International Journal of Advanced Engineering Research and Science (IJAERS)

Peer-Reviewed Journal

ISSN: 2349-6495(P) | 2456-1908(O)

Vol-12, Issue-6; Jun, 2025

***Journal Home Page Available:** https://ijaers.com/*

***Article DOI:** https://dx.doi.org/10.22161/ijaers.126.7*

Optimizing Wind Power Forecasting Using Machine Learning: A Comparative Study with Emphasis on LightGBM for Predictive Maintenance

Jean Marc Fabien Sitraka Randrianirina, Lovasoa Feno Fanantenana Rakotomalala, Bernard Andriamparany Andriamahitasoa, Zely Arivelo Randriamanantany, Liva Graffin Rakotoarimanana

[1]Laboratory of Thermodynamics, Thermal, and Combustion (LTTC), University of Antananarivo, Madagascar.
randrianirinajeanfabien@gmail.com
[2]Department of Physics, Antananarivo University, Madagascar

Received: 27 May 2025,

Receive in revised form: 18 Jun 2025,

Accepted: 23 Jun 2025,

Available online: 27 Jun 2025

***Keywords*— Machine Learning comparison, Predictive Maintenance, Wind Energy Forecasting, Wind Turbines**

***Abstract*— *The abstract should summarize the content of the paper. The variability of wind resources makes wind power forecasting challenging, which limits its integration into the electrical grid. To address this challenge, several machine learning models are compared to identify the most accurate solution for short-term forecasting. A one-year database, with a ten-minute time step, is used, including environmental variables such as wind speed and direction. An in-depth correlation analysis is performed, outliers are removed, and dimensionality reduction is applied using principal component analysis. Next, seven regression models are compared, including artificial neural networks, support vector machines, k-nearest neighbors, linear regression, decision trees, random forests, and LightGBM. Results show that LightGBM offers the best performance, with a normalized mean squared error of 4.36%, compared to 12.71% for linear regression. Thanks to its ability to model complex nonlinear relationships, LightGBM constitutes a reliable and robust solution for wind power forecasting. This approach significantly improves forecasting accuracy and facilitates the planning of predictive maintenance for wind turbines, which contributes to more efficient management of wind power systems.*

I. INTRODUCTION

Forecasting the active power of a wind turbine is a major challenge in the field of renewable energy. Since wind resources are inherently intermittent and unpredictable, accurate forecasting is essential for optimizing power grid management and increasing wind energy integration into the global energy mix. Over the decades, researchers have explored various methods to improve prediction accuracy, utilizing both traditional statistical models and modern machine learning approaches.

The limitations of early neural networks led researchers to explore other methods. By 2000, Support Vector Machines (SVMs) emerged as a robust solution due to their ability to handle high-dimensional data and non-linear patterns. For example, Z. Chen and X. Zhang [1] used SVMs to predict wind power, demonstrating significant improvement compared to Artificial Neural Networks (ANNs), especially in environments with limited or noisy data.

The 2010s marked the emergence of ensemble models, such as Random Forests (RF) and Gradient Boosting

]

Methods (GBM). These models combine multiple predictors to reduce errors and improve the robustness of forecasts. Y. Camara, X. Chesneau, and C. Kante [2] demonstrated the effectiveness of RF in solving complex problems, such as thermal comfort in bioclimatic buildings, highlighting their ability to capture intricate interactions between variables.

Moreover, energy forecasting reached new heights with the introduction of Deep Neural Networks (DNNs) and their variants, such as Long Short-Term Memory networks (LSTMs). These models can capture long-term temporal dependencies to model wind power generation in dynamic environments. X. Luo and J. Wang [3] found that LSTM networks often outperform traditional models in terms of prediction accuracy.

A study conducted by Dione (2020) [4] explored various machine learning methods for wind power forecasting, including Random Forests, Bagging, and Boosting. Dione's work showed that integrating spatio-temporal derivatives into Random Forest models helps reduce the Normalized Mean Absolute Error (NMAE); for example, the NMAE decreased from 7.75% to 7.70% for short-term forecasts. These results highlight the importance of using enriched data to improve prediction accuracy.

In 2021 [5], a study explored the modeling of wind turbine power curves using artificial intelligence tools and predefined inference criteria. A new method was proposed to combine fuzzy inference systems (Mamdani and Sugeno) with neural networks to improve the accuracy of power curve estimation.

Around 2023 [6], a study evaluated four artificial intelligence approaches (SVM, decision trees, ANFIS, and ANN) to predict short-term wind energy at a wind farm in Yalova, Turkey. The results showed that SVM outperformed the other models, with a Mean Absolute Error (MAE) of 71.21 and a coefficient of determination (R^2) of 0.95. Although ANN also showed high accuracy, decision trees and ANFIS performed less well.

Also in 2023, Liu et al. [7] combined LightGBM with digital filtering techniques to improve short-term forecasting accuracy. Using environmental sensor data collected over a six-month period, their model achieved an RMSE of 0.12, outperforming traditional methods. Kazmi et al. (2023) [8] 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, their model outperformed LightGBM and other tree-based regressors, although exact RMSE values were not specified.

In 2024 [9], a study examined the application of Artificial Intelligence (AI), including fuzzy logic, heuristic algorithms, deep learning, and reinforcement learning, in offshore wind systems. The study explored the control of wind systems, including maximum power point tracking and wake management. Finally, in the study by Dhungana (2025) [10], 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.

Therefore, selecting the optimal model for wind power forecasting involves a systematic evaluation of various algorithms, including ANN, KNN, SVM, Linear Regression, Decision Trees, Random Forest, and LightGBM, in order to identify the one that yields the lowest RMSE. This comprehensive approach not only improves prediction accuracy but also strengthens the reliability of forecasting systems in the context of variable and complex wind energy data.

