A Novel Hybrid Machine Learning Framework for Wind Speed Prediction

Mohamed Yassine Rhafes[1*], *Omar* Moussaoui[1], *Maria Simona* Raboaca[2], *and Traian Candin* Mihaltan[3]

[1]MATSI Laboratory, ESTO, Mohammed First University, Oujda, Morocco
[2]ICSI Energy Department, National Research and Development Institute for Cryogenics and Isotopic Technologies, Romania
[3]Faculty of Building Services, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania

Abstract. The growing urgency of environmental challenges and the depletion of fossil fuels have accelerated the search for sustainable and renewable energy sources. Wind energy, for example, is an important source of green electricity. However, using wind power is challenging due to the variability and unpredictability of wind patterns. Consequently, the ability to predict wind power in advance is crucial. The integration of artificial intelligence within the renewable energy sector could provide a viable solution to this challenge. In this study, we investigate the potential of machine learning to improve wind power forecasting by conducting a comparison of three regression models: K-Nearest Neighbor regression, Random Forest regression, and Support Vector regression. These models are combined with a feature selection technique to forecast wind power. Additionally, we propose a novel hybrid approach that combines these machine learning models with Multiple Linear Regression to address the complexities of wind energy forecasting. The performance of the models is evaluated using the R^2 score, Mean Absolute Error, and Root Mean Squared Error. The dataset for this study was generated from a numerical simulation conducted at a location with a latitude of 22.55° N and a longitude of -14.33° E. The findings demonstrate that the proposed hybrid model outperforms the individual machine learning models in terms of prediction accuracy. This study provides a solid foundation for future research and development in wind energy forecasting.

Keywords: Artificial Intelligence, Machine Learning, Hybrid Framework, Exhaustive Feature Selection, Wind Speed Prediction, Wind Energy

1 Introduction

In recent years, the renewable energy sector has experienced significant growth in research and development activities, driven by the growing need for sustainable energy products and

* Corresponding author: mohamedyassine.rhafes@ump.ac.ma

solutions [1]. Using renewable energy sources for electricity generation presents a viable approach to meet the increasing electricity demand and climate change challenges [2] [3].

Electricity generation from wind energy is challenging due to its dependence on unpredictable weather patterns [4]. Artificial intelligence, especially machine learning, has become a useful tool in the renewable energy field to address these challenges. Machine learning models excel in analyzing large volumes of data, including meteorological and geographical information, which enhances their effectiveness in managing the uncertainties of wind energy. Consequently, numerous studies have focused on improving wind energy forecasting through machine learning. Study [5] examines the use of ensemble models such as Random Forest Regression, Gradient Boosted Regression, and Extreme Gradient Boosting. Study [6] compares LASSO regression, K-Nearest Neighbor regression, XGBoost regression, and Support Vector Regression. Additionally, study [7] compares Gradient Boosting Machine, K-Nearest Neighbor Regression, Decision Tree, and Extra Tree Regression. Similarly, study [8] compares Random Forest Regression, Neural Networks, and Extreme Gradient Boosting, while study [9] focuses only on Gradient Boosting Machine. All these studies employ standard machine learning techniques without introducing new methods for predicting wind power.

This research introduces a novel approach by not only comparing standard machine learning models for wind power prediction but also integrating them into a new hybrid machine learning framework. This hybrid method leverages the strengths of each individual model to enhance the accuracy and reliability of wind power forecasts. The primary objective is to offer a robust solution to the challenges in wind energy forecasting.

The paper is structured as follows: Section 2 describes the materials and methods employed in the study. Section 3 presents a detailed discussion of the results. Finally, the paper concludes with a summary of the main findings.

2 Materials and Methods

2.1 Dataset

In this research, the dataset for training, validating, and testing the models was generated from numerical simulations using the NASA Data Access Viewer [10]. We selected specific geographic coordinates to generate our dataset. The chosen coordinates, a latitude of 23.70° N and a longitude of -15.94° E correspond to Dakhla City in Morocco, a region known for its high wind speeds [11]. The dataset includes Year, Month, Day, Hour, Temperature at 2 meters (T2M), Relative Humidity at 2 meters (RH2M), Precipitation (PR), Wind Direction at 10 meters (WD10M), and Wind Speed at 50 meters (WS50M). It contains 9,504 records, spanning from 01/04/2023 to 30/04/2024. The training and validation sets span from 01/04/2023 to 28/02/2024, and the test set spans from 01/03/2024 to 30/04/2024.

