Neural Network-Based Wind Speed Prediction and Evaluation Using the NREL Wind Integration National Dataset

Haibo Peng[1*], Xinwei Xie[2], Tao Zou[3]
[1]Yunnan Open University, Kunming 650500, Yunnan, China
[2]Zhuhai College, Jilin University, Zhuhai 519090, Guangdong, China
[3]School of Management, Northwestern Polytechnical University, Xi'an 710021, Shaanxi, China
E-mail: penghaibo@ynou.edu.cn
[*]Corresponding author

Keywords: WS forecasting, machine learning models, ensemble techniques, renewable energy prediction, feature importance analysis

Received: August 7, 2025

Accurate wide speed (WS) forecasting is an important factor in optimizing wind energy production and keeping power in a stable condition with lower operation costs. Therefore, this study is going to compare different machine learning (ML) models such as Support Vector Regression (SVR), Decision Trees (DT), Gradient Boosting Machines (GBM), Multi-Layer Perceptron (MLP), and ensemble techniques using Voting and Stacking. These models are trained and tested by taking data from the National Renewable Energy Laboratory's Wind Integration National Dataset, also known as the WIND Toolkit, which is a high-resolution turbine location dataset of wind resources for nearly all locations within the contiguous United States and significant portions of Canada and Mexico. The key performance metrics include testing using MSE, RMSE, and R^2 for each model. In fact, the results confirm that the relative performance of ensemble methods has greater stability and forecasting accuracy, with the Voting model outperforming all the other individual models, thus capturing complex patterns in the WS data. Consequently, SHAP analysis underlined capacity factor and longitude as geographic location features among the most impactful, entailing the importance of such variables in the prediction tasks related to wind. These results underline the potentiality of ensemble techniques in view of robust estimation of WS, which would help the management of renewable energy in a more reliable way. Further, the work will focus on some deep learning (DL) approaches and extra meteorological data in order to improve model performance for the sustainable integration of wind energy into the power grid.

Povzetek: Študija primerja več ML-modelov in ansambelske metode (voting/stacking) za napoved hitrosti vetra na podatkih WIND Toolkit ter pokaže, da ansambli dosežejo stabilnejše in natančnejše napovedi, pri čemer SHAP izpostavi faktor izkoriščenosti in geografsko lego kot ključna vpliva.

1 Introduction

One of the most promising sources of renewable energy is wind power, serving as a very sustainable substitute for fossil fuels. High-precision WS forecasting is an essential requirement for optimal wind energy production, grid stability, and, very crucially, operational cost reduction. These require sophisticated methods for the prediction of this intermittently distributed resource, capable of modeling complex nonlinear patterns in the behavior of winds. ANNs are quite suitable for the task of learning from an enormous quantity of data and generalizing patterns that might not be easily recognizable by traditional statistical model bracelets [1]. The work leverages the comprehensive wind data set, National Renewable Energy Laboratory's WIND (WIND Toolkit), which comprises wind resource geographic coordinates, wind capacity, and capacity factors. Such highly detailed datasets constitute one of the key inputs into a neural network model that can very accurately predict WS across a wide range of locations and conditions. The high-resolution wind data that the WIND Toolkit offers enables researchers to study the variations in wind patterns and can, therefore, increase the resolution of predictive models, addressing one of the primary challenges of renewable energy forecasting[2]. These are traditional and ML algorithms, like SVR, Nu SVR, DT, K-Nearest Neighbors (KNN), GBM, and MLP. Each method has advantages with respect to both linear and nonlinear data; however, the ML models seem much more powerful for high-dimensional and complex structures. The ensemble methods, like the stacking ensemble and voting ensemble, for combining models are intended to improve their accuracy further by capitalizing on each other's strengths

in order to offset their individual weaknesses and to make more robust predictions [3]. The major metrics allowing assessment of the model performance were calculations of MSE, RMSE, and Coefficient of Determination (R^2). MSE and RMSE give the average magnitude of an error for a forecasted variable. R^2 describes how well the model explains the variance of WS. These are the metrics that, upon analysis, the best model emerged as Voting Ensemble, and that goes without saying how such ensemble models will always do better on jobs that involve the prediction of complex datasets like WS [4]. Primary metrics to assess the performance of the model involved computations for MSE, RMSE, and R^2, RMSE and MSE provide insight into the absolute error margins of a predicted value, whereas R^2 measures how well the model describes the variance in WS data. After analysis, the metrics identified the Voting Ensemble as the best model, demonstrating how ensemble models often excel in tasks involving the prediction of complex datasets, such as WS [5]. The findings of the study thus outline that correct WS forecasting models are of prime importance in integrating the variable resource of wind energy into the grid in a sustainable way. Applications in the development of more sophisticated models through ML techniques continue to evolve; this will lead to improved operational efficiency and stability within renewable energy systems. The present work, therefore, overviews, by adapting both traditional and ML models, the practical applications of predictive analytics on renewable energy. For similar methodologies, access to the WIND Toolkit data; see relevant research that has utilized this dataset [6].

2 Literature review

With the increasing demand for clean and renewable sources over the past years, research into wind energy as a friendly and sustainable solution to the environment has increased. Precise forecasting of WS is essential for maximizing wind energy production, cutting operating expenses, and preserving power system stability. Over the years, there have been trials with various models, from mere statistical models to advanced hybrid ones, through which experts seek to improve the accuracy and reliability of WS predictions. This literature review discusses 15 relevant works, detailing the objectives, methodologies employed, findings, and contributions of each author to the field. The work of Blanchard and Samanta [7] was directed at the use of neural networks in the forecasting of WS. It was dedicated to testing whether neural networks are adequate for proper forecasting of WS data, including under conditions of changing meteorological parameters. With the inclusion of meteorological and geographical input, this model proved substantially improved for the purpose of good estimation as compared to the conventional statistical models, especially the short-term forecasts. The findings demonstrated how complicated, nonlinear patterns in the WS data may be learned by neural networks, which could then be used to develop better real-life applications of wind energy resource management and operational efficiency. Nielson et al. [8]focused on improving the prediction of wind turbine power by incorporating atmospheric inputs in ANNs. Such a study was aimed at improved accuracy in the integration of variables like temperature, humidity, and atmospheric pressure with WS. The ANN model resulted in a significant improvement in the accuracy of power prediction compared to models considering solely the WS variable, especially in meteorologically unpredictable conditions. Results showed that atmospheric data helped not only in refining the prediction but also in providing key insight into wind turbine optimization for efficiency and energy management.

Rahman et al. [9] gave a succinct overview of deep neural network forecasting strategies for wind time series. The purpose of this research was to examine how well DNN models performed, more so those that rely on techniques such as LSTM and GRU, in predicting WS over different forecast horizons. The research also proved that DNN-based approaches resulted in a significant reduction of the error rate in WS forecasting compared to traditional models, basing the findings on predictive accuracy and the robustness of the models. These findings highlighted the potential of LSTMs and GRUs to adapt nonlinear and complex temporal patterns that are deemed characteristic in wind data, and hence their effective applications in areas where precise long-term forecasts of wind are needed. Dolatabadi et al. [10]proposed a novel hybrid DL WS forecasting model developed by combining the DWPT technique with a Bi-LSTM network. This work was designed to provide improved accuracy on the basis of the combined use of DWPT for the decomposition of WS signals and the Bi-LSTM network, which would learn temporal dependencies in both forward and backward flows. It performed better than conventional methods in handling the inherent non-stationarities and fluctuations in WS data. The model was effective in producing short- and long-term wind forecasts of high accuracy, thus having great potential for applications in energy management and wind farm operations. Lydia et al. [11]reviewed various forecast models on WS and wind power with an emphasis on enhancing accuracy and reliability for renewable energy. In the paper, the performance comparison of several forecasting models, such as the statistical models, the ML model, and the hybrid models, was pursued to forecast power output and WS. The outcomes showed that the hybrid model, which integrated both statistical and ML approaches, achieved higher accuracy compared to a single model, especially when the wind conditions were fluctuating. In this way, results showed the major role played by hybrid techniques in efficient forecasting, management of wind power, and grid stability to support the integration into variable renewable sources in power systems. Routray et al. [12]had different models adopted for the forecasting of wind power and WS; they drew conclusions on the best model to improve accuracy for renewable energy applications. This study will compare the various models' performance in forecasting WS and output power, including multi-models like statistical, ML, and hybrid models. From the outcomes, it was observed that the hybrid model, when combined with the statistical method, ML, gives better accuracy than a single model, especially when the wind is variable. The findings

demonstrated that the integration of renewable energy sources into power networks and the effective management of wind power and grid stability depend on hybrid forecasting methodologies. Malik et al. [13]had different models adopted for the forecasting of WS and wind power; he drew conclusions on the best model to improve accuracy for renewable energy applications. This study will compare the various models' performance in forecasting WS and output power, including multi-models like statistical, ML, and hybrid models. From the outcomes, it was observed that the hybrid model, when combined with the statistical method, ML, gives better accuracy than a single model, especially when the wind is variable. The outcomes indicated that the hybrid forecasting techniques are very vital for efficient management of grid stability and wind power, and therefore, incorporating renewable energy sources into power systems. Wang et al. [14] developed an ANN-based model that could effectively predict the extreme response of floating offshore wind turbines under different operating conditions. The study was performed to enhance safety and durability through sensible forecasting of extreme load responses due to harsh environmental factors. Using ANN in the modeling allowed for complex nonlinear interactions between wind, wave, and turbine dynamics with reliable predictions for extreme responses. The results brought out that the model can be an enabler in improving both the risk assessment and maintenance planning of the offshore wind energy infrastructure, in view of ensuring stable performance under adverse conditions at sea.

Ahmad et al. [15]have proposed an intelligent control system based on fuzzy logic integrated with the artificial neural network model to enhance stability and efficiency for an offshore wind turbine that floats and has four water columns that oscillate. The objective is to optimize performance dynamically against fluctuating sea and wind conditions. The fuzzy logic-based control system interacted with the ANN in order to enhance the turbine response due to variations in the ambient conditions, thus providing a smoothing effect toward fluctuations both in power output and mechanical stresses. Results obtained from this study indicated that such an integrated approach provided better adaptability and control, while its potential was assessed as one for reliable energy generation to be employed in offshore environments. Assaf et al. [16] examined in depth how well the neural network-based models performed for short-term solar irradiance forecasts. Then, all the neural network models, Convolutional Neural Networks, and Recurrent Neural Networks have been accurate for short-term forecasting, especially when they have been trained based on meteorological and environmental data. Hence, it is underlined that the potential of neural networks to enable solar energy management is to optimize the performance of the photovoltaic system, given valid and reliable forecasts of irradiance. Nielson et al.[8] studied the effect of adding atmospheric inputs to ANNs in improving wind turbine power predictions. There was an attempt to make the predictions more realistic, actually feeding the ANN model atmospheric variables like temperature, humidity, and pressure, in addition to WS. Those additional atmospheric parameters have greatly improved the model's predictive accuracy when compared against models that were driven exclusively by WS. This was effective, in particular, in the capturing of much atmospheric interaction. It showed potentials in optimizing the generation of energy from the wind toward supporting more efficient operations of turbines. This was especially effective in capturing a lot of atmospheric interaction. It showed its potential in optimizing wind energy generation and supporting more efficient turbine operation.

Jiang et al. [17]proposed an automated framework for wind power prediction that integrated multi-time scale analysis and temporal attention mechanisms. The proposed framework leverages the Temporal Fusion Transformer model, which effectively captures temporal dependencies at different scales. This study found that such a method drastically enhances the prediction accuracy of both short- and medium-term forecasts and, therefore, holds great potential to allow better grid reliability and energy management.

Liu et al. [18]proposed the GAT-based model combined with the frequency-enhanced mechanism to improve short-term WS forecasting. The GAT component captures the complex spatial dependencies between WS stations, while the frequency-enhanced mechanism deals with temporal patterns. This hybrid model performed distinctly better than the state-of-the-art approaches, especially in finding the dynamic behavior of WSs within the small-scale forecasting horizon. Wang et al. [19]proposed a hybrid model by incorporating CNN features with the Informer architecture for short-term wind power prediction. Due to the fact that CNN extracts the spatial features, the informer can handle long-sequence time series with effective learning of the dependencies in space and time. It reported superior performance compared to traditional methods for the non-stationary and fluctuating nature of wind power data. Huang et al. [20]explored the application of Transformer-based models for wind power forecasting. The study demonstrated how wind power forecasting jobs can benefit from the use of transformers, which are well-known for their ability to capture long-range dependencies in sequential data. The suggested model achieved higher accuracy and efficiency compared to traditional recurrent neural network models, indicating its potential for improving wind power prediction.