II. METHODOLOGY

1. Wind Turbine Characteristics

The wind turbine under investigation possesses a rated capacity of 3618.73 kW. It is designed to operate optimally at a wind speed of approximately 7.10 m/s, achieving its maximum power output (rated power) at 13 m/s.

Physically, the turbine has an estimated weight of 120 metric tons, which is typical for a turbine of this capacity. Its rotor diameter measures approximately 120 meters, sweeping a surface area of 11,310 square meters. The turbine is equipped with three blades, each constructed from glass fiber-reinforced composite material.

The generator features a slotless, brushless permanent magnet alternator, providing an electrical output of 120/240 VAC with a frequency ranges of 59.3 to 60.5 Hz. The yaw control system is passive, and the rotational direction of the rotor is clockwise when viewed from upwind.

The dataset used for this analysis was collected from this specific wind turbine. It is provided in CSV format and includes key operational parameters such as actual wind speed, theoretical power output, and wind direction. These parameters were recorded continuously over a one-year period, with a ten-minute sampling interval, offering a high-resolution temporal profile of the turbine's performance.

2. Data Visualization

Data visualization is a crucial step in the analysis process. It helps to identify trends, understand relationships between variables, and detect potential anomalies. This step plays an essential role in interpreting results intuitively and is especially important for exploratory data analysis prior to any modeling.

The dataset used in this study was filtered to include one year of data from 20XX. Figure 1 illustrates the distribution of active power, wind direction, and wind speed (labeled as "LV Active Power (kW)"). This visualization is particularly useful for analyzing extreme behaviors in energy production, which can occur in wind turbines, and for identifying data points that may require further treatment before applying a predictive model.

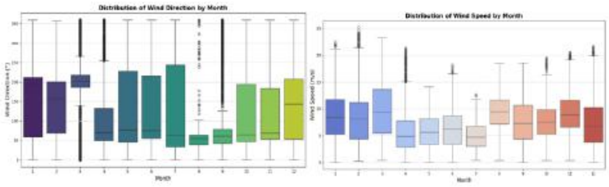

Fig. 1: Distribution of active power

(LV ActivePower (kW))

The visualization of active power throughout the year reveals clear fluctuations, with notable increases observed particularly in the middle and at the end of the year. These variations likely stem from external factors influencing energy production or consumption patterns.

Furthermore, the seasonality of active power demonstrates a high frequency, strongly suggesting a periodic behavior, likely on a daily or weekly cycle. This indicates that active power generation is significantly influenced by recurring variations, such as daily energy demand cycles or weekly weather patterns.

Finally, the analysis of residuals (random component) highlights irregular fluctuations, which represent unexpected variations or anomalies in the active power data. High variability within these residuals may indicate the occurrence of exceptional events or potential measurement errors.

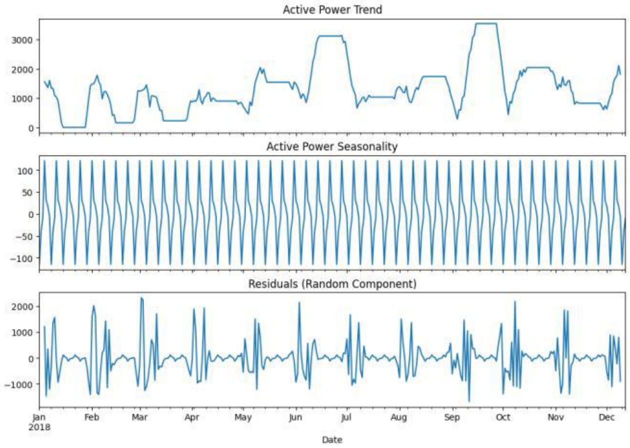

Fig. 2: Decomposition of the Active Power Time Series.

This decomposition approach allows for the separate analysis of long-term trends, regular cyclical patterns, and unexpected variations, providing a comprehensive understanding of the factors influencing wind turbine active power output

3. Correlation Analysis Between Numerical Variables

The function "*df[num_cols].corr()*" is used to compute the correlation coefficients between the various numerical variables in the dataset. This correlation matrix is then visualized as a "*heatmap*" using "*sns.heatmap()*", where the color intensity represents the strength and direction of the correlation between pairs of variables. This visualization provides valuable insights into linear relationships among features, which can guide feature selection, highlight multicollinearity, and support interpretation of modeling results.

To understand the relationships between the numerical variables in our dataset, we computed the correlation coefficients using *"df[num_cols].corr()"*. This correlation matrix was then visualized as a *"heatmap"* using *"sns.heatmap()"*. The color intensity in the *"heatmap"* represents the strength and direction of the correlation between pairs of variables. This visualization provides valuable insights into linear relationships among features, which can guide feature selection, highlight multicollinearity, and support the interpretation of modeling results.

$$P = \frac{1}{2} \rho\, A\, V^3 C_p \qquad (1)$$

Where: P is the standardized active power, ρ is the air density under standardized conditions, A is the standardized swept area of the blades, V^3 is the standardized wind speed and C_p is the power coefficient.

$$C_p = \frac{P_{mécanique}}{P_{total}} \qquad (2)$$

Where:$P_{\text{mechanical}}$ is the mechanical power extracted by the wind turbine and P_{Total} : the total wind power passing through the swept area.

Figure 3 presents a correlation matrix, visually representing the strength and direction of the linear relationships among the variables. Specifically, the correlation between active power and wind speed is 0.94. This indicates a strong and positive correlation, confirming that an increase in wind speed is directly associated with an increase in the power output of the wind turbine, up to a certain operational threshold.

That's a solid explanation of your correlation findings and their implications! You have clearly articulated why certain relationships (or lack thereof) exist.