2.2 Machine Learning Algorithms

We concentrated on supervised learning, specifically targeting regression problems. This approach is crucial in the field of wind energy, where machine learning models are employed to predict wind speeds. Table 1 presents the details of the algorithms used in this study.

Table 1. Machine learning algorithms used in this study

Ref	Algorithm	Short Description
[12]	Ensemble Method	Is a technique that combines multiple machine learning algorithms. The objective is to obtain better predictive performance. There are several types of ensemble methods, the most commonly used are bagging, boosting, and stacking. In our analysis, we used both bagging and stacking techniques.
[13]	Stacking	Is an ensemble technique that aggregates predictions from various machine learning models to create a new dataset. This dataset serves as the training and testing ground for another algorithm, known as the meta-model. The meta-model then makes the final prediction based on this aggregated data.
[14]	Bagging	Is a type of ensemble method that creates multiple subsets of the original dataset and trains multiple instances of the same machine learning algorithm on these subsets. The final prediction is the average of the predictions from multiple instances of the algorithm.
[15]	K-Nearest Neighbors (KNN)	is a machine learning algorithm that predicts the outcome for a new data point by considering the values of its 'k' closest neighbors. The prediction is based on averaging or voting among these neighbors. One of the key challenges with this algorithm is selecting the optimal value for 'k', which directly impacts its accuracy and performance.
[16]	Support Vector Regression (SVR)	This algorithm, derived from Support Vector Machines (SVM) [16], is designed for regression tasks. Its goal is to find the optimal hyperplane that fits the data within a specified tolerance while maximizing the margin around the hyperplane. The approach seeks to capture as many data points as possible within this margin to improve the algorithm's prediction accuracy.
[17]	Random Forest (RF)	Is a machine learning algorithm that works by building multiple decision trees and outputting the average of their predictions.
[18]	Linear Regression (LR)	Is a statistical technique used to analyze the relationship between one dependent variable and independent variables. It works by fitting a linear equation to the data, allowing for the prediction of the dependent variable based on the input values of the independent variables.

2.3 Proposed Framework

Our framework combines stacking and bagging techniques. For the stacking approach, we use KNN, SVR, and a bagged RF as base models. these three different types of algorithms trained on the same training set. Each model might capture different patterns in the data due

to their unique methodologies. The predictions from the KNN, SVR, and RF models are combined to form a new training set. For the meta-model, which integrates the predictions from these base models to make the final prediction, we use multiple linear regression. Figure 1 presents our proposed framework.

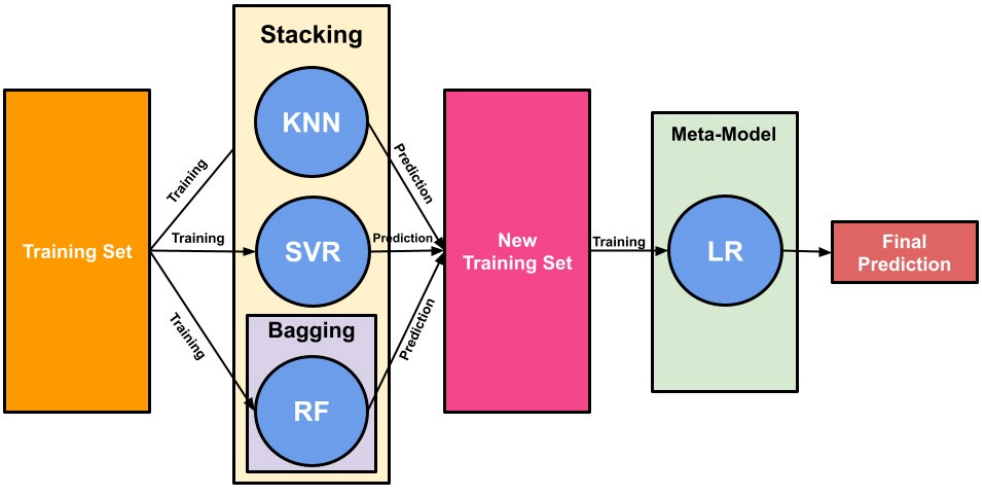

Fig. 1. Proposed hybrid machine learning framework for wind speed prediction

2.4 Measurements of forecasting performance

Forecasting performance is evaluated using a range of statistical methods to measure a model's accuracy. The metrics used to assess the models in this study are outlined in Table 2.