Mukendi et al. [21]have proposed a new methodology regarding WS and power forecasting using shape-wise feature engineering. This work significantly improved the robustness of both CNN-LSTM and auto-regressive models against noise by changing the supplied data's form. The results of this work indicate that this technique gives consistent accuracy for different forecast horizons and, hence, is promising for turbine control systems and resource planning in energy production. Keisler and Le Naour [22] proposed WindDragon, an automated DL framework to enhance short-term wind power forecasting. Numerical weather predictions as well as automated ML techniques are employed in model selection and

hyperparameter optimization for this approach. It has been proved that the predictability from WindDragon ensures better wind energy integration into the power grid as well as helps various efforts toward net-zero carbon emissions.

3 Methodology

This section describes the methodology applied to the development and evaluation of predictive models for WS forecasting by availing ML algorithms and ensemble techniques of the WIND from the National Renewable Energy Laboratory. It involves data preparation, training of models, procedure evaluation, and optimization in order to achieve the best result in terms of accuracy and reliability for WS forecasts.

3.1 Data collection and preprocessing

The data utilized in this analysis were fetched from the WIND Toolkit, a source of excellent wind data that also includes geographic coordinates and wind capacity, along with capacity factors. These inputs form a very important set of variables for training predictive models. First, the raw wind data were normalized to make them more consistent and thus improve the model's performance by reducing biases that could arise because of the presence of different scales among the data. Then, the following data was randomly split into two subsets: 80% for training and 20% for testing, enabling the evaluation and validation of this robust model.

3.2 Model development and training

Several ML algorithms were implemented, ranging from conventional models such as SVR and Nu SVR to more advanced algorithms, including DT, KNN, GBM, and MLP. Each of the models was then trained on the WIND Toolkit dataset to capture both linear and nonlinear patterns that could be helpful in improving predictive accuracy. Hyperparameter tuning for each model was performed by systematic grid searches and also by means of nature-inspired algorithms such as genetic algorithms that refine the most important hyperparameters: C, ϵ, and γ-when it came to SVR models. As for the ensemble models, there were two main techniques that improved the prediction accuracy of algorithms. In its process, the Voting Ensemble method pools predictions from each individual model involved, whereby a vote is assigned to each based on the output given. However, the Stacking Ensemble technique takes various models as base learners whose outputs have to go through another meta-learner for the actual prediction. In that respect, this stacked approach leverages the strengths of individual models, whereby it places the ensemble in a position to make up weaknesses that could be common in just one model. A model of wind power forecasting begins with six steps in order, as depicted in Fig. 1. First, Step 1 is data collection and the preparation of that data to analyze; Step 2 is the determination of necessary input features and target values that one would need from the dataset. It then further divides the dataset into training and testing subsets in a ratio of 80:20, respectively, on which model training and evaluation are based. Step 4 describes the wind power forecast by using different ML algorithms: SVR, Nu SVR, DT, KNN, GBM, and MLP. It involves two ensemble techniques, namely Voting and Stacking, applied in Step 5 to improve accuracy by aggregating the strengths represented by the different models. Lastly, Step 6 identifies the best-performing model according to several performance metrics, which is Voting Ensemble.

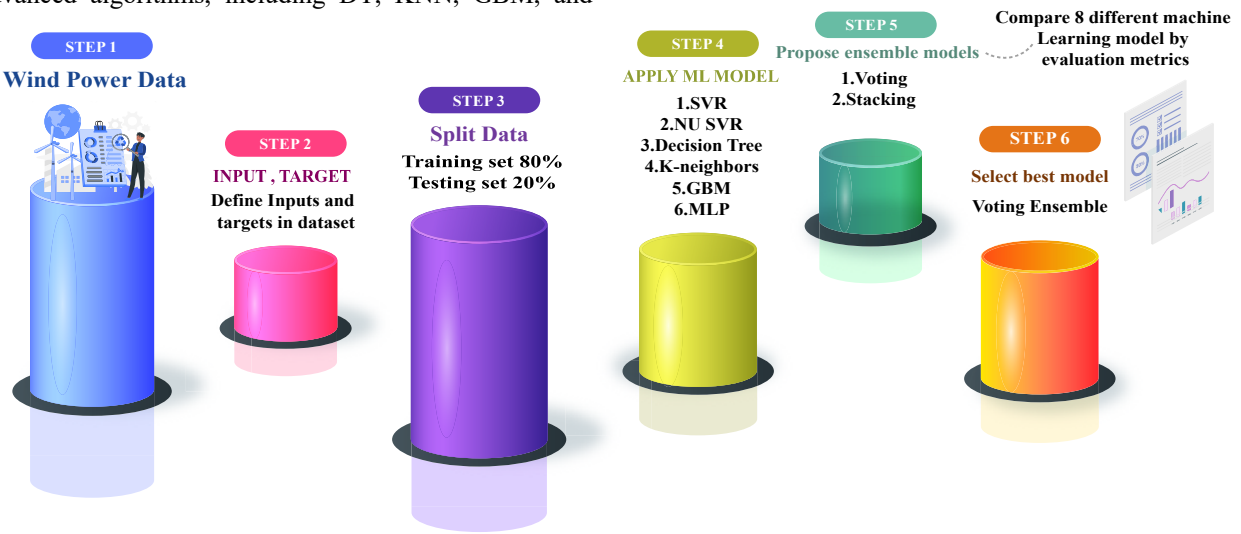

Figure 1: Workflow for wind power prediction using ML models

The organized methodology for this is presented in Fig. 1 and represents one of the most holistic wind power prediction methods wherein both ML models and ensemble techniques have been combined to provide strong and precise forecasts. Fig. 2, The KNN classification process for a target point within a two-class dataset. The figure above shows the identification of the target point and segregation of the dataset into two classes: Class 1 in red and Class 2 in blue. The second panel shows the KNN distances computed from the target point to its nearest neighbors, drawing on the target to connect to the nearest points in both classes. In the third panel, it does the classification by finding the majority class in a defined neighborhood. This neighborhood is usually taken to be a

circle around the target point, within which it searches for the majority of the classes of the points that fall there. Accordingly, based on the majority of neighbors from those points, the target point will fall into one of the two classes, hence illustrating how KNN assigns a class label based on spatial proximity.

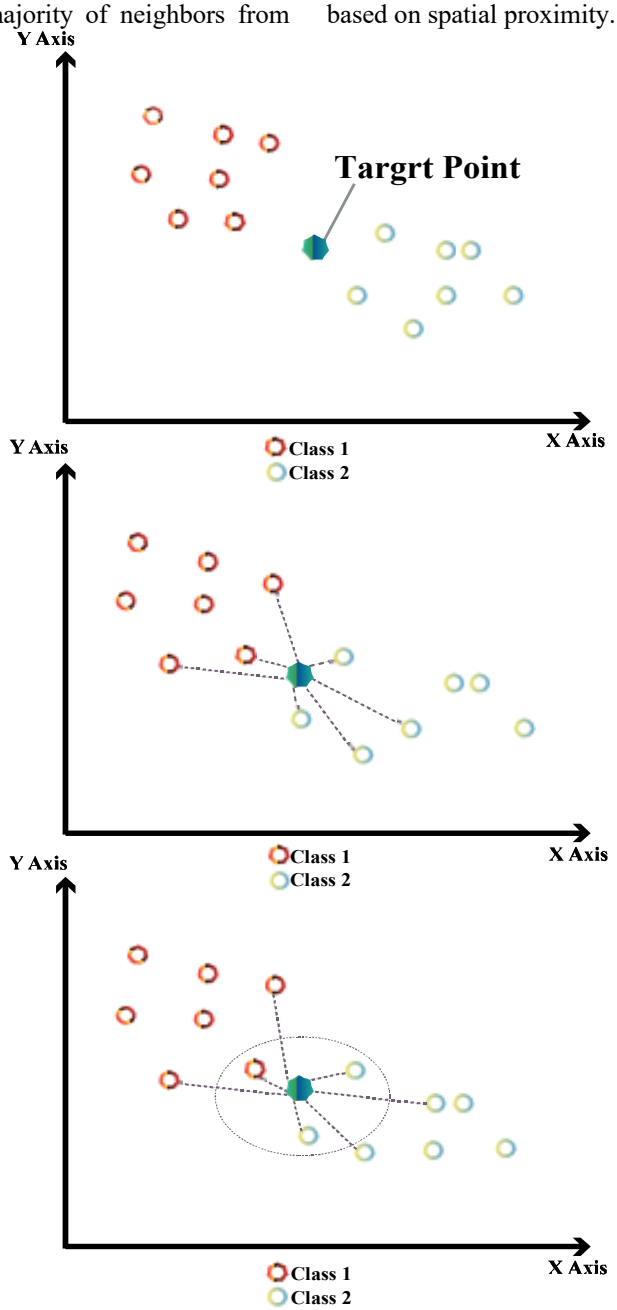

Figure 2: KNN classification process for the target point

Fig. 3 shows the architecture of the MLP, a general class of ANNs used for predictive modeling and classification tasks. The major constituents of the network are three layers, including the Input Layer, which is the layer fed with raw data input; one or more Hidden Layers, where neurons or nodes process and transform inputs through weighted connections and an activation function, respectively; and the Output Layer, which provides a prediction or classification based on the transformed data. Each neuron connects itself to neurons in the adjacent layer in a fully connected manner. This enables the network to learn complex nonlinear relationships within data. Generally speaking, through training, the weights within an MLP will be adjusted, which will enable the model to generalize patterns in order to be more accurate. Thus, this is a diverse model that can be applied to many different ML applications.

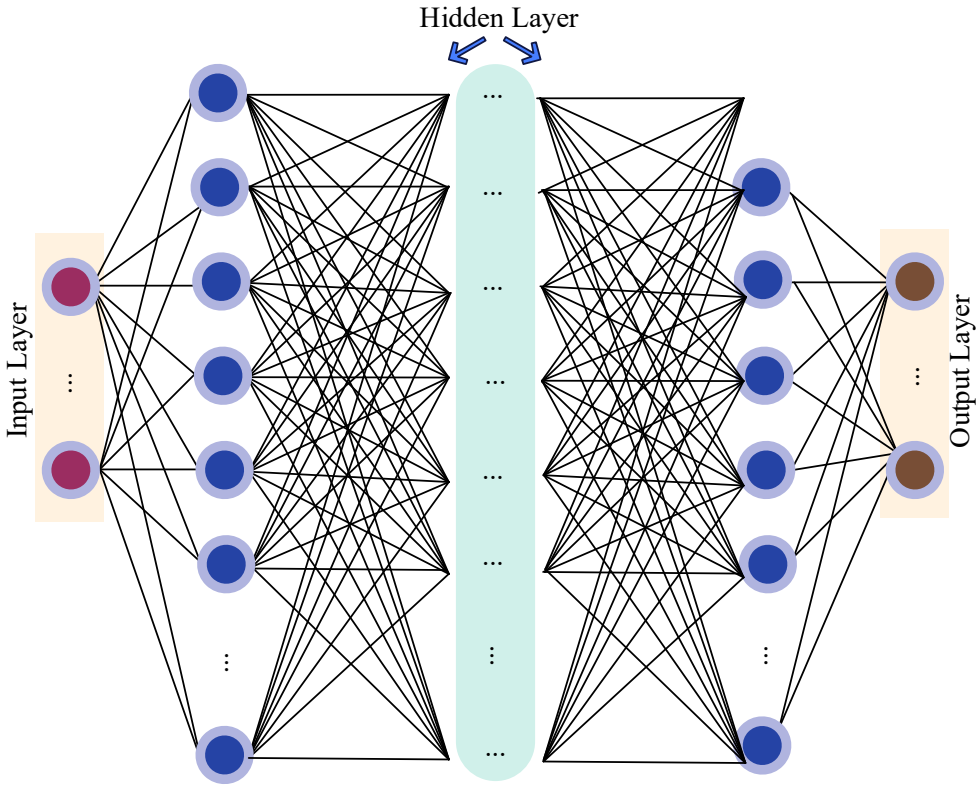

Figure 3: The architecture of an MLP network

Fig. 4 shows the structure of one variety of DT, which serves as a guide for sequential decisions. It is attached to the top node, and each decision node- a diamond-shaped icon- offers a 'yes'/'no' choice, each of which then splits into two further divisions. Each individual pathway illustrates one possible stream of decisions contingent upon some condition or criterion. This process continues down through the tree until terminal nodes, labeled as Steps, such as Step 1 and Step 2, are reached, from which the ultimate outcome or action is determined by the previous decisions. The structure can also be straightforward: a step-by-step breakdown into complex decision processes for intuitive, organized problem-solving, classification, or predictions. DT finds broad applications in ML, especially in applications where decisions should be interpretable and structured in rules.