The correlation between active power and wind direction is -0.07. This value, being close to zero, indicates an almost non-existent linear relationship. This suggests that wind direction doesn't significantly impact the turbine's power output, likely because the turbine's yaw mechanism automatically aligns it with the prevailing wind.

Similarly, the correlation between wind speed and wind direction is also -0.07. This reflects a weak linear relationship, suggesting these two variables are relatively independent within the analyzed dataset.

This correlation analysis is crucial for identifying the most relevant variables for modeling. Wind speed clearly emerges as a key predictor for active power output. In contrast, wind direction appears more suitable for secondary analyses or for understanding turbine operational dynamics rather than as a primary direct predictor of power generation.

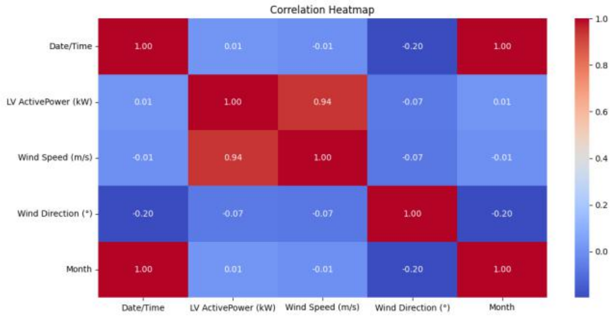

Fig. 3 : Correlation Analysis

4. Outlier Removal and Polar Analysis

The detected outliers are subsequently removed from the dataset using the *"remove_outlier()"* function. To conduct the polar analysis, based on established methodologies, the following relation is applied, [11, 12]:

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

Where : $f(\theta,V)$ is the relative frequency of winds from direction and at speed V, $n(\theta,V)$ is the number of occurrences of winds measured in direction θ and at speed V, N is the total number of wind observations across all directions and speeds, θ is the wind direction (N, NE, E, SE, S, SW, W, NW) expressed in (°), V is the wind speed ($V \epsilon$ [0,5], [5,10], [10,15], [15,20] et [20,25]) expressed in m/s.

f(θ,V) is the relative frequency of winds from direction θ, and at speed V. n(θ,V) is the number of occurrences of winds measured in direction θ, and at speed V. N is the total number of wind observations across all directions and speeds. θ is the wind direction (N, NE, E, SE, S, SW, W, NW) expressed in degrees (°). V is the wind speed ($V \epsilon$ [0,5], [5,10], [10,15], [15,20] and [20,25]) expressed in m/s.

Figure 4 presents a wind rose diagram that visualizes the joint distribution of wind direction and wind speed over the observation period. Each segment represents a sector of wind direction, while the concentric circles indicate the frequency of wind occurrence in percentage. The colors within each segment correspond to predefined ranges of wind speed, as detailed in the legend.

From this diagram, it is evident that the dominant wind directions are from the North (N), North-East (NE), and East (E). These directions exhibit the highest frequencies, with the NE direction showing the most significant contribution, exceeding 22% of all recorded wind events. Moreover, the wind speeds associated with these dominant directions span a broad range, from low to high velocities, with a notable proportion falling between 10.1 and 20.2 m/s.

This polar representation clearly highlights the anisotropy in wind behavior, confirming that the site experiences prevailing winds predominantly from the NE quadrant. Such insights are critical for various applications, including wind energy harvesting, pollution dispersion modeling, and structural orientation in civil and environmental engineering projects.

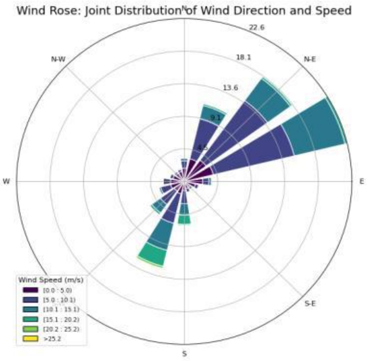

Fig. 1 : Wind rose diagram

5. Principal Component Analysis (PCA)

Principal Component Analysis (PCA), illustrated in Figure 5, is a statistical method aimed at reducing the dimensionality of data while maximizing the preservation of their variance. Its main steps consist of:

- Covariance Matrix: The PCA method begins by calculating the covariance matrix, which measures the relationships of variation between the data variables.

- Eigenvectors and Eigenvalues: The eigenvectors indicate the main directions of variation in the data, while the eigenvalues assess the importance of these directions. These elements are derived from the covariance matrix.
- Data Projection: The data is then projected onto the first principal components. This step helps reduce dimensionality while preserving as much of the information and variance of the initial data as possible.

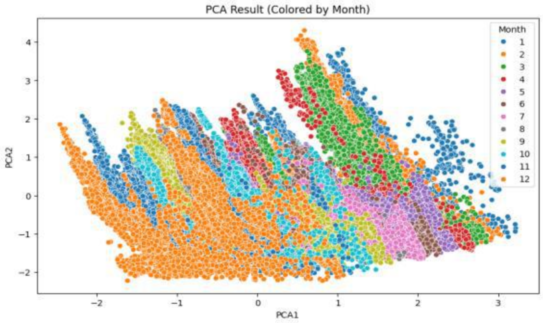

Fig. 5: PCA Results

Figure 6 displays the cumulative explained variance as a function of the number of principal components derived from our Principal Component Analysis (PCA). The graph illustrates that:

- The first principal component alone explains approximately 30% of the total variance in the dataset.
- The first two components combined explain around 60% of the total variance.
- The first three components together capture roughly 90% of the variance.
- Finally, the first four components account for 100% of the total variance, indicating that the complete original information is retained when considering these four components.

