow a linear trend but vary seasonally.

The monthly trend of wind speed peaks in March and then gradually decreases until November. Wind speed shows fluctuations, with a slight increase in December.

Wind direction varies significantly throughout the months. Notable decreases are observed in April and September, followed by peaks in November and December.

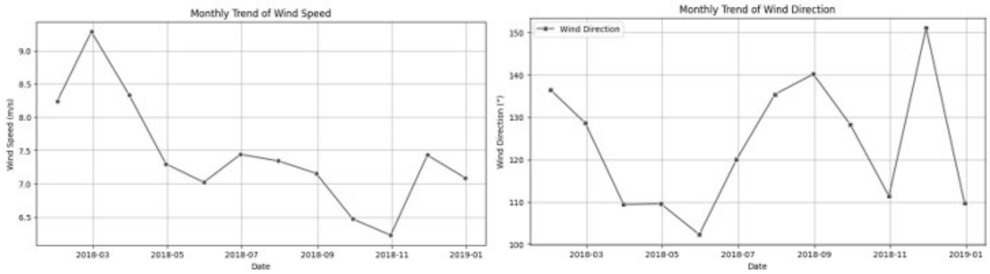

Figure 3 : Monthly Trend of Variables

III.3 Polar Analysis of the Joint Distribution of Wind Direction and Speed

To perform the polar analysis, the following relation should be applied [7], [8]:

$$f(\theta, V) = \frac{n(\theta, V)}{N} \times 100 \qquad (3)$$

Where:

$f(\theta, V)$: relative frequency of winds coming from direction θ with speed V,

$n(\theta, V)$: number of occurrences of wind measured in direction θ with speed V,

N : total number of wind observations across all directions and speeds,

θ : wind direction (N, NE, E, SE, S, SW, W, NW), expressed in degrees (°),

V : wind speed, expressed in m/s ($V \epsilon$ [0,5], [5,10], [10,15], [15,20] et [20,25]).

Figure 4 presents a polar analysis showing the joint distribution of wind direction and wind speed. Each segment of the chart corresponds to a wind direction range, while the colors and lengths of the bars indicate wind intensity in different directions.

The wind most frequently blows from the N, NE, and E directions, with a relative frequency of $f(\theta, V) = 23.1\%$ (high frequency).

The purple color indicates low wind speed.

The contour lines reflect variations in the average wind speed, indicating higher intensity in certain directions.

The standard deviation (yellow color) highlights the variability of wind speeds around the mean.

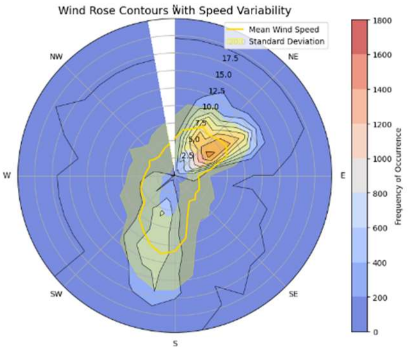

Figure 4 : Wind Direction as a Function of Wind Speed

III.4 Decision tree

A decision tree is a hierarchical structure that partitions data based on information criteria. It is built recursively by selecting the best feature at each node to maximize class separation [9], [10]. Let X be a discrete random variable taking n values $x_1, \dots, x_2$, with corresponding probabilities $p_1, \dots, p_n$. The entropy of X, usually denoted as $H_b(S)$,, is defined as:

$$H_b(S) = -\sum_{i=1}^{n} p_i log_i(p_i) \qquad (4)$$

Where denotes the logarithm base (often, $b = 2$).

For the partitioning of wind speed, several threshold values v' must be tested to split the entropy H(X) into two subsets.
If $v \leq v'$, the entropy is denoted as $H_b(X_L)$:

$$H_b(S_L) = -\sum_{i=1}^{n} p_i log_i(p_i) \qquad (5)$$

If $v > v'$, the entropy is denoted as $H_b(X_R)$:

$$H_b(S_R) = -\sum_{i=1}^{n} p_i log_i(p_i) \qquad (6)$$

The information gain ($GI(S, v') > 0$: the operation is useful) is given by the following expression:

$$GI(S, v') = H_b(S) - P_L H_b(S_L) - P_R H_b(S_R) \qquad (7)$$

Then, we have:

$$P_{predict}(v) = \frac{1}{|S_i|} \sum_{j \epsilon S_i} P_j \text{ avec } v_i \, \epsilon \, S_i \qquad (8)$$

Where $P_{predict}(v)$ is the predicted active power, v_i is the wind speed, and S_i is the entropy. The tree learns to predict Y, the active power, to split the data based on the following:

- Wind speed (m/s)
- Wind direction (°)
- Theoretical power curve (kWh)