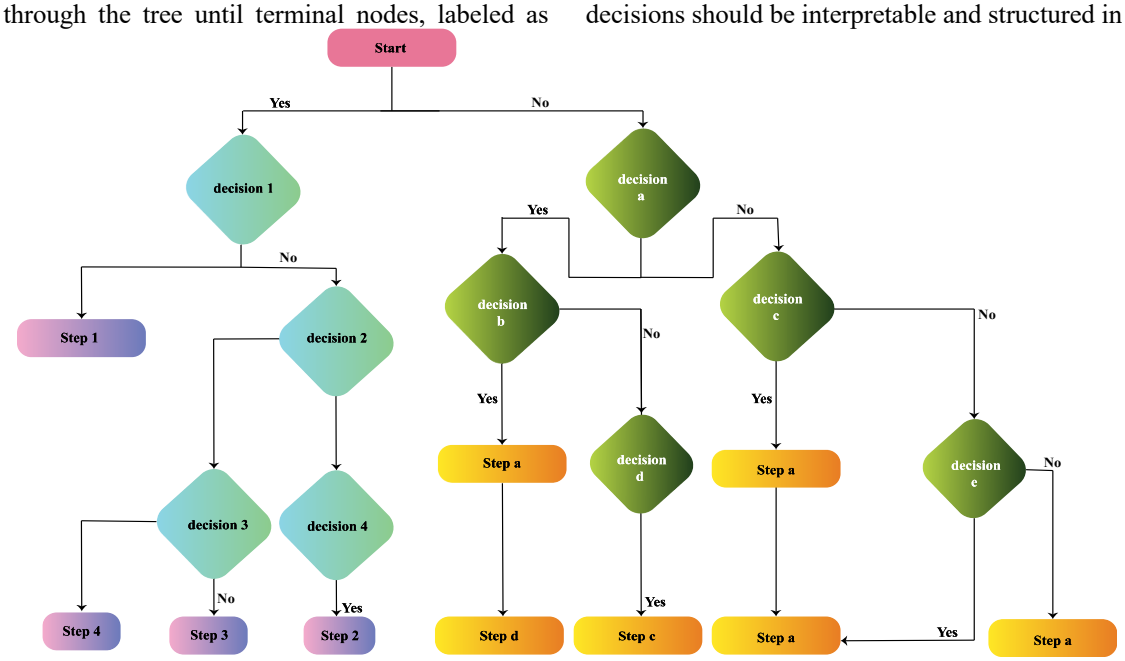

Figure 4: DT structure for sequential decision-making

Fig. 5 represents the training procedures of one of the most robust ensemble learnings in predictive modeling, which is the GBM. Where GBM constructs a lot of DTS in a greedy manner, each one of the trees tries to correct

mistakes made by the previous one. It starts by letting the model make predictions, using Tree 1 on the input data (X, Y), thereby returning residual errors $r_1 = y - \hat{y}_1$. In subsequent iterations, each tree (Tree 2, Tree 3, etc.) is trained on the residuals from the previous step (e.g., $X, r_1, , X, r_2, and\ so\ on$). This process is repeated recursively until the final tree, say Tree N, has been trained in the hope of further minimizing the prediction error. In progressively reducing these errors, GBM constructs one strong model that amalgamates the predictions of all the trees to make an accurate prediction. Hence, it works extremely effectively to handle complicated non-linear data relationships.

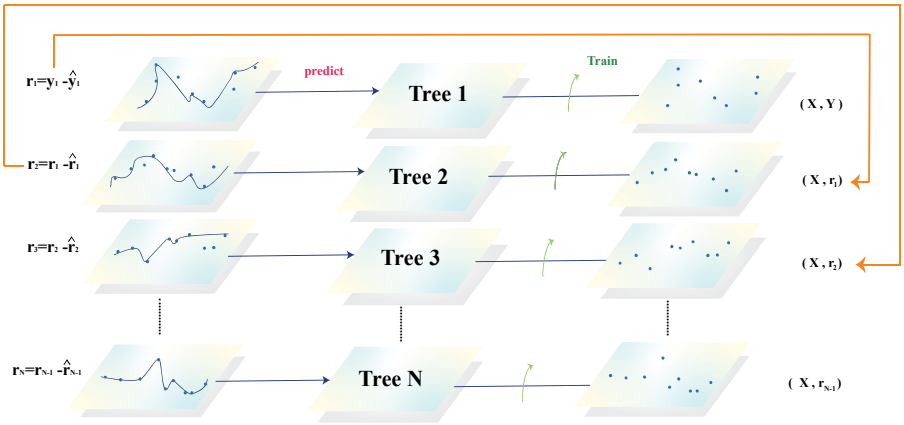

Figure 5: GBM's training process

Fig. 6 is the flowing process for hyperparameter optimization of the Nu SVR model. The flow starts with a normalization of data, followed by splitting the data into training and test sets at random. Initial parameters for the SVR are set, and the model is trained, followed by an evaluation. The approach utilizes nature-inspired algorithms, like genetic algorithms and particle swarm optimization, to adjust the SVR hyperparameters if it fails to achieve the stopping criterion. The optimized parameters are then experimented with on the SVR model. This will be an iterative process until it meets the stopping criteria, where one can obtain the final optimized SVR hyperparameters. It is a structured approach meant to ensure a well-tuned model that is able to fit the outcomes with the least possible error.

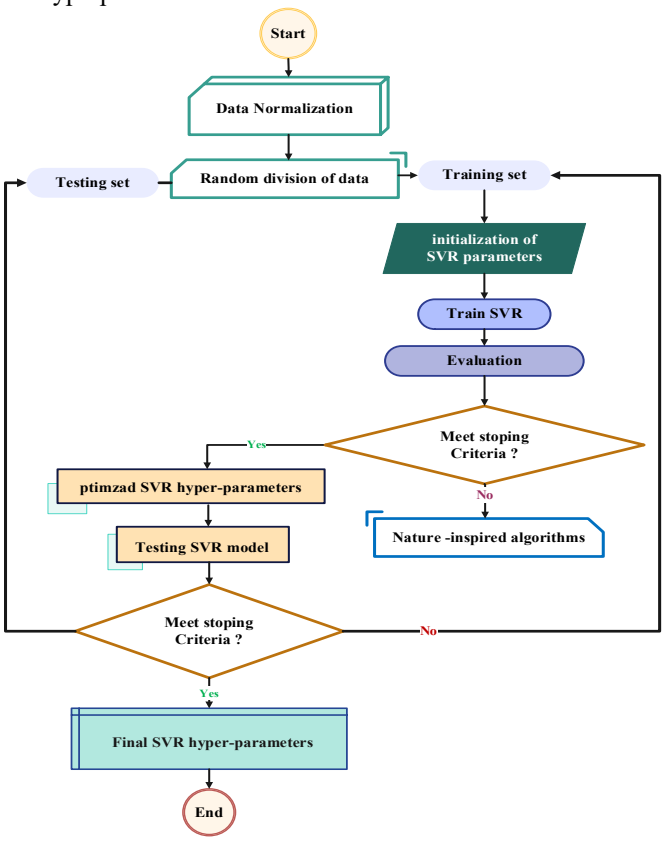

Figure 6: The hyperparameter optimization process for Nu SVR

Fig. 7 describes the steps to develop and test the SVR model. The steps start with the collection of experimental data. The following step is data preprocessing, which cleans the experimental data to make it ready for modeling. The resulting set of data is split into two subsets: Training data and Testing data. Identifying the critical hyperparameters of the SVR model, namely C, ϵ, and γ, is then performed by choosing those that optimize the performance of the model. The model goes through both SVR Training and SVR Testing to check its accuracy. Further, the model is subjected to measurement of correlation value between predicted and actual data to assess its prediction capability. This systematic workflow allows the SVR model to be well-trained and tested in order to yield reliable results.

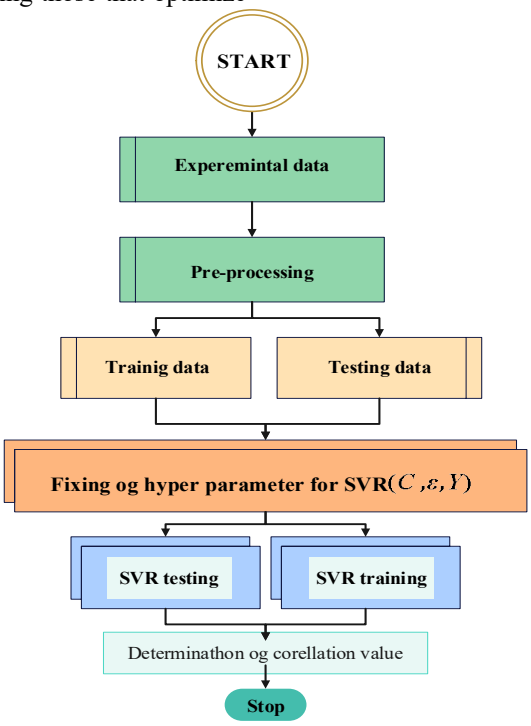

Figure 7: Workflow for SVR model training and evaluation

Fig. 8 provides an overview of the architecture in a Stacking Ensemble model. It presents a very general strategy in ML, where predictions of multiple base models are combined into one general model that normally has better performance. Given a dataset D, several separate models may be individually trained on this data, each learning something different from it. These models' outputs then form the inputs to a meta-learner or meta-model that learns to weigh and combine these predictions to come up with a final output. The Stacking Ensemble approach tends to result in even better predictive accuracy and robustness since it leverages the strengths of multiple models. This kind of structure is particularly useful when complex prediction tasks are at issue since it boosts performance by reducing biases specific to single models.

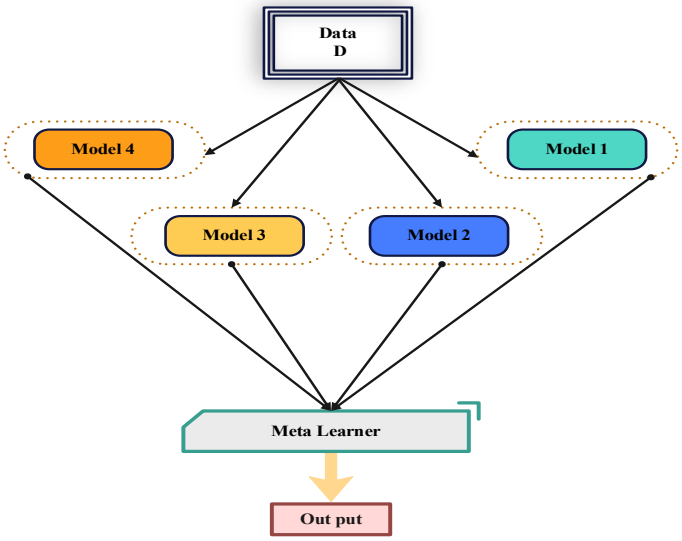

Figure 8: Stacking ensemble model structure

Fig. 9 is a multi-step image processing workflow using a voting mechanism to localize a target position within an image. Step 1 is whole scales the image through a low-pass filter to blur noise, then calculates at every location in the image the gradient to detect edges or intensity changes; Step 2 votes-sets an initial vote image to zero and starts a voting process were. The algorithm, therefore, calculates votes based on directional updates and decreases the angle parameter ϕ, gradually approaching a stopping criterion. Step 3 (Localize) completes localization by marking the final local maxima, followed by a threshold, in order to highlight the most probable positions regarding the target location. This technique comes in handy for object detection and position tracking applications where accuracy for such tasks is improved by iterative voting and thresholding.