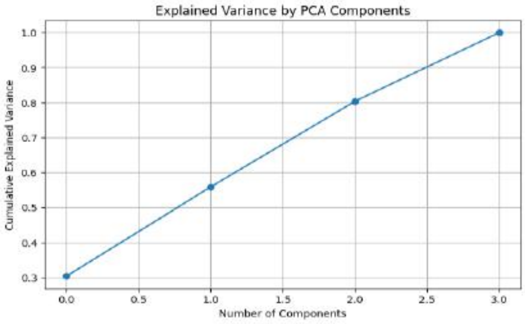

Fig. 2: Explained Variance by PCA components

6. Deep Neural Networks

Before processing the dataset and applying advanced modeling techniques, it's essential to follow necessary data preparation steps to ensure clear and actionable results. Our modeling approach involves the use of Deep Neural Networks (DNNs), which are powerful machine learning models capable of learning complex patterns and relationships within data.

Figure 7 illustrates the characteristics of the optimized neural network, which was developed using an Artificial Neural Network (ANN) approach.

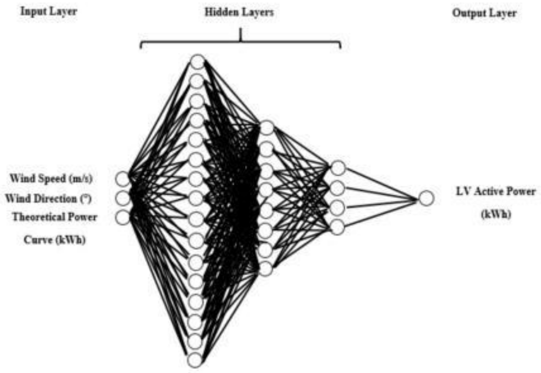

Fig. 3 : Characteristics of the Optimized Neural Network Using ANN

Random Forests improve a model's generalization ability by reducing variance compared to a single decision tree. This method is particularly well-suited for situations where the data is noisy or when the relationships between features are complex [13]. The overall prediction ŷ of a random forest can be expressed as:

$$\hat{y} = \frac{1}{N}\sum_{j=1}^{N} h_j(x) \qquad (4)$$

Decision trees and random forests are commonly used to analyze feature sets derived from acoustic or vibrational data, as well as in other domains like energy forecasting. These algorithms make it possible to detect patterns and anomalies in the data, thereby contributing to the early identification of potential machine failures or complex operational dynamics.

7. Comparison of SVM Regression

Support Vector Machine (SVM) Regression (SVR), while rooted in the principles of SVM classification, is specifically designed for regression tasks. Unlike classification SVMs that aim to find a hyperplane to separate distinct classes, SVR seeks to identify a hyperplane that best fits the data points within a specified margin of tolerance (ϵ-tube) in a multidimensional space [14]. The primary objective is to minimize the error by ensuring that as many data points as possible fall within this margin, thereby enhancing the model's ability to generalize to unseen data.

$$y = \omega^T . x + b \qquad (5)$$

Where: $\omega^T \epsilon\, X$, $b\ \epsilon\ R$ and $(\cdot,\cdot)$: denotes the inner product in X.

Finding a function y involves minimizing the value of ω. The prediction is calculated using the following expression:

$$y = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(x_i, x) + b \, (6)$$

Where: x_i is the standardized wind speed, x is the new standardized wind speed value, b is the model bias, y is the predicted standardized active power and $K(x_i, x)$ is a kernel function that measures the similarity between an observed wind speed x_i and a new wind speed x.

With,

$$K(x_i, x) = \ e^{-\gamma \| x_i - x \|^2} \qquad (7)$$

$$b = y_i - \omega \, . x_i - \epsilon \qquad (8)$$

The surface representing the prediction of the standardized SVM model (Figure 5, right) shows significant divergence from the actual data in certain value ranges, indicating a suboptimal fit. The visible gap between the experimental data points and the predictive surface highlights the importance of data normalization in enhancing the performance of the SVM model.

In this case, we have:

$$Y = \omega_1 X_1 + \omega_2 X_2 + \omega_3 X_3 + b \ (9)$$

The value of the kernel function is:

$$K = \begin{bmatrix} 1.00000000 & 5.02988819e-01 & 3.73359944e-06 & \dots & 7.26117308e-02 \\ 5.02988819e-01 & 1.00000000 & 4.78675171e-05 & \dots & 1.89641406e-01 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 7.26117308e-02 & 1.89641406e-01 & 2.43227190e-08 & \dots & 1.00000000 \end{bmatrix}$$

The value of the standardized matrix is:

$$X = \begin{bmatrix} -1.12525318 & 0.60054166 & -1.09166059 \\ -0.61467742 & -0.00729909 & -0.85285238 \\ \vdots & \vdots & \vdots \\ -1.41008267 & -0.9936669 & -1.09166059 \end{bmatrix}$$

And

$$Y = \begin{bmatrix} -0.70679964 \\ -0.65062865 \\ \vdots \\ \vdots \end{bmatrix}$$

We obtain,

$$\omega = \begin{bmatrix} 0.10677864 \\ 0.01298735 \\ 0.90329082 \end{bmatrix}$$

And the value of the standardized bias is:

$$b = 0.043266308382584225$$

Hence,

$$Y = 0.106 X_1 + 0.012 X_2 + 0.90 X_3 + 0.0432 \qquad (10)$$

8. Comparison of KNN Regression

K-Nearest Neighbors (KNN) Regression is a non-parametric method used for prediction. Its performance heavily relies on the optimal choice of k (the number of nearest neighbors) and the distance metric employed [15, 16].

$$d(x, x_j) = \sqrt{\sum_{i=1}^{n} (x_i - x_{j,i})^2} \qquad (11)$$

where: x_i is the standardized wind speed, $x_{j,i}$ is a point among the -nearest neighbors, n is the number of dimensions in the feature space, x_i and $x_{j,i}$ is the i-th components of x_i and $x_{j,i}$.