Table 2. Performance metrics used in this study

Metric	Formula	Components
Mean Absolute Error (MAE)		is the number of observations is the actual value is the predicted value is the mean of the actual values
Root Mean Squared Error (RMSE)		
R2 Score (Coefficient of Determination)		

2.5 Feature Selection Technique

Feature selection techniques are methods used to identify and select a subset of relevant features that enhance the performance of the algorithm. In this research, we employed exhaustive feature selection [19], a technique that evaluates all possible combinations of features to determine the combination that yields the highest performance.

2.6 Pipeline

To ensure the integrity of our research, we followed the methodology outlined in Figure 2, which includes the following steps:

- Data Collection: This first step involves collecting the necessary data.
- Exploratory Data Analysis: Once the data collected, a preliminary analysis to determine their characteristics and structure of the data.
- Data Preprocessing: Before modeling, the data is cleaned and transformed. This process involves addressing missing data and normalizing values.
- Training and validation sets: The dataset is split into training and validation sets, which are used for model learning and parameter tuning.
- Predictions: Finally, the model makes predictions based on the testing set, outputting the results for evaluation.

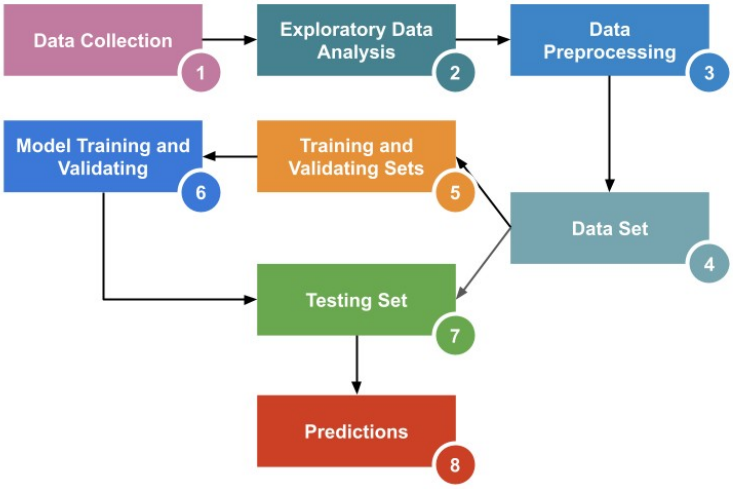

Fig. 2. Methodological approach adopted in this research

3 Results and Discussion

We used Scikit-learn [20], a free and open-source machine learninng framework in our experiments.

After generating the dataset using the NASA Data Access Viewer [10], we performed an analysis to examine the linear relationships between the features and the target. As illustrated in Figure 3.