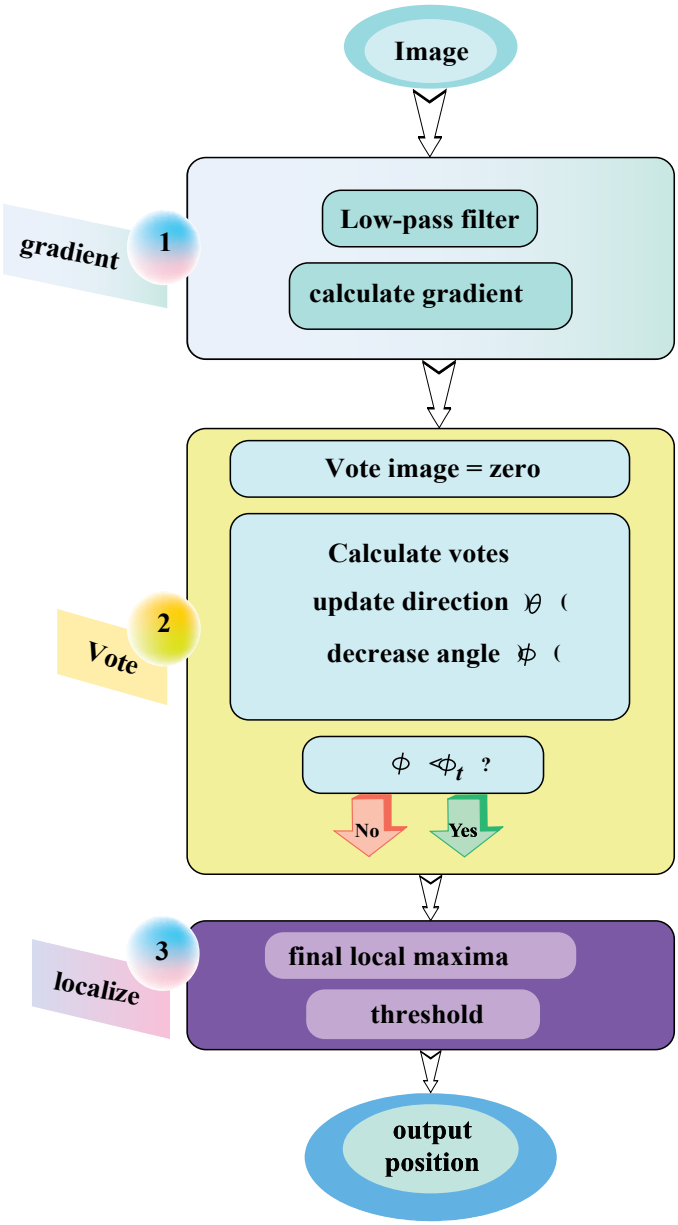

Figure 9: Image processing and voting mechanism for position localization

3.3 Evaluation metrics

All models, including the surrogate models, were tested against each other based on a set of metrics. Fig. 10 presents some common mathematical formulae along with metrics widely used for diagnostic analysis related to predictive model performance. These metrics span from the RMSE, representing the average magnitude of the prediction errors, up to the mean absolute error (MAE), that is, the average absolute difference between the estimated and actual values. MSE and the median absolute percentage error (MDAPE) further quantify the accuracy of the predictions captured by squared deviations and percentage errors, correspondingly. These metrics represent an overall seminal framework for assessing and comparing model accuracy, reliability, and suitability with regard to different tasks of prediction.

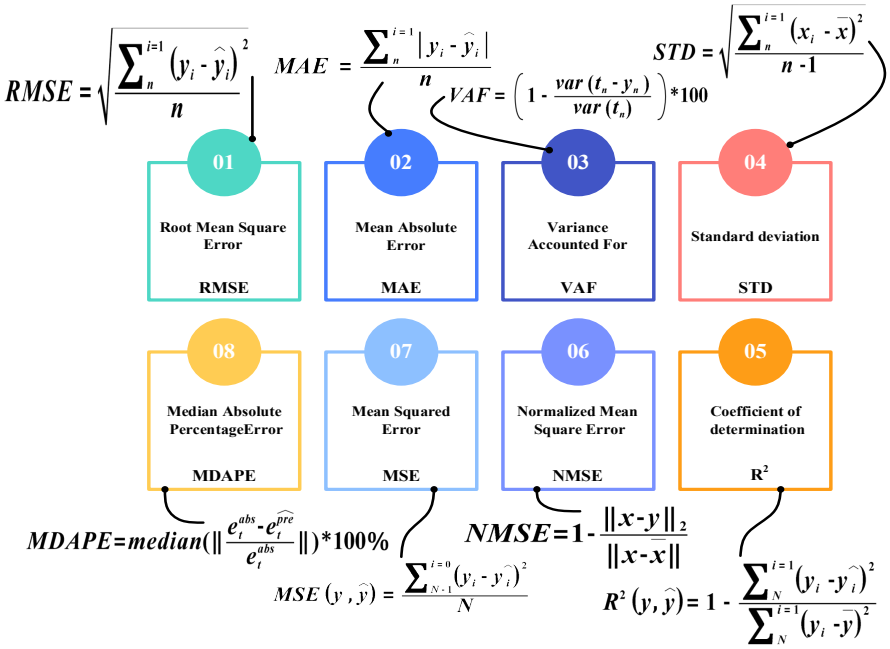

Figure 10: Performance metrics for model evaluation

The Root Mean Square Error (RMSE) is a measure of the magnitude of the prediction errors, heavily penalizing larger errors. MAE, being the mean absolute difference between the values that were anticipated and those that were seen, provides a straightforward measure of the size of the error in the predictions. Variance Accounted For (VAF) measures the proportion of variance in the actual values explained by the model as a percentage. Standard Deviation (STD) reflects the spread or variability of values around the mean. R^2 indicates how much of the variance in the actual data is captured by the model. Normalized Mean Square Error (NMSE) compares the model's error to the data's variance, helping assess relative prediction accuracy. The MSE is the mean of the squared differences between actual and predicted values; it penalizes larger errors. The MDAPE is precisely the median of the absolute percentage errors, being, therefore, less sensitive to outliers.

y_i: The actual (observed) value for the i-th data point in the dataset.

$\hat{y}_i$: The predicted value for the i-th data point is generated by the model.

n: Overall number of observations or data points in the dataset.

$\bar{y}$: The mean of the actual values in the dataset, calculated as $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

t_s and y_s: t_s in the VAF formula refers to the actual value, while the t_s is a predicted value at a certain instance or sample.

Var (t_s): It means the variance of the actual values, which is a measure of the dispersion of data around the mean.

x_i: A particular observed value within the dataset (employed in the formula for STD and NMSE) that will relate to a data point.

$\bar{x}$: The mean of the observed values x_i in the dataset, similar to $\bar{y}$ but used in different contexts in certain formulas.

e_{bs}^{th} and e_{bs}^{ref}: In MDAPE, e_bs^th is the theoretical value, and e_bs^th is the reference value, used in most fields where the relative error of measurement applies

N: Another representation of the total number of data points, sometimes used interchangeably with n.

$\|x - y\|_2$: In NMSE, this is used to refer to the Euclidean distance between two sets of values, x, which are actual, and y, which are predicted, and this captures magnitude information on the error.

C, ϵ, and γ: are the most general hyperparameters in SVR, though it was not shown here. These are parameters governing the margin, tolerance bound, and the kernel of the model.

4 Results and discussion

In this part, experimental outcomes are presented regarding a review that was done on the application of various ML models and ensemble techniques in the forecasting of WS using data from the WIND Toolkit, a product **of** NREL Wind Integration National Dataset. The results will highlight the performance evaluation of every different model, using the Mean Squared Error (MSE), RMSE, and R^2 metrics that will grant the ability to analyze the exactness and reliability of each different approach. The following paper compares traditional models, such as SVR and DT, to advanced ensemble methods for the optimal model configuration that robustly represents nonlinear variability in WS data. This, therefore, serves to bring out various weaknesses and strengths of each individual model while at the same time attempting to provide insight into how several models through ensemble methods could possibly improve predictive accuracy, particularly when it involves dealing with intrinsic wind variability. The results from each model are detailed in subsequent sections, along with their discussion in relation to implications for predicting WS in applications

involving renewable energy. The heatmap of Fig. 11 presents the correlations for five of the variables in the dataset: longitude, latitude, capacity, capacity factor, and WS.

The Pearson correlation coefficient's strength and direction are displayed by color intensity, with lighter hues denoting weaker or negative correlations and deeper blues denoting stronger positive correlations. The capacity factor is strongly positively related to WS at 0.8, and it could be said that, in general, this factor increases with increased WS, as higher WSs would normally imply higher energy output from wind turbines. The capacity factor increases moderately with latitude at 0.32 and longitude at 0.28. These may reflect some geographic influence on wind conditions and, hence, energy production. Other relationships are much weaker-for instance, the near-zero correlations among capacity and both longitude 0.014 and latitude -0.059 would suggest that, for this dataset, geographic location may have little effect on the actual placing of a turbine in terms of its capacity. The heatmap, in general, provides insight into how key variables interact by underlining the importance of WS in determining the capacity factor while showing a minimal geographic dependence on the capacity. This, in turn, can inform feature selection during predictive modeling by focusing on the more strongly correlated variables as one aims to develop an accurate model.

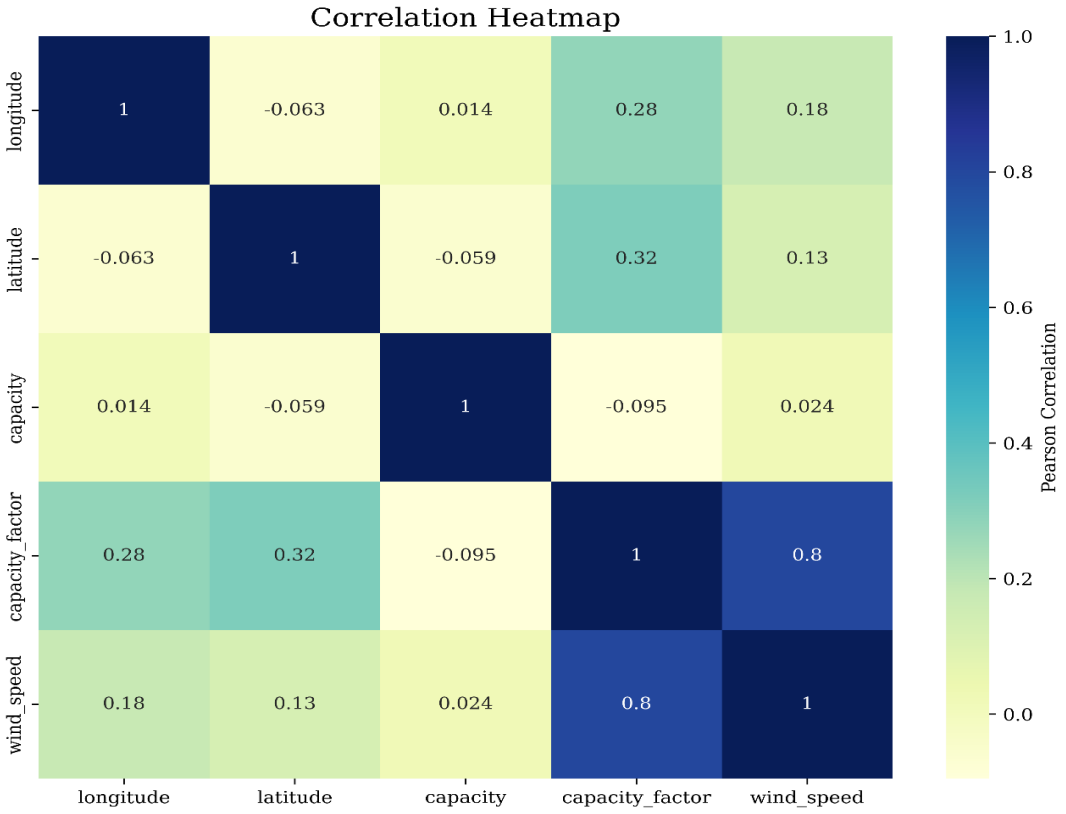

Figure 11: Correlation analysis of key wind energy variables

Fig. 12 shows the forecasted wind velocity values using the various models up to DT, GBM, KNN, MLP, Nu SVR, and SVR plotted against target values across a series of data points, i.e., row numbers. The topmost black line indicates the actual values of the target variable; each model's prediction is plotted as a different color line below it. Another insight that this plot shows is that some of the models tend to be stuck much closer to the target line, like the GBM and DT, and some are very off the target values, such as the SVR and Nu SVR models. This means that the ensemble methods such as GBM, along with tree-based models like DT, turn out to be more accurate; the best result comes in terms of prediction for the WS data. Because performance by SVR and Nu SVR may come out comparably poorer due to non-linear variation, it lacks the capability to capture the WS variations in a non-linear pattern. The above visualization strongly indicates the model performance difference. Hence, GBM and DT are the preferred data sources. Consistency in the predictions from each model on the available data tells us something concerning the resilience and dependability of the models in forecast applications.