The Euclidean distance is essential for defining the proximity between points, with neighbors identified as those for which this distance is minimal. To make predictions, the derivative of the Mean Squared Error (MSE) with respect to the predicted value is set to zero in order to minimize the error.

$$MSE = \frac{1}{k} \sum_{i \epsilon N_k(x)} (y_i - \hat{y}_i(x))^2 \, (12)$$

$$\frac{\partial MSE}{\partial \hat{y}_i(x)} = 0 \qquad (13)$$

Therefore, we obtain:

$$\hat{y}_i(x) = \frac{1}{k} \sum_{i \epsilon N_k(x)} y_i \qquad (14)$$

where: $\hat{y}_i(x)$ is The standardized predicted active power, y_i is The standardized wind speed associated with the neighbor x_i and $N_k(x)$: The index of k the smallest values of $d(x, x_j)$.

9. Linear Regression

Regression differs from correlation analysis, in which all variables are treated symmetrically. Indeed, the existence of a relationship between X and Y does not necessarily imply a causal relationship between the two variables [17].

The variable Y represents the actual random variable to be explained (dependent or response variable). The model assumes that, on average, $E(X)$ can be expressed as an affine function of X. This assumption implicitly suggests a notion of causality.

$$E(X) = f(X) = \beta_0 + \beta_1 X \qquad (15)$$

Where $Y = \beta_0 + \beta_1 X + \varepsilon$ and X is a deterministic variable. When X is considered random, the model is expressed conditionally on the observed values of X:

$$E(Y|X = x) = \beta_0 + \beta_1 x \qquad (16)$$

The assumptions associated with this model are as follows:

The distribution of the error term ϵ is independent of the independent variable X.

The error has a mean of zero and constant variance:

$$\forall i = 1, \ldots, n \; E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$$

β_0 and β_1 from equation 15) are constant, meaning the model does not experience structural breaks.

An additional assumption for inference: $\varepsilon \sim N(0, \sigma^2)$.

The linear regression coefficients are reported in Tables 1 and 2.

The value of the standardized matrix X is:

$$Y = 0.1430X_1 + 0.0293X_2 + 0.81836X_3$$

The value of the non-standardized matrix X is:

$$y = 1311.36 + 44.33x_1 + 0.41x_2 + 0.79x_3 \qquad (17)$$

Table 1: Standardized Linear Regression Coefficients.

	Variable	Standardized Coefficient	Standardized RMSE Error
0	*Wind Speed* (m/s)	1.430113e-01	0.321166
1	*Wind Direction* (°)	2.933001e-02	0.318290
2	*Theoretical Power Curve* (kWh)	8.183626e-01	0.361815
3	*Intercept*	3.878953e-15	---

Table 2 : Unstandardized Linear Regression Coefficients.

	Variable	Unstandardized Coefficient	Unstandardized RMSE Error
0	*Wind Speed* (m/s)	44.331674	42.523935
1	*Wind Direction* (°)	0.413111	140.765042
2	*Theoretical Power Curve* (kWh)	0.785241	655.471999
3	*Intercept*	1311.361866	403.429879

10. Decision Trees

10.1 Construction of the Tree

Decision trees are a versatile and effective technique for analyzing data and making predictions. While commonly used for predicting a qualitative variable (classification) based on predictors of any type (qualitative or quantitative), they are equally adept at regression tasks (predicting a quantitative variable). Their inherent flexibility, allowing them to handle diverse predictor types, represents a significant advantage over other classification or regression tools that may be limited in this regard [18].

Let X be a discrete random variable, taking n distinct values x_1,x_2,…,x_n with respective probabilities p_1, p_2,…, p_n. The entropy of X, generally denoted as $H(X)$, quantifies the impurity or uncertainty in the data and is given by the following expression:

$$H_b(X) = -\sum_{i=1}^{n} p_i log_i(p_i) \qquad (18)$$

where b is the base of the logarithm (often $b = 2$).

For the partition of wind speed, several thresholds v' of the wind speed must be tested to separate the entropy $H(X)$ into two subsets. If $v \leq v'$, we denote the entropy as $H_b(X_L)$:

$$H_b(S_L) = -\sum_{i=1}^{n} p_i log_i(p_i) \qquad (19)$$

If $v > v''$, we denote the entropy as $H_b(X_R)$:

$$H_b(S_R) = -\sum_{i=1}^{n} p_i log_i(p_i) \qquad (20)$$

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

$$GI(S, v') = H_b(S) - P_L H_b(S_L) - P_R H_b(S_R) \qquad (21)$$

Thus, we have:

$$P_{predict}(v) = \frac{1}{|S_i|} \sum_{j \epsilon S_i} P_j \qquad (22)$$

Where : $P_{predict}(v)$ is the predicted active power, v_i : is the wind speed and S_i represents the entropy.

Entropy is interpreted as a measure of disorder. It is zero when all the values of X are perfectly known. Moreover, it reaches its maximum (positive) when X follows a uniform distribution. It achieves its maximum value (positive) when X is uniformly distributed, which means that $p_i = \frac{1}{n}$ for all i.

10.2 Visualization of the Tree

The tree learns to predict active power (Y) by recursively splitting the dataset based on key features, including:

- Wind speed;
- Wind direction;
- Theoretical power curve.

This visualization will demonstrate how the tree segments the data space to make its predictions.

This section presents the decision tree for data segmentation.

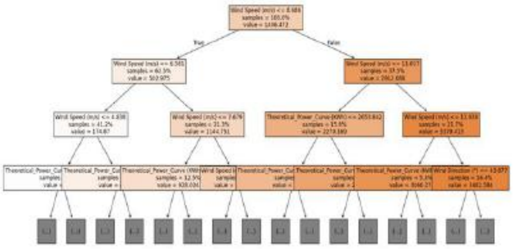

Fig. 4: Decision Tree Flowchart

Figure 8 illustrates the samples and the division created by the threshold function for specific wind speeds within the decision tree. The intermediate nodes of the tree further refine these predictions by incorporating additional factors such as the theoretical power curve and wind direction. This hierarchical splitting process allows the tree to segment the data space effectively, leading to more precise predictions of active power.