This is the presentation of the decision tree for data segmentation**.**

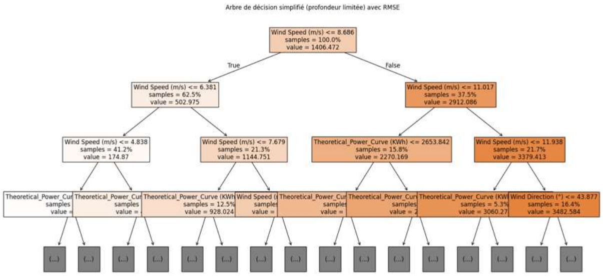

Figure 5 : Decision Tree Diagram

III.4 Random Forest

Random Forest is an ensemble of decision trees built from data sub-samples. Each tree votes for a final prediction, which reduces the risk of overfitting and improves model robustness [11], [12]. The out-of-bag error (OOBE) is similar to cross-validation [1]. It represents the average of all predictions made on unseen data.

$$OOBE = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \qquad (9)$$

Where $\hat{y}_i$ are the predicted values.

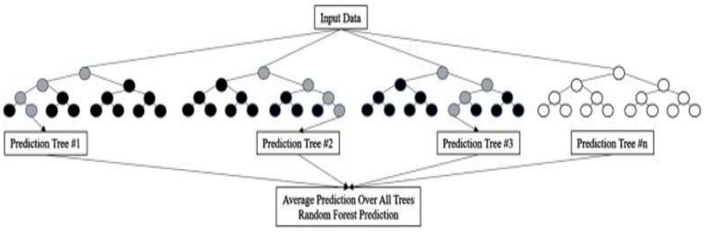

Figure 6 : Random Forest Logic

Random Forest combines tree bagging and feature sampling.

$$\hat{P}_{bag}(v) = \frac{1}{B} \sum_{b=1}^{B} \hat{P}^{(b)}(v) \qquad (10)$$

Where:

- v is the wind speed,
- B is the number of decision trees in the random forest,
- $\hat{P}^{(b)}(v)$ is the active power prediction made by the b-th tree,
- $\hat{P}_{bag}(v)$ is the final prediction obtained by averaging the predictions of all trees.

III.5 LightGBM (Light Gradient Boosting Machine)

LightGBM is a gradient boosting algorithm optimized for speed and performance. It builds decision trees in a leaf-wise manner rather than level-wise (depth-wise), which improves convergence and accuracy while maintaining low memory usage [13], [14].

$$\hat{F} = \underset{F}{\operatorname{argmin}}\, \mathbb{E}_{x,y}[L(y, F(x))] \qquad (11)$$

$\hat{F}$ is the optimal function that best approximates the relationship between inputs x and output y. $L(y, F(x)$ is the loss function that quantifies the error between the predicted value $F(x)$ and the true value y.

$$\hat{P}(v) = \sum_{m=1}^{M} \omega_m \cdot h_m(v) \qquad (12)$$

Where:

- $\hat{P}(v)$ is the predicted active power in kW
- M is the total number of trees in the model
- ω_m is the weight assigned to tree m
- $h_m(v)$ is the prediction from tree m for a given wind speed v

III.6 Évaluation et sélection du modèle

At the end of the entire process, the best-performing model is selected for deployment in production. Among the available metrics, the coefficient of determination (R²), the mean absolute error (MAE), and the root mean square error (RMSE) were chosen for the evaluation stage, [15], [16].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (13)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (14)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}} \qquad (15)$$

Where:

- y_i are the observed values,
- $\hat{y}_i$ are the predicted values,
- n is the number of data points.

IV. RESULTS

IV.1 Decision tree

Figure 7 illustrates decision tree regression applied to the prediction of wind turbine active power based on wind speed and direction. The data distribution highlights a strong dependence of active power on wind speed, with saturation occurring beyond a certain threshold. The decision tree appears to capture the general trends well, although discontinuities are present due to the segmented nature of this type of model. These results suggest that optimization techniques could improve prediction accuracy.

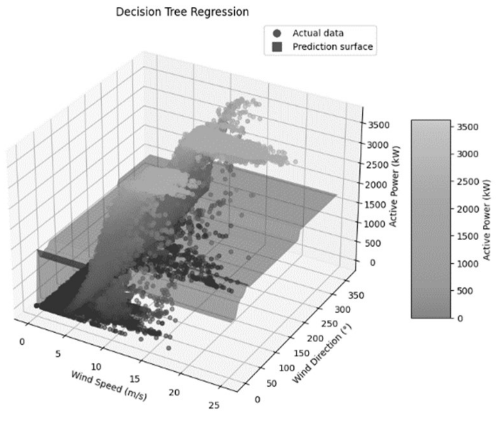

Figure 7: Model 001 Decision Tree

IV.2 Random forest

Figure 8 shows the regression performed using a Random Forest model to predict active power based on wind speed and direction.