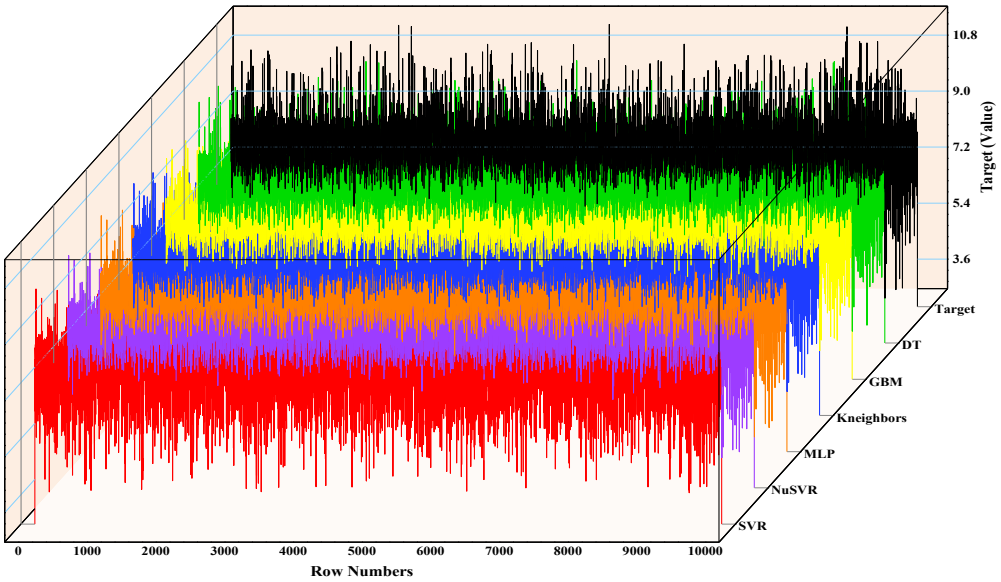

Figure 12: Comparison of model predictions against target WS values

Fig. 13 shows the comparisons between different models: DT, GBM, SVR, KNN, MLP, and Nu SVR. Different plots are drawn to show forecasted WS against actual values for all models. In each plot, the R^2 score of training and testing is also shown to present the strength of the model in explaining the variance in data. From the DT model, even though the R^2 score is very high on training, reaching 0.999, which depicts an almost perfect fit, there is a depression in the R^2-in-test at 0.861, which can be a possible sign of overfitting. The NBA with Theenza GBM, KNN, and MLP all depict performance that is balanced between the training and test sets since the respective R^2 scores are close to each other, therefore generalizing with less overfitting. Among them, KNN presents a test R^2 of 0.895, while GBM follows close at 0.887 and the MLP at 0.881. In contrast, SVR and Nu SVR depict very low R^2 values for both training and testing sets, which reflects the overall relatively poor performances of these two algorithms in modeling the underlying pattern of variation within the data. Overall, KNN, GBM, and MLP emerge as the most reliable models with firm generalization capabilities, while DT may not be that reliable for real-world applications due to overfitting, though with high training accuracy. The above analysis gives an idea that the ensemble and neighborhood-based methods might turn out to be more appropriate for the given task of WS prediction.

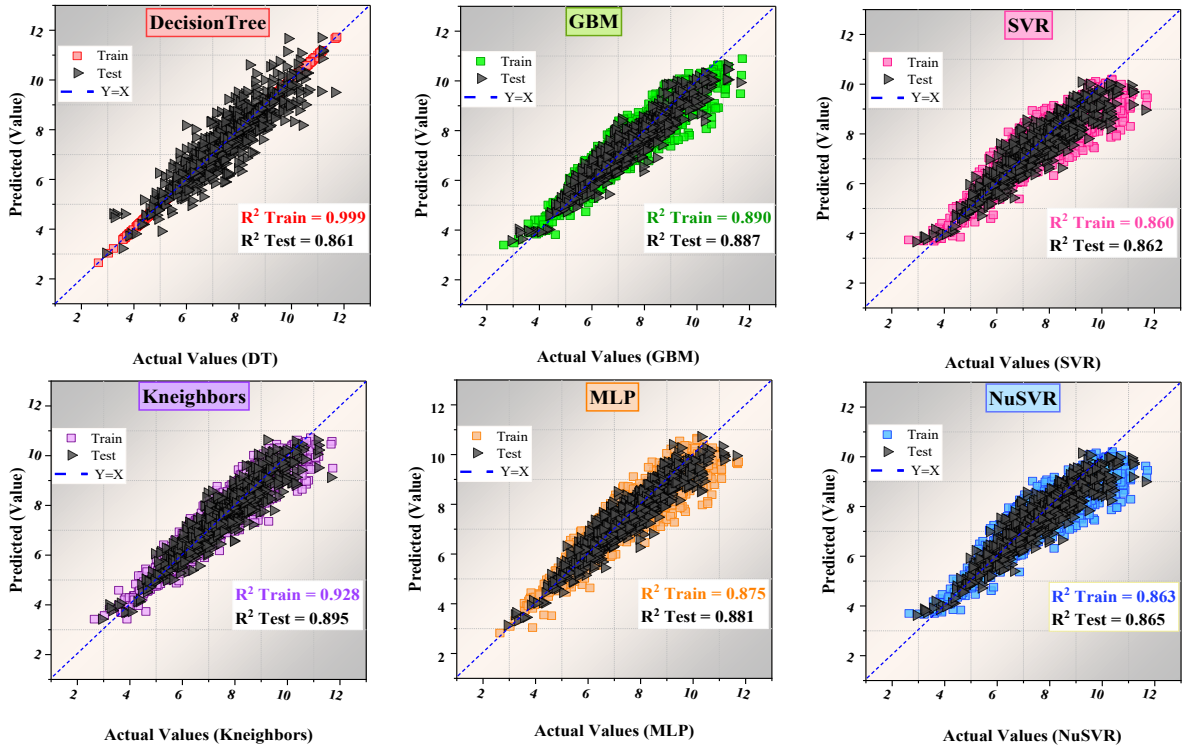

Figure 13: Model comparison based on predicted vs. actual WS values (R^2 analysis)

Fig. 14 is a distribution plot of the prediction error for each model, split into training and testing datasets. Each model error is color-coded as horizontal lines; zero error is colored red and indicates perfect predictions. First, the narrow error spread around zero in the DT and GBM models demonstrates that the model is highly accurate with less deviation in predicting actual values. On the other hand, some of the model results show a wider distribution of errors, such as Nu SVR and SVR models, reflecting larger deviations from actual values more often, especially within the test set. In particular, all the models show smaller error dispersion in the training set, while the test set can widen such dispersions, which is the typical case of loss when the models generalize to unseen data. In particular, the performance is generally maintained by models like DT and GBM but fails to generalize in the case of others, like Nu SVR. In general, the narrow distribution of the errors around zero obtained for both DTs and GBM models denotes their robustness and suitability to the variance in WS prediction tasks represented in this dataset.

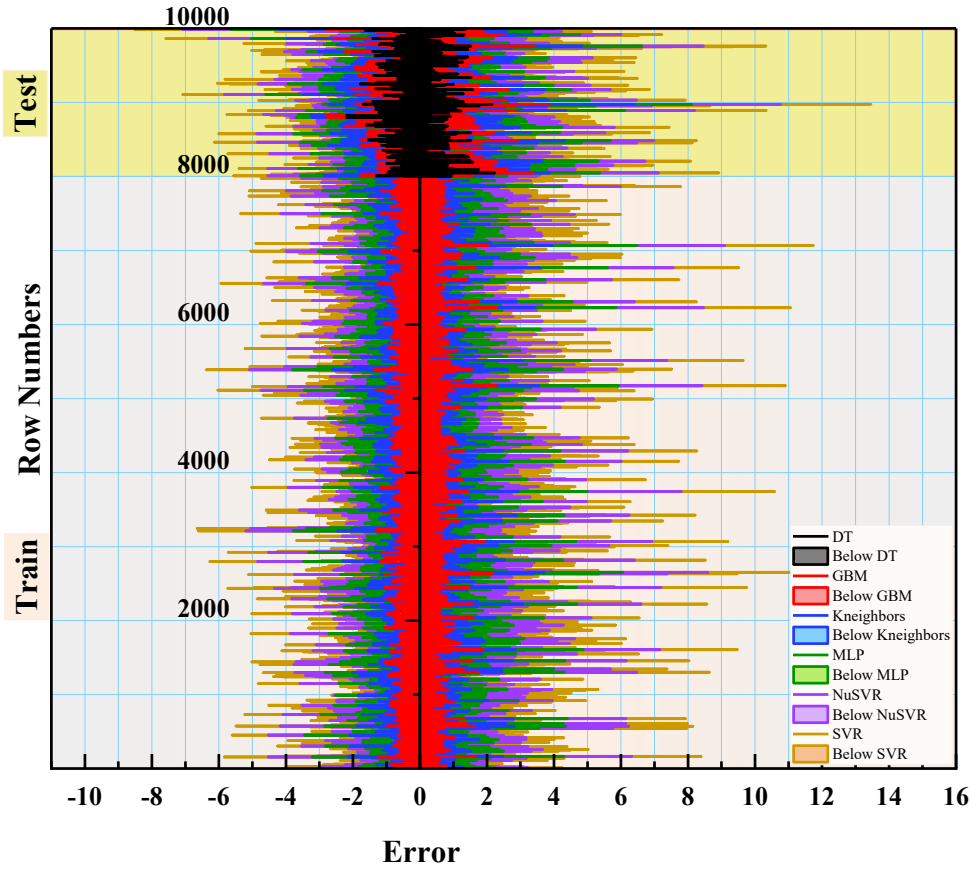

Figure 14: Error distribution of predicted WSs across training and testing sets

In Fig. 15, the boxplot displays the range and distribution of prediction errors for different models divided into training and testing sets. Since the boxplot IQRs center on zero, with their whisker extensions holding the outliers, this suggests that most of each model's error range is centered around zero. Models like DT and GBM have relatively narrow error distributions, especially in the training set, reflecting through the small values of IQRs that most prediction errors are close to zero. In contrast, models such as Nu SVR and SVR had a larger spread in the distribution of error, which had more evident outliers, indicating that their type of models might have generalization issues and be more sensitive to the great deviations of predictions on unseen data. However, KNN, MLP, and GBM presented very similar performance in the training and test sets, with a minimum spread of error, which turned into robust and reliable results. Boxplot analysis shown here suggests that ensemble and neighborhood-based models like GBM and KNN might provide more robust and really accurate predictions, while models like Nu SVR are more prone to variability, especially on new sets.

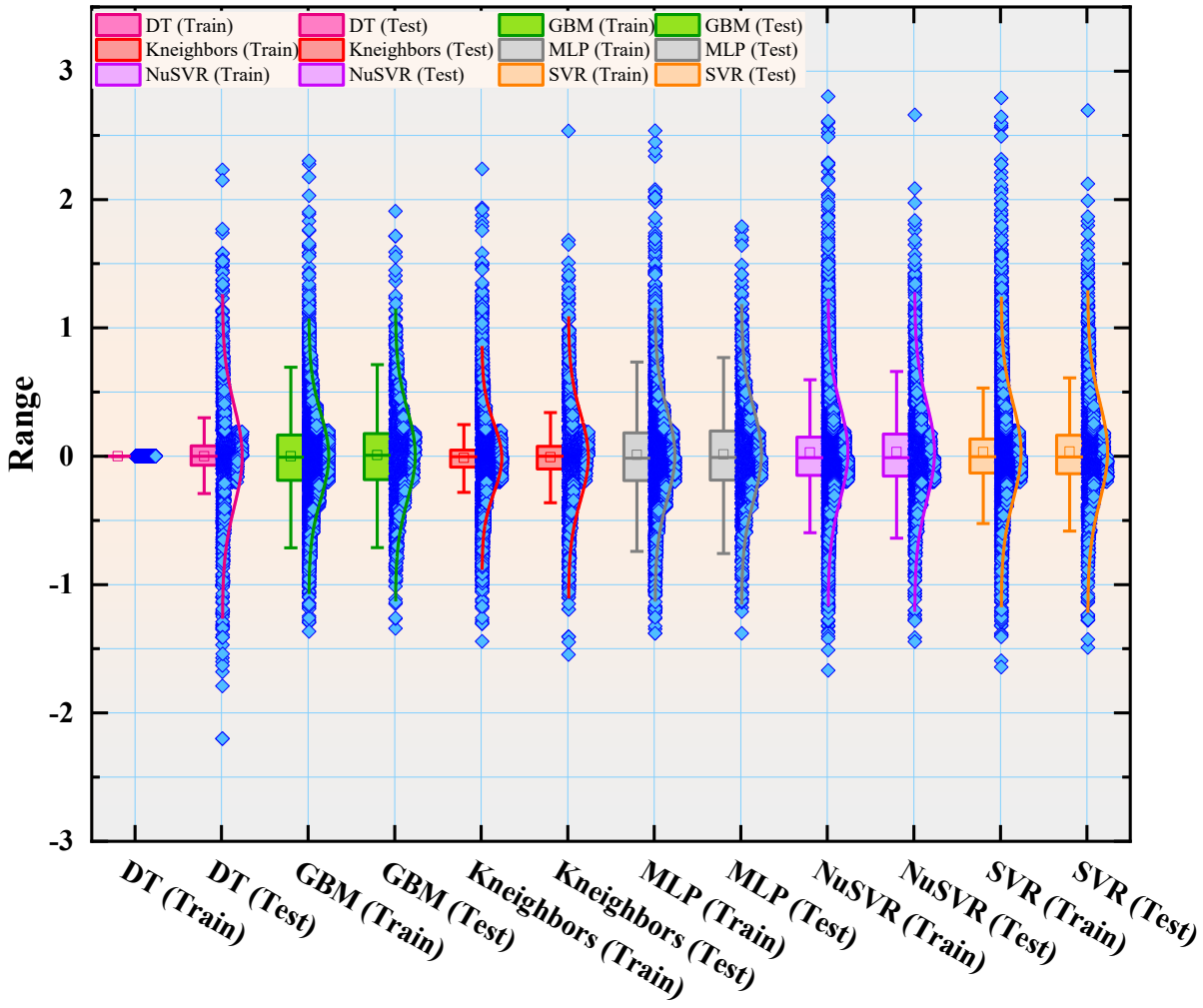

Figure 15: Boxplot comparison of prediction error range for training and testing sets across models

Fig. 16 is a comparative bar chart of several important performance measures: MDAPE, VAF, R^2, and STD for six models based on training and testing sets. DT and MLP models show high R^2 and low MDAPE when training but undergo strong declines on the test set, indicative of overfitting. KNN and GBM plots showed balanced R^2 and VAF between training and testing and very low MDAPE and STD _dev values, hence with more robust generalization and relatively equal precision within sets. Nu SVR and SVR models showed considerably lower R^2 and VAF values, hence capturing the pattern underlying the data effectively. On generalizing and keeping accuracy, overall, the most reliable models are KNN and GBM; DT and MLP train well but unfortunately considerably lose their reliability on unseen data, while SVR-based models are the weakest in this lot.