11. Random Forest

A Random Forest (RF) is a classifier composed of an ensemble of classifiers structured as trees, $\{h(x, \theta_k), k = 1, \ldots, n\}$, where the $\{\theta_k\}$ are independent and identically distributed random vectors. Each tree assigns a unit vote to the most popular class for a given input x, [19] [20]. The ensemble of classifiers $h_1(x)$, $h_2(x)$,..., $h_k(x)$ and the learning set S, randomly sampled from the distribution of the random vector (X, Y), define the margin function as given in the following relation:

$$mg(X, Y) = av_k I(h_k(X) = Y) - max_{j \neq Y} av_k I(h_k(X) = j) \quad (23)$$

Where, $I(.)$ is the indicator function.

The Random Forest (RF) algorithm is an ensemble learning method composed of multiple decision trees that operate in parallel. This ensemble approach significantly enhances the model's robustness and effectively reduces overfitting compared to using a single decision tree. Each tree in the forest is trained on a random subset:

- Random Sampling (Bootstrapping): Trees are constructed using data samples drawn with replacement from the original dataset;
- Random Feature Selection: At each split within a tree, a random subset of variables is used to determine the best threshold.

12. LightGBM

LightGBM is a highly efficient gradient boosting algorithm specifically optimized for both speed and performance. A key differentiator of LightGBM is its approach to tree construction: it builds decision trees in a leaf-wise (best-first) manner rather than a level-wise (depth-wise) fashion. This leaf-wise growth strategy typically leads to faster convergence and higher accuracy, while simultaneously maintaining lower memory consumption. [21].

$$\hat{F} = \underset{F}{argmin}\, \mathbb{E}_{x,y}[L(y, F(x)] \quad (24)$$

Where: $\hat{F}$ is the optimal function that best approximates the relationship between the input x and the output y. $L(y, F(x))$ is the loss function that quantifies the gap between the prediction $F(x)$ and the actual value of y.

$$\hat{P}(v) = \sum_{m=1}^{M} \omega_m \cdot h_m(v) \quad (25)$$

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 and $h_m(v)$ is the prediction of tree mmm for a wind speed v.

13. Normalized Root Mean Squared Error (nRMSE)

The nRMSE is obtained by normalizing the RMSE using the range of the observed values. This allows for performance comparisons across different datasets [22].

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \quad (26)$$

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

III. RESULTS AND DISCUSSIONS

1. Impact of Data Standardization on ANN Model Accuracy

Figure 9, which is based on your dataset, illustrates that active power is distributed between 0 and 3500 kW. The intermediate distribution appears relatively sparse, suggesting irregular system operation. This observation highlights potential opportunities for optimization, particularly in balancing the load and improving overall energy efficiency.

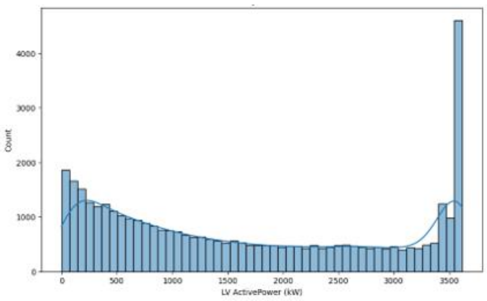

Fig. 5 : Distribution of Active Power (kW)

Figure 10 visually demonstrates the significant impact of data standardization on the stability and convergence of neural network models.

The right-hand plot in Figure 10, representing the standardized ANN, clearly displays a smoother regression surface. This smooth surface signifies that the model is better able to fit the actual data, leading to more stable and reliable predictions.

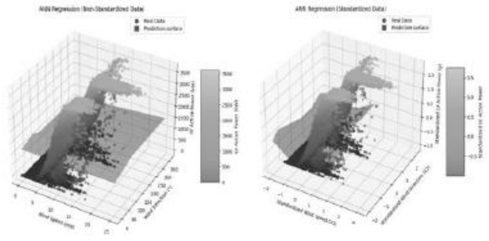

Fig. 6 : Model 001 ANN (Standardized and Unstandardized)

2. Support Vector Machine Performance and Global Trend Modeling

The real data, as previously illustrated in Figure 1 (left), exhibit a clear overall trend. However, this data also shows significant dispersion, revealing a complex underlying relationship that the SVM model attempts to replicate.

The prediction surface generated by the SVM model appears noticeably smoother. This smoothness indicates that the SVM model tends to capture global trends within the data rather than precisely accounting for every local variation.

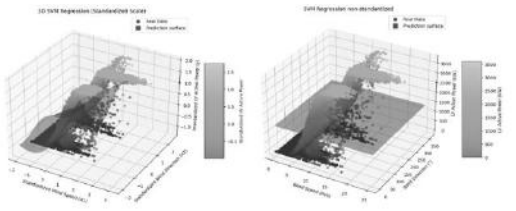

Fig. 7 : Model 002 SVM (Standardized and Unstandardized)

3. Impact of Standardization on KNN Model Accuracy

Figure 12 demonstrates that standardization significantly alters the scale of the independent variables, such as wind speed and wind energy. Consequently, standardization improves the data distribution and facilitates better model generalization by preventing biases related to the initial measurement units. Without standardization, predictions appear more scattered and less accurate, primarily due to the magnitude discrepancies between the variables.