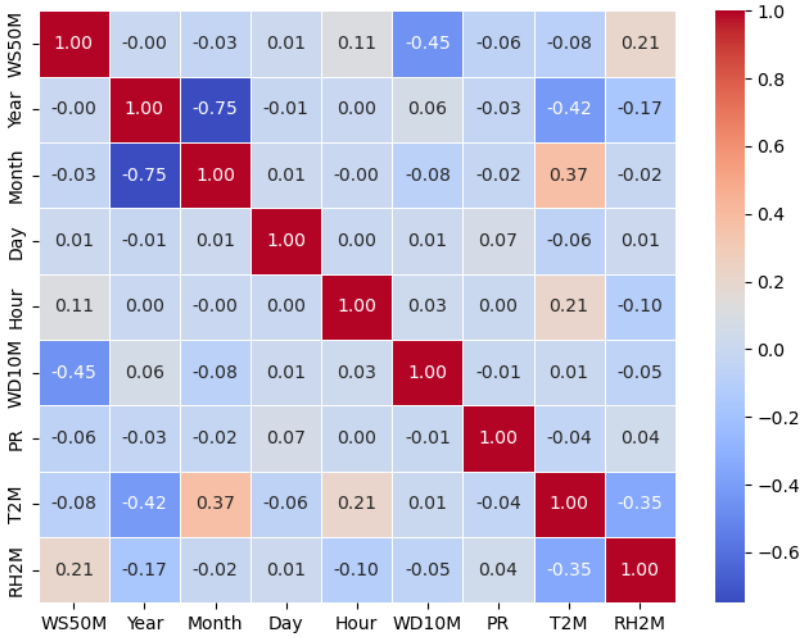

Fig. 3. Correlation matrix heatmap of initial dataset variables

The results displayed in Figure 3 indicate the absence of a linear relationship between the features and the target. This suggests that the *RBF* kernel would be a suitable choice for SVR, as it is capable of handling the non-linear patterns present in the data.

As outlined in Section 2, our new hybrid approach integrates four machine learning models: KNN, SVR, RF, and LR. We compared the results of our proposed model with those of the individual KNN, SVR, and RF models using the R2 score, RMSE, and MAE.

To ensure a fair comparison between the models, we used cross-validation with the training set to determine the optimal hyperparameters for each model. The results are as follows:

- KNN: number of neighbors = 2
- SVR: regularization parameter = 100, and kernel coefficient = 1, kernel = *RBF*
- RF: number of trees = 131

Additionally, exhaustive feature selection was used as a technique to determine the best combination of features that provides the optimal performance for a model. Table 3 presents the features selected for each model.

Table 3. Features selected using the exhaustive feature selection technique for each model

Features / Model	Year	Month	Day	Hour	T2M	RH2M	PR	WD10M
KNN	X	X	X		X		X	X
SVR	X	X	X	X	X	X		X
RF	X	X	X	X				

After determining the optimal hyperparameters and selecting the features for KNN, SVR, and RF, the results of training and testing the models are presented in Table 4. The results indicate that RF delivers the best performance.

Table 4. Performance of standard machine learning models for wind power prediction

	Training			Testing		
Metrics / Models	**R2 score**	**RMSE**	**MAE**	**R2 score**	**RMSE**	**MAE**
KNN	0.95	0.54	0.31	0.85	0.95	0.57
SVR	0.96	0.48	0.24	0.84	1.00	0.65
RF	0.99	0.24	0.16	0.93	0.62	0.42

The objective is to combine KNN, SVR, RF, and LR into a hybrid model to leverage the strengths of each, thereby improving the accuracy and reliability of wind power predictions. The new dataset includes predictions from KNN, SVR, RF, and the actual wind speed as the target. The final prediction is made using LR, which was chosen due to the high correlation among the independent variables in the new dataset. Figure 4 presents the correlation matrix heatmap of the new dataset, and Table 5 shows the results of training and testing our proposed model.

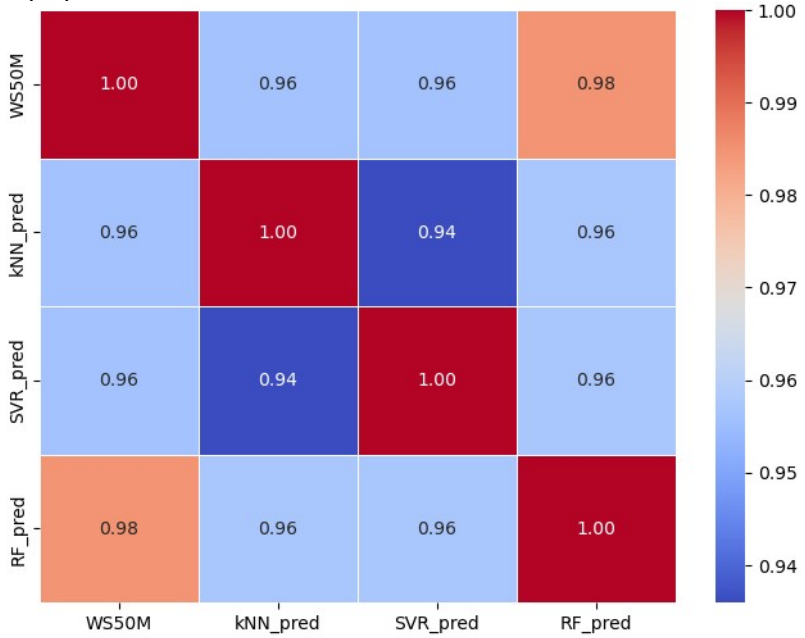

Fig. 4. Correlation matrix of the new dataset

Table 5. Performance of the proposed hybrid model for wind power prediction

	Training			Testing		
Metrics / Models	**R2 score**	**RMSE**	**MAE**	**R2 score**	**RMSE**	**MAE**
KNN-SVR-RF-LR	0.99	0.08	0.06	0.94	0.58	0.39