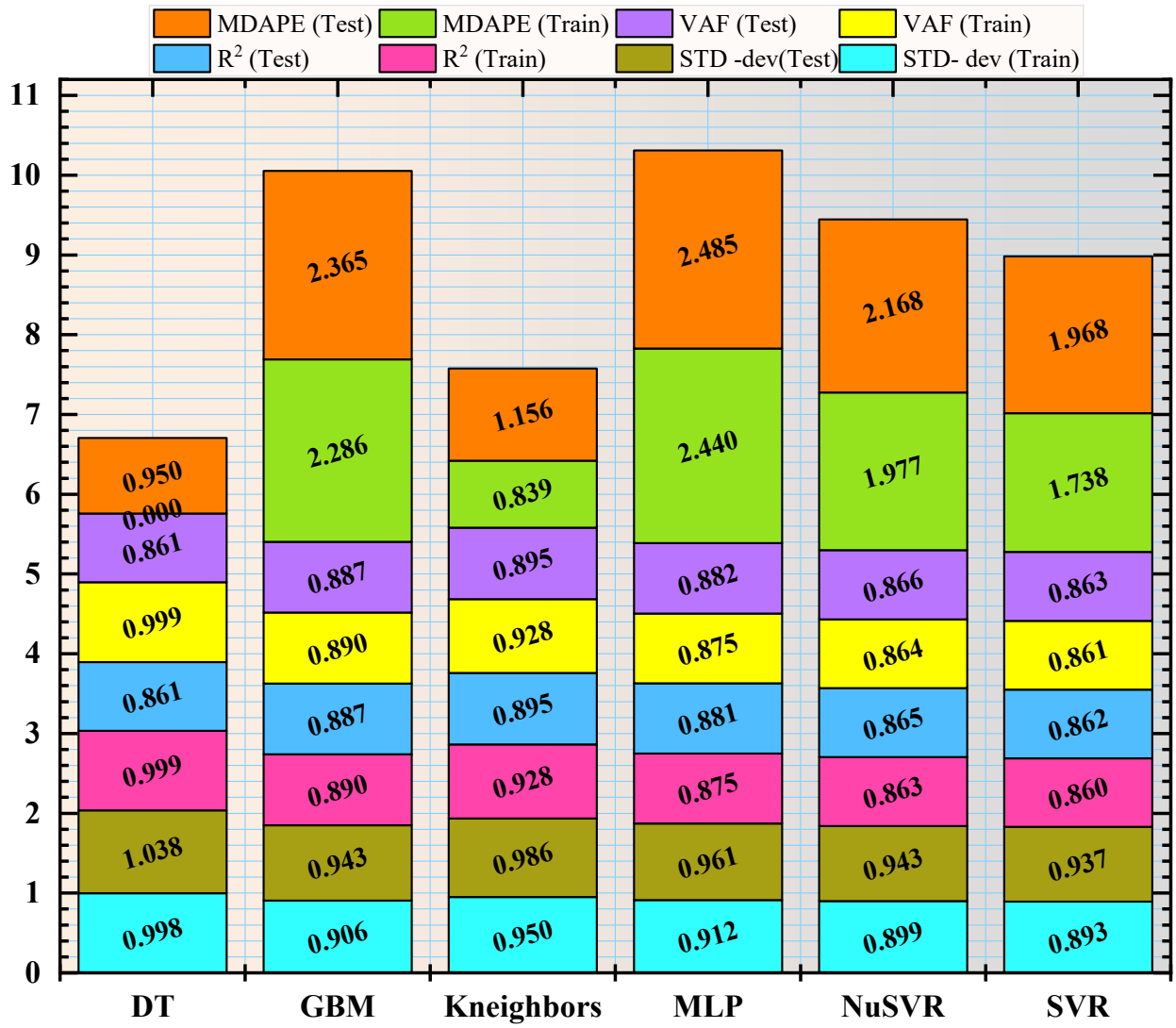

Figure 16: Comparative evaluation of model performance metrics across training and testing sets

Fig. 17 is a performance comparison of six models, namely DT, GBM, KNN, MLP, Nu SVR, and SVR, on four error metrics MAE-Mean Absolute Error, NMSE, MSE, and RMSE for both the testing and training datasets in a stacked bar chart. The DT model gives the minimum errors of both training and testing with very low RMSE and MAE values, indicating high accuracy. The KNN and GBM models also result in relatively balanced training and test set errors, thus indicating good generalization capability. On the other hand, Nu SVR and SVR models rank among those with the highest value of RMSE and MAE, especially for the test set, reflecting weak predictive accuracy with not-so-effective generalization capability. MLP performs moderately, with relatively higher errors within the test set than in the training set, but still outperforms the models based on SVR. Overall, DT, KNN, and GBM are the most accurate and reliable models, while Nu SVR and SVR present relatively high errors, which may thus be less appropriate for this data set.

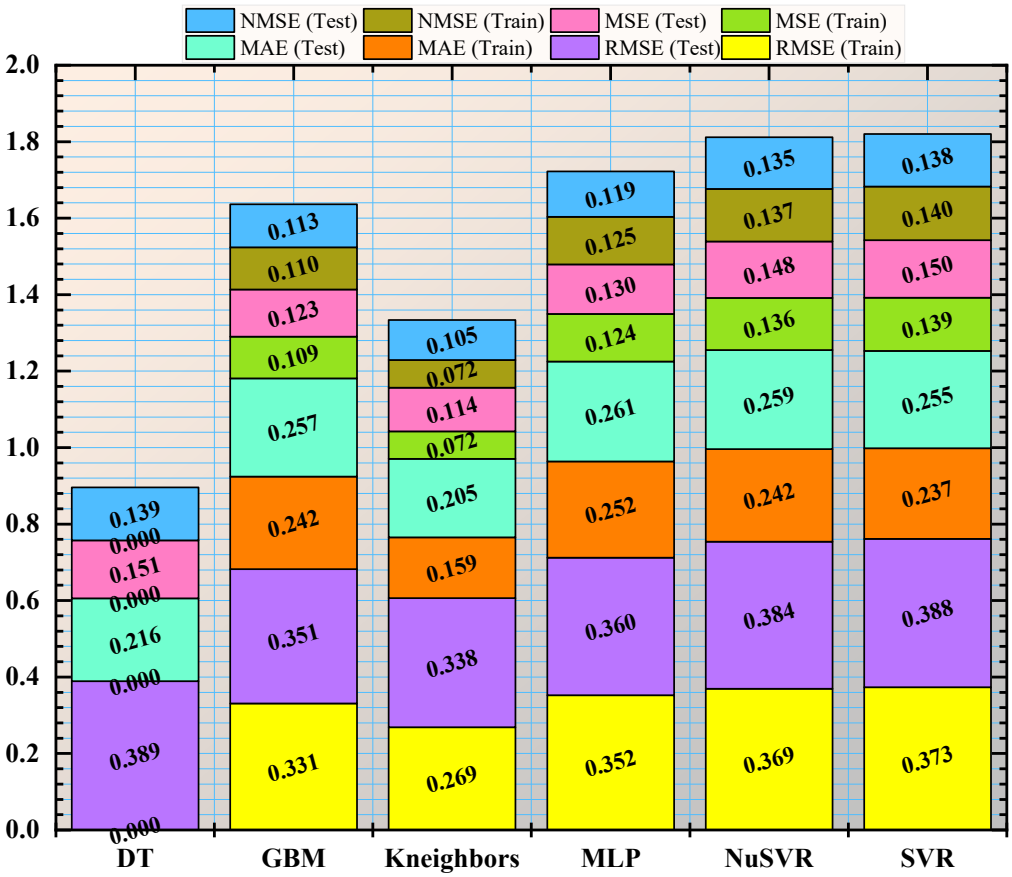

Figure 17: Comparative analysis of error metrics across models for training and testing sets

Fig. 18 is a Taylor diagram, which shows the comparison of the different models' performance visually by plotting their performance regarding the STD, correlation coefficient, and centered RMSE for the WS variable against the observed data point (yellow star). The ideal model should align with the observed data point at the correlation coefficient of 1.0, the STD of 1.0, and the smallest RMSE. The models that are closest to the observed point, such as the Voting and Stacking ensembles represented by triangles, have highly correlated and low STD, meaning they are highly predictive with reliable performance. Other models, such as DT and GBM, have high correlation coefficients but slightly higher STD, hence being relatively accurate yet with more variance. This places the SVR and NuSVR models farther away from this observed point, indicating a weaker correlation and higher RMSE values to show poorer prediction accuracy. In general, what this diagram pinpoints is that the two ensemble models, Voting and Stacking, provide the best balance of correlation and minimization of error as the most robust choices for WS predictions on this dataset.

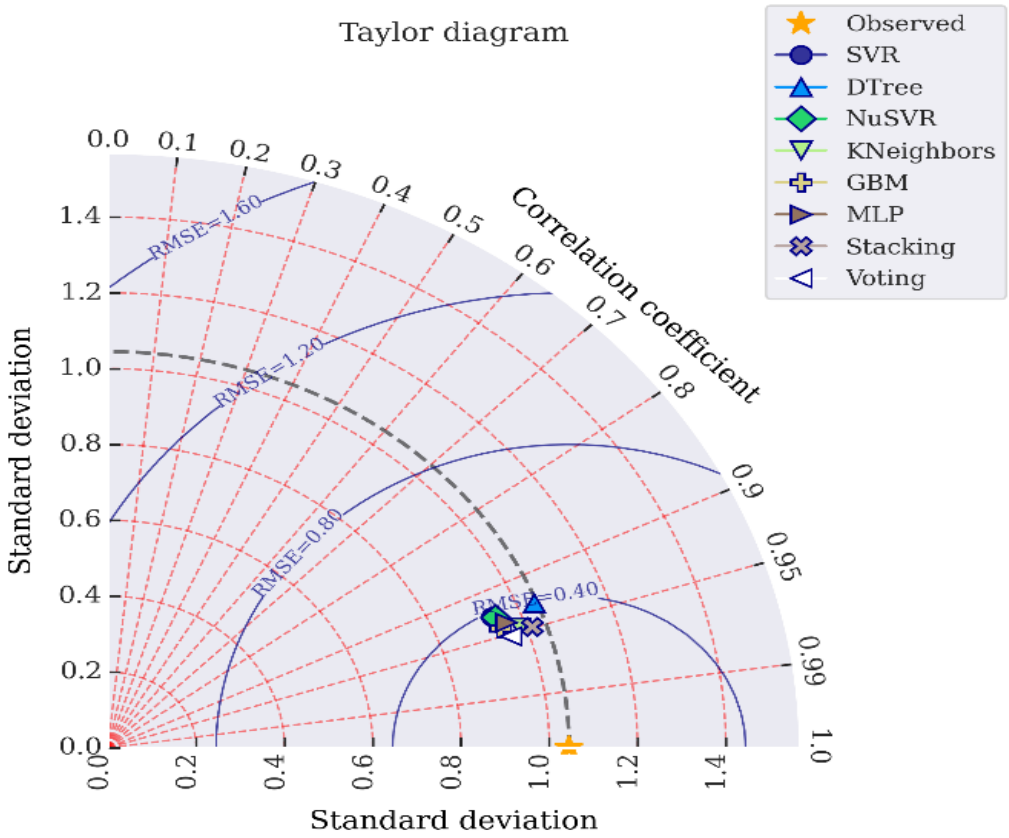

Figure 18: Taylor diagram of model performance in WS prediction

Table 1 presents the performance of two ensemble models, Stacking and Voting, based on a number of performance metrics, both for the testing and training sets. The results of the Stacking model outperform those obtained by the Voting model in the training set: the former has a lower MSE. These lower error values, in addition to the higher value of R^2, which is 0.951 compared with 0.931, suggesting that stacking fitting is more accurate for training data. Furthermore, lower NMSE and MDAPE for Stacking mean that the model is more accurate.