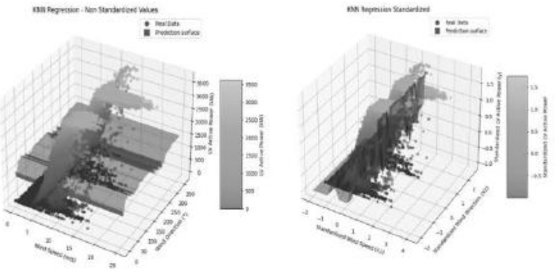

Fig. 8 : Models 003 KNN (Standardized and Unstandardized).

4. Effect of Standardization on Linear Regression Coefficients and Prediction Surface

As illustrated in Figure 13, the regression equation significantly changes between standardized and non-standardized models. This demonstrates that standardization primarily alters the scale of the coefficients while crucially maintaining the underlying structure of the relationships between the variables. Consequently, standardization facilitates an easier comparison of the relative importance of the coefficients, thereby improving the interpretation of the contributions of the independent variables to the model.

Furthermore, the prediction surface of the non-standardized model is unduly influenced by the raw scale of the variables. This disparity in scales can bias the model's learning process, particularly if the units or magnitudes of the independent variables are vastly different.

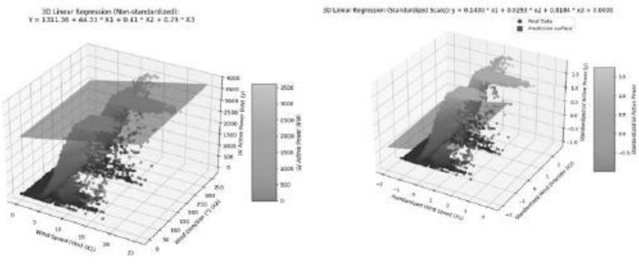

Fig. 9 : Models 004 Linear regression (Standardized and Unstandardized)

5. Decision Tree Regression for Wind Turbine Power Prediction

Figure 14 illustrates the application of decision tree regression for predicting the active power of a wind turbine based on wind speed and direction. The data distribution clearly shows a strong dependency of active power on wind speed, with power output saturating beyond a certain threshold. The decision tree effectively captures these general trends; however, discontinuities emerge due to the segmented nature of this model type. These results suggest that applying optimization techniques could further improve prediction accuracy.

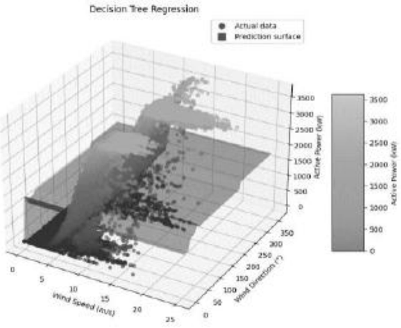

Fig. 10 : Model 005 Decision Tree

6. Random Forest Regression for Enhanced Predictive Accuracy

Figure 15 displays the regression performed using a Random Forest (RF) model to predict active power based on wind speed and direction. Compared to the single decision tree, the RF model's prediction surface significantly better aligns with the actual data's trend.

The Random Forest effectively reduces the discontinuity inherent in individual decision trees by smoothing the prediction surface through the aggregation of multiple trees. This closer alignment between predictions and actual data suggests the model's superior ability to capture the complex nonlinearities of the wind power phenomenon. These results confirm that ensemble methods like Random Forest notably improve the reliability of predictions crucial for wind turbine predictive maintenance.

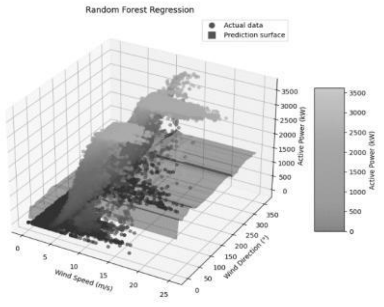

Fig. 11: Model 006 Random Forest

7. LightGBM Regression for Capturing Nonlinear Relationships

The LightGBM model excels at capturing complex and nonlinear relationships between input variables and the target variable. As Figure 16 illustrates, its prediction surface closely follows the distribution of the actual data, demonstrating that LightGBM effectively learns the intricate trends of the studied phenomenon. However, some observed deviations indicate that there are still variations not fully accounted for by the model.

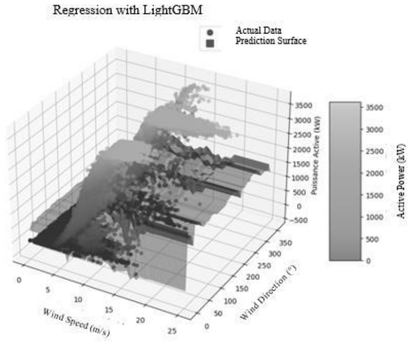

Fig. 12 : Model 007 – LightGBM

8. Evaluation of Regression Models: ANN, SVM, KNN, Linear Regression, Decision Tree, RF, and LightGBM

We tested several regression models to capture the relationships between environmental variables and the generated active power. Each model was chosen for its ability to handle complex data and its recognized performance in similar contexts.

Figure 17 compares the actual values against the predicted values obtained from these various machine learning models. Each colored dot represents a prediction made by a specific model, plotted against its corresponding actual value. The red dashed line indicates the ideal trend, where predictions perfectly match the real values. A significant deviation from this line suggests a higher error for that corresponding model.