The performance metrics of standard machine learning models and the proposed hybrid model, as detailed in table 4, showcase the efficacy of each approach in predicting wind power. The standard models, comprising KNN, SVR, and RF, demonstrated robust performance, with RF leading in training accuracy with an R2 score of 0.99 and maintaining strong testing performance with a score of 0.93. The proposed hybrid model, combining KNN, SVR, RF, and LR, significantly enhanced predictive accuracy, achieving an impressive training R2 score of 0.99 and a testing score of 0.94. This hybrid model outperformed the individual models in terms of both RMSE and MAE during testing, indicating its superior capability to leverage the strengths of its constituent models for more reliable and accurate predictions.

4 Conclusion

This study presents that integrating KNN, SVR, RF, and LR into a composite machine learning model enhances the accuracy of wind power prediction. The hybrid model, evaluated using R2 score, RMSE, and MAE metrics, outperformed the individual models, providing improved precision. The findings indicate that the combined strengths of each model provide a robust solution for wind energy forecasting.

Future research could aim to refine this hybrid model by incorporating additional algorithms or applying it to more diverse datasets to increase its applicability and efficiency. Furthermore, the study of deep learning models is crucial in the field of renewable energy, as their ability to handle the non-linear characteristics of weather components can lead to more accurate forecasts.

References

1. D. Gielen, F. Boshell, D. Saygin, M. D. Bazilian, N. Wagner, and R. Gorini, Energy Strategy Reviews 24, 38 (2019)
2. A. Midilli, I. Dincer, and M. Ay, Energy Policy **34**, 3623 (2006)
3. B. I. Cook, J. S. Mankin, and K. J. Anchukaitis, Curr Clim Change Rep **4**, 164 (2018)
4. L. Lledó, V. Torralba, A. Soret, J. Ramon, and F. J. Doblas-Reyes, Renew Energy **143**, 91 (2019)
5. A. Torres-Barrán, Á. Alonso, and J. R. Dorronsoro, Neurocomputing **326–327**, 151 (2019)
6. H. Demolli, A. S. Dokuz, A. Ecemis, and M. Gokcek, Energy Convers Manag **198**, 111823 (2019)
7. U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, Energies (Basel) **14**, 5196 (2021)
8. B. Bochenek, J. Jurasz, A. Jaczewski, G. Stachura, P. Sekuła, T. Strzyżewski, M. Wdowikowski, and M. Figurski, Energies (Basel) **14**, 2164 (2021)
9. S. Park, S. Jung, J. Lee, and J. Hur, Energies (Basel) **16**, 1132 (2023)
10. S. Sayago, G. Ovando, J. Almorox, and M. Bocco, Int J Remote Sens **41**, 897 (2020)
11. I. Merini, A. Molina-García, M. Socorro García-Cascales, M. Mahdaoui, and M. Ahachad, Energies (Basel) **13**, 5979 (2020)
12. Ensemble Machine Learning (2012)
13. S. A. N. Alexandropoulos, C. K. Aridas, S. B. Kotsiantis, and M. N. Vrahatis, IFIP Adv Inf Commun Technol **559**, 545 (2019)

14. Q. Sun and B. Pfahringer, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **7106 LNAI**, 251 (2011)
15. F. Acito, Predictive Analytics with KNIME 209 (2023)
16. R. G. Brereton and G. R. Lloyd, Analyst **135**, 230 (2010)
17. T. H. Lee, A. Ullah, and R. Wang, Advanced Studies in Theoretical and Applied Econometrics **52**, 389 (2020)
18. X. Gang Su, *Linear Regression Analysis: Theory and Computing* (World Scientific Publishing Co., 2009)
19. C. M. T. Khan, N. A. A. Aziz, J. E. Raja, S. W. Bin Nawawi, and P. Rani, Emerging Science Journal **7**, 147 (2023)
20. O. Kramer, Studies in Big Data **20**, 45 (2016)