On the test set, both models behave similarly, with the Voting model only marginally outperforming Stacking on some metrics. This is in reaching the minimum MSE of 0.105 against 0.111 and the RMSE of 0.324 against 0.333. Also, the R^2 is marginally higher for the Voting model, 0.904, than for Stacking, 0.898, and so is VAF. This would then indicate that it generalizes slightly better to unseen data. However, both models have a similar consistency, with the STD for the two data sets spreading relatively comparably, as can be seen from the values that the STD has taken. Overall, Stacking generalizes slightly better on the training set data, while Voting creates a slight advantage in terms of performance on the test data and is, therefore, more reliable for generalization. From this, a clear idea of the strength of ensemble techniques in making robust and correct predictions has emerged, with Voting being slightly preferable in this context.

Table 1: Comparative performance analysis of stacking and voting ensemble models

Optimizer	Stacking	Voting
Train		
MSE	0.049	0.069
RMSE	0.221	0.263
MAE	0.130	0.176
R2	0.951	0.931
NMSE	0.049	0.069
MDAPE	0.838	1.478
STD _dev	0.968	0.918
VAF	0.951	0.931
Test		
MSE	0.111	0.105
RMSE	0.333	0.324
MAE	0.206	0.219

R2	0.898	0.904
NMSE	0.102	0.096
MDAPE	1.356	1.778
STD _dev	1.012	0.959
VAF	0.898	0.904

Table 2 compares the error metrics for various ML models, such as DT, GBM, KNN, MLP, Nu SVR, and SVR, in both training and test datasets. The main performance metrics like MSE, RMSE, MAE, and R^2 all suggest that KNN is doing the best on the Test set generally, with the lowest MSE and RMSE along with the highest R^2, which can indicate better generalization on previously unseen data. DT, by contrast, has extremely low errors on the training set, which may indicate overfitting, as its test performance falls comparatively. On the whole, the KNN and GBM models have a very good balance regarding stability and accuracy for both testing and training; hence, these models can be used to make reliable predictions on new data.

Table 2: Error metrics derived from the application of hybrid models

Optimizer	DT	GBM	KNN	MLP	Nu SVR	SVR
	Train					
MSE	1.97E-33	0.109	0.072	0.124	0.136	0.139
RMSE	4.44E-17	0.331	0.269	0.352	0.369	0.373
MAE	2.22E-18	0.242	0.159	0.252	0.242	0.237
R2	0.999	0.890	0.928	0.875	0.863	0.860
NMSE	1.98E-33	0.110	0.072	0.125	0.137	0.140
MDAPE	0	2.286	0.839	2.440	1.977	1.738
STD _dev	0.998	0.906	0.950	0.912	0.899	0.893
VAF	0.999	0.890	0.928	0.875	0.864	0.861
Test						
MSE	0.151	0.123	0.114	0.130	0.148	0.150
RMSE	0.389	0.351	0.338	0.360	0.384	0.388
MAE	0.216	0.257	0.205	0.261	0.259	0.255
R2	0.861	0.887	0.895	0.881	0.865	0.862
NMSE	0.139	0.113	0.105	0.119	0.135	0.138
MDAPE	0.9501	2.365	1.156	2.485	2.168	1.968
STD _dev	1.038	0.943	0.986	0.961	0.943	0.937
VAF	0.861	0.887	0.895	0.882	0.866	0.863

Fig. 19 is Actual vs. predicted scatter plots for Voting and Stacking ensemble model predictions. Markers distinguish between training and testing points. Both the models show a high correspondence with the Y=X line, suggesting that the models have resulted in good predictions. Similarly, the R^2 value of the Voting model is 0.931 for the training data and 0.904 for the test data, indicating very good predictive accuracy and generalization. On the other hand, the Stacking model is relatively a bit higher at 0.951 for the training data, which means that it has fitted better for the training data, whereas the generalization has faced a small drop, as evidenced by its R^2 of 0.898 in the test data. This pattern would hint that while Stacking may capture more detail within the training data, and Voting maintains slightly better consistency across both training and testing datasets. Overall, both ensemble methods yield reliable predictions. Still, voting is slightly more robust and, thus, perhaps better suited for situations where the generalization of new data is paramount.

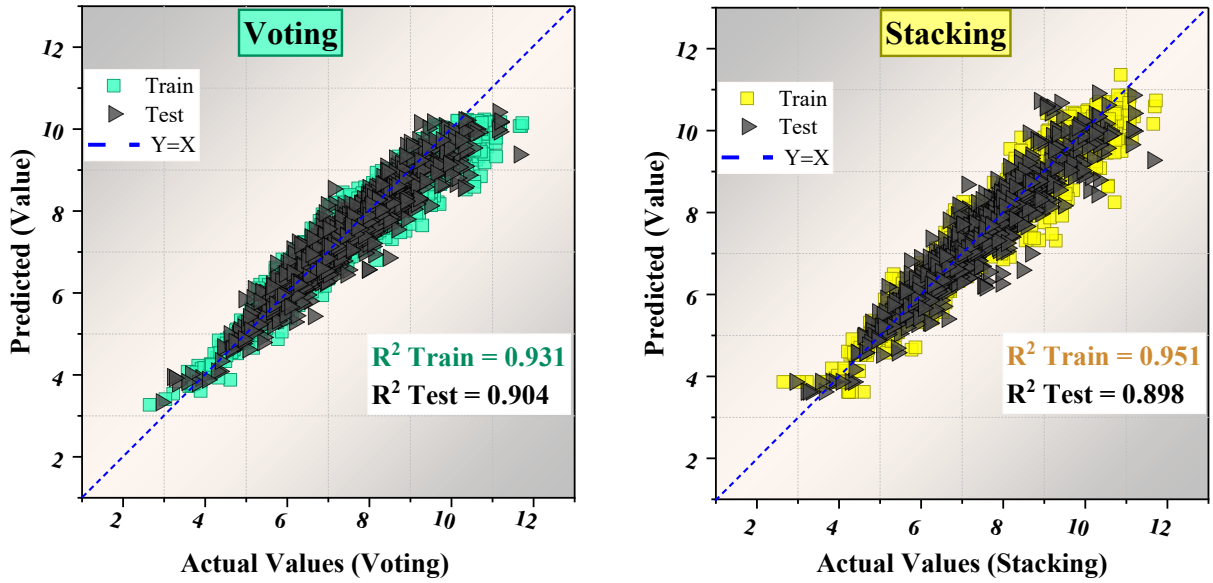

Figure 19: Comparing projected and actual values for voting and stacking ensemble models

The error distribution of the ensemble models for voting and stacking is displayed in Fig. 20. The errors are concentrated around zero, indicating the quite accurate predictions made by both models. The green color bars represent the error distribution for the Voting model, and the yellow color bars represent the error distribution for the Stacking model, which have a similar central peak, reflecting that most of the predictions are quite close to the actual values. This tells us that both models tend to keep their prediction errors low without strong bias, and the spread of both curves is quite similar. The model has overall error distributions being much larger or much smaller in comparison to the competing model. However, the Voting model seems to be somewhat more centered or somewhat narrower in distribution compared to the Stacking model's distribution, though it may indicate that Voting produces more consistent predictions with fewer extreme deviations. This would be a reason for the following conclusion: both ensemble methods yield high accuracy, although Voting has a minor advantage in terms of error consistency.

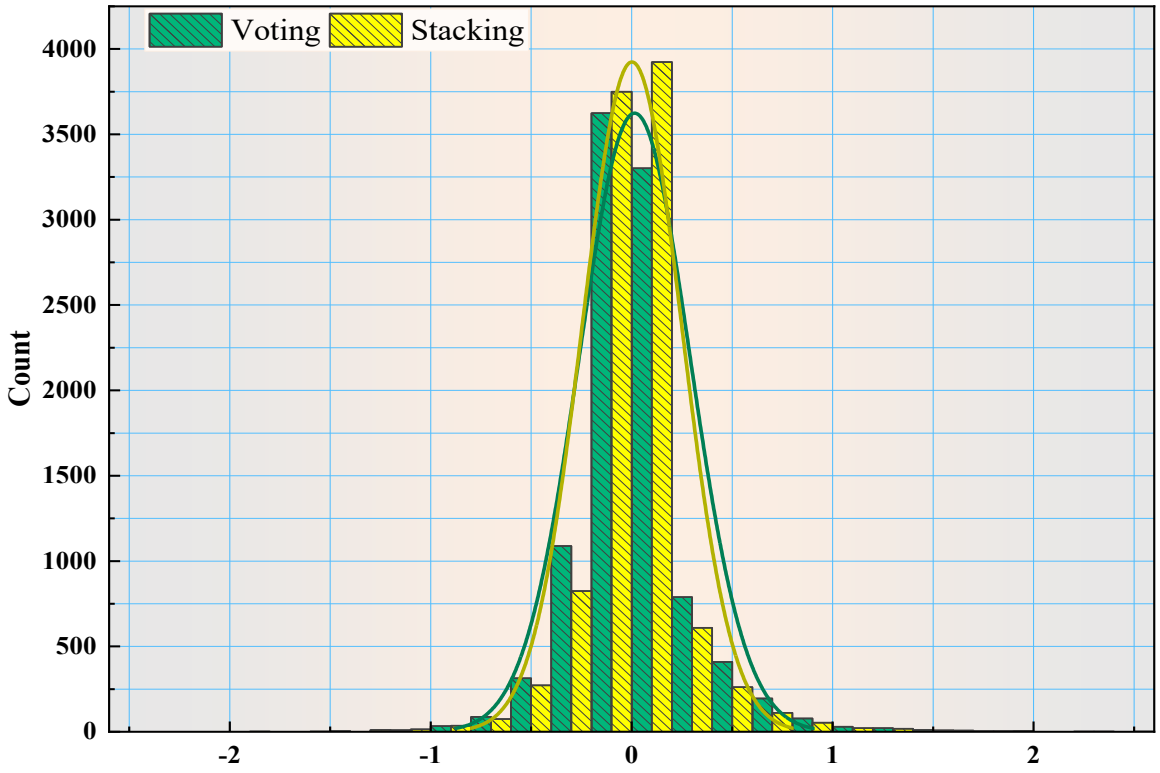

Figure 20: Error distribution for voting and stacking ensemble models

Fig. 21 is a SHAP analysis of the dependence of longitudinal features, latitude, capacity, and capacity factor on WS forecast, with different strengths. The most influencing features are the longitude and capacity factor. Longitude has a nonlinear but huge influence in some ranges of this variable, while the capacity factor has shown quite a clear positive linear relationship with WS predictions-higher predicted WSs at locations with higher capacity factors. Latitude holds variability in its influence from one region to another, whereas capacity is more scattered, which would argue for a less direct relationship with WS. This overall analysis suggests that longitude and capacity factors seem to be strong predictors of wind velocity, whereas latitude and capacity are more profoundly interconnected.