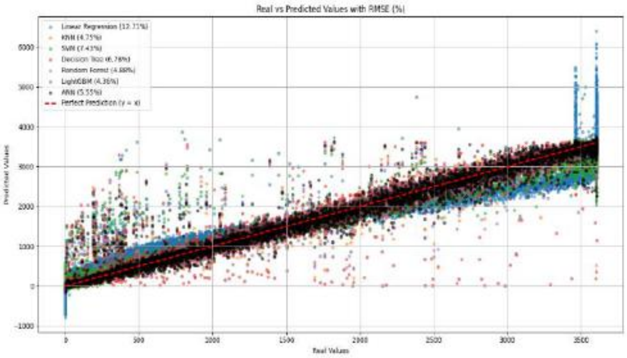

Fig. 13 : Comparison of RMSE for Each Model

Figure 18 presents a comparative analysis of the Root Mean Square Error (RMSE) values across the different machine learning models evaluated. As seen in the figure, the Linear Regression model exhibits the highest error, with an RMSE of 12.71%. This indicates its comparatively lower performance in terms of prediction accuracy for this dataset. In stark contrast, the LightGBM model achieved significantly lower errors, with an RMSE of just 4.36%. This result strongly suggests that LightGBM provides a much better fit to the data and superior predictive accuracy.

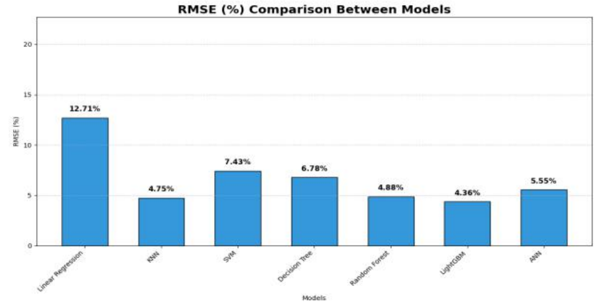

Fig. 14 : RMSE (%) Comparison between models

The LightGBM model emerged as the most effective among all approaches tested, demonstrating a significantly lower RMSE compared to other methods (Figure 18). Its superior ability to capture complex relationships between environmental variables, such as wind speed, wind direction, and temporal components, directly translated into highly accurate predictions. In line with previous studies, Random Forests also proved effective in both cases but delivered slightly lower performance than LightGBM in this particular analysis. These findings, grounded in experimental data and standard evaluation metrics, confirm LightGBM's superiority in complex scenarios and its more reliable application in wind power forecasting contexts.

IV. CONCLUSION

This study successfully demonstrated the effectiveness of machine learning models in improving the accuracy of wind power forecasting. By comparing various regression algorithms, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Linear Regression, Decision Trees, Random Forest, and LightGBM, our analysis revealed that LightGBM achieved the best performance with the lowest prediction error.

The strategic application of data standardization, outlier filtering, and dimensionality reduction proved essential in enhancing both model stability and interpretability. Among all tested approaches, LightGBM's superior ability to model nonlinear relationships allowed it to capture the complex dynamics of wind power generation most effectively.

These results underscore the significant potential of advanced machine learning algorithms. They can not only optimize wind energy integration into power systems but also robustly support predictive maintenance strategies, ultimately improving the operational reliability and efficiency of wind turbines.

REFERENCES

[1] A. Kaur, S. Gupta (2016), « Artificial neural networks for wind power prediction : A review and comparative analysis with traditional models », *Renewable and Sustainable Energy Reviews*, 60, 487-496.

[2] Y. Camara, X. Chesneau, C. Kante (2014), « Random Forests for complex problems : Application to thermal comfort prediction in bioclimatic buildings », *Energy and Buildings*, 82, 160-168.

[3] Luo, X. et Wang, J. , (2023). “Deep Neural Networks and Long Short-Term Memory Networks for Wind Energy Forecasting.” Journal of Renewable Energy.

[4] Dione, M. (2020). *Prévision court terme de la production éolienne par Machine Learning*. Thèse de doctorat, Institut Polytechnique de Paris, École nationale de la statistique et de l'administration économique.

[5] J. C. de Albuquerque, R. R. B. de Aquino, O. N. Neto, M. M. S. Lira, A. A. Ferreira, M. A. de Carvalho Jr.(2021). « *Power Curve Modelling for Wind Turbine Using Artificial Intelligence Tools and Pre-established Inference Criteria* », Journal of Modern Power Systems and Clean Energy, 9(3), 526-533.

[6] Bouabdallaoui D., Haidi T., Elmariami F., Derri M., Mellouli E. M. (2023) « *Application of four machine-learning methods to predict short-horizon wind energy* », Global Energy Interconnection, 6(6), 726-737.

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

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

[9] D. Song, G. Shen, C. Huang, Q. Huang, J. Yang, M. Dong, Y. H. Joo, N. Duić. (2024). « *Review on the Application of Artificial Intelligence Methods in the Control and Design of Offshore Wind Power Systems* », Journal of Marine Science and Engineering, vol. 12, n° 424, pp. 1-20.

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

[11] A. Carta, J. Velázquez, and D. Cabrera. (Jun. 2009) "*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.

[12] J. L. Garcia and F. O. S. Nunes. (Jul. 2009.) "*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.

[13] Haykin S. (2009), Neural Networks and Learning Machines, 3rd ed., Pearson.

[14] Kharroubi, J. (2002) *Étude de techniques de classement « Machines à Vecteurs Supports » pour la vérification automatique du locuteur.* Thèse de doctorat, École Nationale Supérieure des Télécommunications, France.

[15] Hastie, Trevor. (2001). *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations.* Tibshirani, Robert., Friedman, J. H. (Jerome H.).

[16] F. Wang, H. Zhang, J. Li, X. Liu, Y. Zhao. (2017) « *K-Nearest Neighbor (KNN) Based Classification Algorithm for Data Mining Applications* », International Journal of Modern Education and Computer Science, 9(5), 23-29.

[17] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). “*Introduction to Linear Regression Analysis* (5th ed.)”. Wiley.

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

[19] Y. Camara, X. Chesneau, and C. Kante. (2018) « *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.

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

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

[22] G. Zhang, « Neural networks for time-series forecasting : Current status and future directions », International Journal of Forecasting.