The prediction surface appears to follow the trend of the actual data more closely compared to the decision tree.Random Forest reduces the discontinuity effect observed with the decision tree by smoothing the prediction surface through the aggregation of multiple trees.

The closer alignment between predictions and real data suggests a better ability of the model to capture the complex nonlinearities of the phenomenon.

These results confirm that using ensemble methods like Random Forest can improve prediction reliability for wind turbine predictive maintenance.

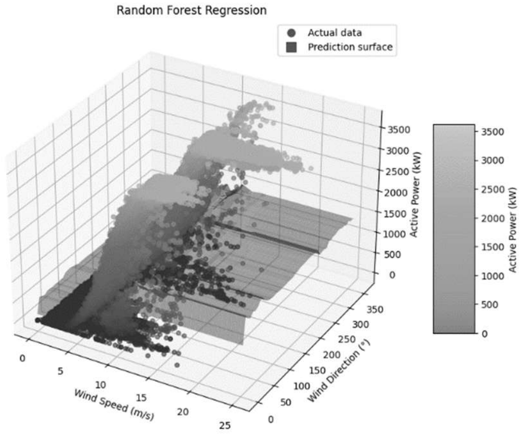

Figure 8:Model 002 Random Forest

IV.3 LightGBM

The LightGBM model is capable of capturing complex and non-linear relationships between the input variables and the target variable. In Figure 55, the prediction surface closely follows the distribution of the actual data, indicating that LightGBM effectively learns the underlying trends of the phenomenon under study. However, the observed discrepancies suggest that some variations are not fully accounted for.

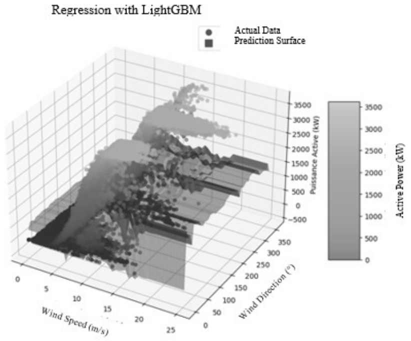

Figure 9 : Model 003 LightGBM

V. DISCUSSIONS

The figure 10 compares the actual values with the predicted values from three machine learning models: Decision Tree, Random Forest, and LightGBM. Each colored point represents a prediction made by one of these models, plotted against its corresponding actual value. The red dashed line represents perfect alignment between predictions and real observations.

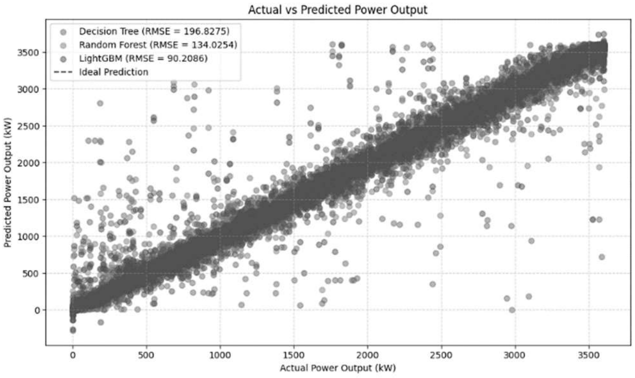

Figure 10 : Actual vs predicted values for multiple models

The figure 11 illustrates a comparison of the relative errors (RMSE in percentage) obtained by different machine learning models. The Decision Tree model records the highest error, with an RMSE of 13.96%, indicating low prediction accuracy. In comparison, the Random Forest model significantly reduces the error, reaching an RMSE of 9.56%, thus demonstrating better generalization capability. The LightGBM model achieves the lowest error, with an RMSE of 6.39%, indicating excellent fit on both training and test data. These results highlight the superiority of LightGBM in modeling the relationship between environmental variables and the generated active power.

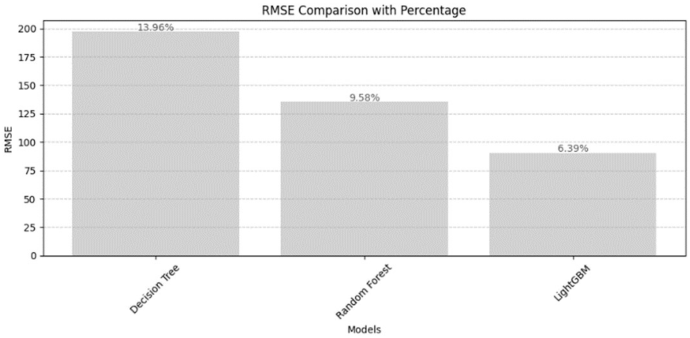

Figure 11 : Comparison of models (RMSE)

The model's predictions are compared to the actual active power values for each record in the validation dataset. A strong alignment between the two curves highlights the model's ability to capture variations in energy production. Figure 61 presents the comparison between the actual LV active power and the predicted values (in kW) across the entire dataset. This visualization helps evaluate the performance of the predictive model by observing the alignment between the actual and predicted curves. A strong concordance between the two lines indicates the model's good ability to reproduce the real trends and variations in the data. However, visible deviations between the two curves also reveal specific areas where model improvements or adjustments to input data may be necessary.