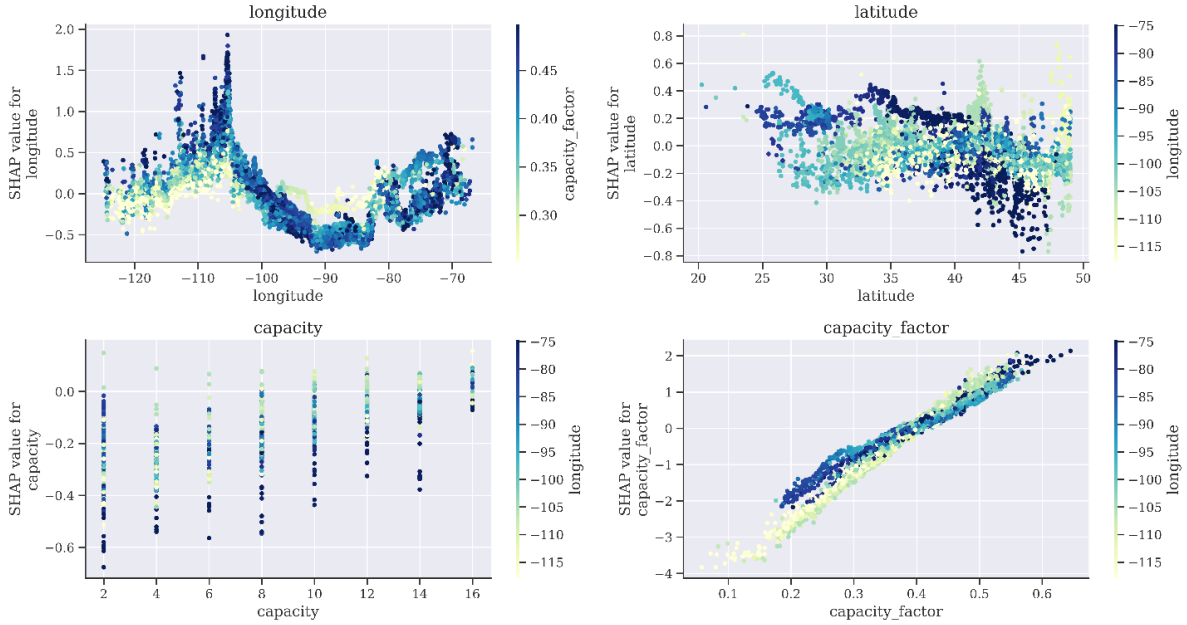

Figure 21: SHAP analysis of feature influence on WS prediction

Fig. 22 represents a summary plot by SHAP, ranking the influence of four features: capacity factor, longitude, latitude, and capacity towards the WS prediction model. The horizontal position of each point depicts the SHAP value, correlating with the impact that each feature had on the model output, where positive values push predictions higher and negative ones push them lower. The most informative feature is the capacity factor, which has a large spread of SHAP values and higher values on the right of the distribution, showing that larger magnitudes of the capacity factor indeed strongly raise the predicted WS. The second most impactful feature value is longitude, which reveals partly positive SHAP values at both the lower and upper bounds of its respective range, suggesting that some special values in longitude positively or negatively influence predictions based on geographical location. Latitude and capacity are relatively smaller in their effects, as indicated by centering their SHAP values around zero, hence their limited influence on the model output. Color gradient represents the feature's actual value: high values appear in dark blue, and low values appear in light yellow. This plot underlines the fact that capacity factor and longitude are the most influential drivers in the prediction of WS, while latitude and capacity are less part of the model's predictions.

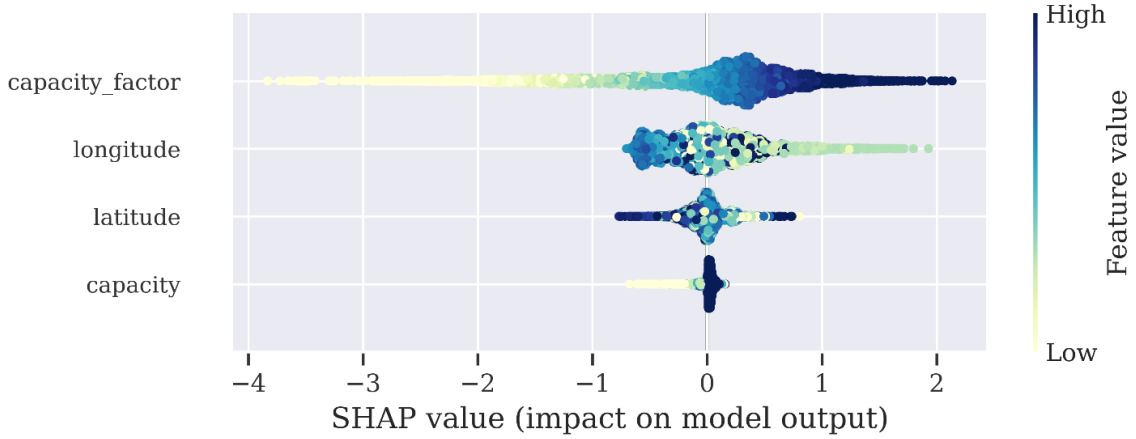

Figure 22: SHAP summary plot of feature impact on WS prediction model

This was a comparison analysis between ML models and ensemble techniques for WS prediction. According to the metrics of R^2, MSE, and RMSE, all showed that the ensembles-the Voting and Stacking methods-are the most robust and accurate predictions. The aforementioned methods work best in capturing the complex patterns in wind data since there is a combination of many models and better generalization on unseen test data. Voting slightly outperformed others for test data, balancing the accuracy with the stability of the results. Besides, SHAP analysis has shown that capacity factor and geographic location longitude are the strongest predictors of WS; therefore, both environmental factors and turbine-specific parameters are critical in determining the output of wind energy. Overall, these results suggest that the application of ensemble techniques with the selected features, such as capacity factor and longitude, will significantly improve the accuracy of the WS forecast, consequently allowing for better and more effective renewable energy prediction. This information will be of paramount importance in further work on wind power integration into the grid for a more sustainable energy infrastructure.

5 Conclusion

It shows the precision of neural network-based and ML models, particularly ensemble methods, in forecasting WS for data from the NREL Wind Integration National Dataset. The research underlines that new, sophisticated ensemble approaches like Voting and Stacking can enable more reliable and precise WS forecasting than basic, classic, stand-alone models. In these advanced ensemble methods, the strengths of multiple algorithms are combined to achieve improved predictive performance. According to the performance metrics of the model-MSE, RMSE, and the R^2, it was observed that the ensemble models capture non-linear relationships in wind data and generalize better to new, unseen data. The SHAP analysis also highlighted the most informative features for the prediction of WS: capacity factor and geographic location (longitude) and, therefore, pointed to the importance of those variables in renewable energy forecasting. This is a very important insight that can be used to optimize the production of wind energy with a view to integrating renewable energy into the grid in such a way as to ensure operational efficiency and grid stability. These findings suggest that a combined approach using both neural networks and ensemble techniques provides one robust solution to handle the intrinsic variability of wind patterns, hence contributing toward sustainable energy management and planning. Future studies will be focused on more sophisticated DL models like RNN and LSTM, which are architecture designs for temporal dependencies in sequences that could yield even better forecasting performance of WS variability across horizons. Besides, the inclusion of more diversified meteorological and atmospheric variables can lead to further refinements of the models to help check problems associated with the dynamic and intermittent properties of wind. This, therefore, forms the basis of continued improvement in renewable energy prediction analytics, improving the development of resilient and sustainable energy systems.

Authorship contribution statement

Tao Zou: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.
Xinwei Xie: Methodology, Software

Data availability

On Request

Author statement

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

Ethical approval

All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

References

[1] M. L. Hossain, S. M. N. Shams, and S. M. Ullah, “Comparative study of machine learning algorithms for wind speed prediction in Dhaka, Bangladesh,” Sustainable Energy Research, 11(1):23, 2024. https://doi.org/10.1186/s40807-024-00109-z

[2] C. Draxl, A. Clifton, B.-M. Hodge, and J. McCaa, “The wind integration national dataset (wind) toolkit,” Appl Energy, 151, 355–366, 2015. https://doi.org/10.1016/j.apenergy.2015.03.121

[3] Y. Ren, P. N. Suganthan, and N. Srikanth, “Ensemble methods for wind and solar power forecasting—A state-of-the-art review,” Renewable and Sustainable Energy Reviews, 50, 82–91, 2015. https://doi.org/10.1016/j.rser.2015.04.081

[4] Y. Hao and C. Tian, “A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting,” Appl Energy, 238, 368–383, 2019. https://doi.org/10.1016/j.apenergy.2019.01.063

[5] M. Liu, Z. Cao, J. Zhang, L. Wang, C. Huang, and X. Luo, “Short-term wind speed forecasting based on the Jaya-SVM model,” International Journal of Electrical Power & Energy Systems, 121, 106056, 2020. https://doi.org/10.1016/j.ijepes.2020.106056

[6] G. Maclaurin, C. Draxl, B.-M. Hodge, and M. Rossol, “Wind Integration National Dataset (WIND) Toolkit,” DOE Open Energy Data Initiative (OEDI); National Renewable Energy Laboratory …, 2014. https://doi.org/10.1016/j.apenergy.2015.03.121

[7] T. Blanchard and B. Samanta, “Wind speed forecasting using neural networks,” Wind Engineering44(1): 33–48, 2020. https://doi.org/10.1177/0309524X19849846

[8] J. Nielson, K. Bhaganagar, R. Meka, and A. Alaeddini, “Using atmospheric inputs for Artificial

Neural Networks to improve wind turbine power prediction," Energy, 190, 116273, 2020. https://doi.org/10.1016/j.energy.2019.116273

[9] M. M. Rahman et al., "A comprehensive study and performance analysis of deep neural network-based approaches in wind time-series forecasting," J Reliab Intell Environ, 9(2): 183–200, 2023. https://doi.org/10.1007/s40860-021-00166-x

[10] A. Dolatabadi, H. Abdeltawab, and Y. A.-R. I. Mohamed, "Hybrid deep learning-based model for wind speed forecasting based on DWPT and bidirectional LSTM network," IEEE Access, 8, 229219–229232, 2020.DOI: 10.1109/ACCESS.2020.3047077

[11] M. Lydia, G. Edwin Prem Kumar, and R. Akash, "Wind speed and wind power forecasting models," Energy & Environment, p. 0958305X241228515, 2024. https://doi.org/10.1177/0958305X241228515

[12] A. Routray, Y. S. Reddy, and S. Hur, "Predictive Control of a Wind Turbine Based on Neural Network-Based Wind Speed Estimation," Sustainability, 15(12): 9697, 2023. https://doi.org/10.3390/su15129697

[13] H. Malik, A. K. Yadav, F. P. G. Márquez, and J. M. Pinar-Pérez, "Novel application of Relief Algorithm in cascaded artificial neural network to predict wind speed for wind power resource assessment in India," Energy Strategy Reviews, 41, 100864, 2022. https://doi.org/10.1016/j.esr.2022.100864

[14] K. Wang, O. Gaidai, F. Wang, X. Xu, T. Zhang, and H. Deng, "Artificial neural network-based prediction of the extreme response of floating offshore wind turbines under operating conditions," J Mar Sci Eng, 11(9): 1807, 2023. https://doi.org/10.3390/jmse11091807

[15] I. Ahmad, F. M'zoughi, P. Aboutalebi, I. Garrido, and A. J. Garrido, "Fuzzy logic control of an artificial neural network-based floating offshore wind turbine model integrated with four oscillating water columns," Ocean Engineering, 269, 113578, 2023.https://doi.org/10.1016/j.oceaneng.2022.113578

[16] A. M. Assaf, H. Haron, H. N. Abdull Hamed, F. A. Ghaleb, S. N. Qasem, and A. M. Albarrak, "A review on neural network-based models for short term solar irradiance forecasting," Applied Sciences, 13(14): 8332, 2023. https://doi.org/10.3390/app13148332

[17] M. Jiang, J. Shen, X. Jiang, L. Luo, R. Zhou, and Q. Zhou, "A novel automatic wind power prediction framework based on multi-time scale and temporal attention mechanisms," arXiv preprint arXiv:2302.01222, 2023. https://doi.org/10.48550/arXiv.2302.01222

[18] H. Liu, H. Ma, and T. Hu, "Enhancing short-term wind speed forecasting using graph attention and frequency-enhanced mechanisms," arXiv preprint arXiv:2305.11526, 2023. https://doi.org/10.48550/arXiv.2305.11526

[19] H.-K. Wang, K. Song, and Y. Cheng, "A hybrid forecasting model based on CNN and informer for short-term wind power," Front Energy Res, 9, 788320, 2022. https://doi.org/10.3389/fenrg.2021.788320

[20] S. Huang, C. Yan, and Y. Qu, "Deep learning model-transformer based wind power forecasting approach," Front Energy Res, 10, 1055683, 2023. https://doi.org/10.3389/fenrg.2022.1055683

[21] M. M. Christian, Y. S. Kim, H. Choi, J. Lee, and S. You, "Enhancing Wind Speed and Wind Power Forecasting Using Shape-Wise Feature Engineering: A Novel Approach for Improved Accuracy and Robustness," arXiv preprint arXiv:2401.08233, 2024.https://doi.org/10.17703/IJACT.2023.11.4.393

[22] J. Keisler and E. Le Naour, "WindDragon: Enhancing wind power forecasting with Automated Deep Learning," arXiv preprint arXiv:2402.14385, 2024.https://doi.org/10.48550/arXiv.2402.14385