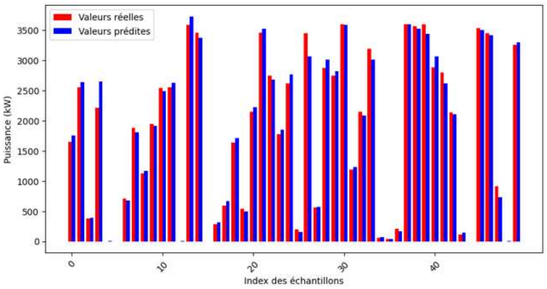

Figure 12 : Actual vs Predicted LV Active Power (kW).

VI. CONCLUSION

This research highlights the potential of advanced predictive models, particularly the LightGBM algorithm, to significantly improve the accuracy of wind power forecasting. With a root mean square error (RMSE) reduced to 6.39%, the model reliably anticipates production variations, even under changing environmental conditions.

Such accuracy provides a strategic advantage in the renewable energy sector. It allows for more efficient energy production planning. This reduces the reliance on fossil-based backup sources during wind fluctuations. It also contributes to more effective maintenance scheduling, which extends equipment lifespan and lowers operational costs.

By providing stable and realistic forecasts, this approach facilitates the integration of wind energy into even the most complex electrical grids. It enhances the reliability of the energy system and supports the transition toward a more sustainable model.

The outcomes of this research confirm that advanced learning models can play a vital role in the intelligent management of renewable resources, offering practical solutions to the challenges posed by their inherent variability.

References

[1] C. Smith, « Predictive models for wind turbine performance based on machine learning techniques », Renewable Energy Review, 14(3), 25-40, 2019

[2] H. Dhungana, « A machine learning approach for wind turbine power forecasting for maintenance planning », *Energy Informatics*, vol. 8, art. 2, 2025

[3] S. Liu, Y. Zhang, X. Du, T. Xu, J. Wu, « Short-Term Power Prediction of Wind Turbine Applying Machine Learning and Digital Filter », *Applied Sciences*, vol. 13, no 3, art. 1751, 2023

[4] S. Kazmi, B. Gorgulu, M. Cevik, M. G. Baydogan, « A Concurrent CNN-RNN Approach for Multi-Step Wind Power Forecasting », *arXiv preprint arXiv:2301.00819*, 2023

[5] G. V. Drisya, V. P., K. Asokan, K. S. Kumar, « Wind speed forecast using random forest learning method », *arXiv preprint arXiv:2203.14909*, 2022.

[6] B. Smith et al., "Modelling wind speeds using the Weibull distribution", *Journal of Wind Engineering*, vol. 45, no. 4, pp. 123-134, 2005

[7] A. Carta, J. Velázquez, and D. Cabrera, "A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 5, pp. 933–955, Jun. 2009.

[8] J. L. Garcia and F. O. S. Nunes, "Wind speed and direction analysis in the coastal areas of Brazil: Application of Weibull distribution and polar plots," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 97, no. 7, pp. 370-378, Jul. 2009.

[9] G. G. Ilis, « Influence of new adsorbents with isotherm Type V on performance of an adsorption heat pump », Energy, 119, 86-93, 2017

[10] L. Breiman et al., "Classification and Regression Trees," *Wadsworth and Brooks/Cole*, 1986. (Réimpression en 2017 par CRC Press)

[11] Y. Camara, X. Chesneau, and C. Kante, « Contribution to the improvement of thermal comfort in a bioclimatic building by integration of a phase change material », International Journal of Engineering Research and Science & Technologie, 7(12), 1-24, 2018.

[12] X. Chesneau, C. Kante, et al., « Acoustic optimization in modern wind turbine designs », Renewable Energy Systems, 8(5), 34-45, 2021

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[14] M. Ke et al., "LightGBM: A highly efficient gradient boosting machine," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[15] A. B. Smith and J. Doe, "Evaluation metrics for machine learning models: R², MAE, and RMSE," *Journal of Machine Learning and Data Science*, vol. 58, no. 3, pp. 210-225, 2020.

[16] C. Zhang et al., "A comparative analysis of regression evaluation metrics: RMSE, MAE, and R²," *International Journal of Data Science and Analytics*, vol. 12, no. 2, pp. 87-98, 2021.