JOURNAL OF RENEWABLE ENERGY AND SMART GRID TECHNOLOGY

Vol. 20, No. 1, January-June 2025

Selecting the Best Regression Model for Wind Power Prediction and Management for VPP

Subhajit Roy*, Devanshu Singh, Nikita Sinha, Dulal Chandra Das, Nidul Sinha

National Institute of Technology, Silchar, Assam 788010, India

***Corresponding author's email**: subhajitroy111@gmail.com

Article info:

Received: 21 February 2025
Revised: 23 April 2025
Accepted: 13 May 2025
DOI:
10.69650/rast.2024.260878

Keywords:

Machine Learning
Virtual Power Plant
Wind Energy
Lasso Regression
ADA Boost Regressor
Random Forest

ABSTRACT

Modern power systems increasingly rely on renewable energy, making effective prediction and management strategies essential, particularly for wind power, known for its variability and intermittency. This study delves into the application of machine learning models to predict wind power generation and optimize power management within virtual power plants (VPPs). It emphasizes key processes such as data preprocessing, feature engineering, and the use of advanced algorithms like Lasso Regression, Support Vector Machine (SVM) Regression, Adaptive Boosting (ADA Boost), and Random Forest Regression. The analysis focuses on critical meteorological and operational factors—wind speed, Low Voltage (LV) Active Power, and wind direction—that significantly impact wind energy output. The study addresses common data challenges, including missing values and feature scaling, to enhance model accuracy and reliability. By developing predictive models, the research enables efficient resource allocation, dynamic energy dispatch, and robust management strategies for VPPs. Through machine learning, the study proposes innovative solutions to improve grid stability, enhance renewable energy utilization, and promote sustainable energy systems. These insights pave the way for resilient and efficient integration of wind energy into modern power infrastructures.

1. Introduction

The shift toward sustainable energy on a global scale has highlighted the importance of renewable energy, especially wind power, in addressing climate change and decreasing dependence on fossil fuels [1-2]. Countries around the world are implementing renewable energy strategies to achieve environmental targets and strengthen energy security, with wind energy playing a pivotal role in these efforts [3]. Wind energy's widespread availability and scalability make it a promising choice for large-scale energy production. However, its unpredictable and intermittent nature creates challenges for integrating it into power grids, requiring advanced methods to improve its reliability and predictability [4].

VPPs have emerged as a groundbreaking approach for effectively managing renewable energy sources such as wind and solar power [5-6]. By combining distributed energy resources (DERs)—such as wind turbines, solar installations, and energy storage systems—VPPs replicate the operations of traditional power plants while providing greater adaptability and operational efficiency [7]. This aggregation enables real-time coordination, optimized energy dispatch, and improved grid stability [8]. Furthermore, integrating microgrids and vehicle-to-grid (V2G) technologies within VPPs enhances energy flexibility by enabling bidirectional power flow, allowing electric vehicles and localized energy hubs to act as dynamic grid-balancing assets [9]. However, the operational success of VPPs heavily depends on accurate forecasting and dynamic management of renewable energy outputs. Wind energy systems play a crucial role in microgrids and VPPs by providing decentralized, scalable, and sustainable power solutions that enhance grid resilience and operational flexibility [10]. Wind energy, with its fluctuating nature influenced by meteorological and geographical factors, demands sophisticated predictive tools to maximize its potential within VPP frameworks [11].

Machine learning (ML) has become a transformative tool for tackling the complexities of wind energy forecasting and management. By analyzing extensive datasets that include weather conditions and operational data, ML algorithms can uncover intricate patterns, forecast energy generation, and refine energy distribution strategies [11]. Techniques such as Lasso Regression, SVM, ADA Boost, and Random Forest have proven effective in modeling non-linear interactions between factors like wind speed, direction, and power output [12-15]. These approaches deliver valuable insights that improve resource utilization, minimize energy waste, and strengthen grid stability. Research has widely investigated the use of ML in renewable energy forecasting. For example, Sun et al. [16] demonstrated the effectiveness of Random Forest in short-term wind power prediction, noting its capability to manage complex, non-linear datasets. Sierla et al. [17] examined how ML mitigates uncertainties in Virtual Power Plant (VPP) operations, while Qureshi et al. [18] explored the use of regression models to predict wind turbine performance. Other works, such as [19] Nooruldeen et al. [19] and Hussain and Zaidi [20], have emphasized the potential of ensemble learning techniques like ADA Boost in improving predictive accuracy. Together, these studies highlight the increasing significance of ML in renewable energy systems. Yet,

there is a lack of exploration into integrated frameworks that tackle data quality challenges, optimize algorithm choices, and incorporate real-time adaptability.

This study advances existing research by proposing a holistic approach to wind power prediction and management within VPPs using machine learning. Critical meteorological variables, including wind speed, wind direction, and LV Active Power, are examined to develop predictive models that enhance operational efficiency and sustainability [21]. The research also emphasizes the role of effective data preprocessing methods, such as managing missing data and feature scaling, to improve model accuracy. By incorporating multiple ML algorithms, the study enables a comparative evaluation, offering insights into the advantages and limitations of each method for wind power forecasting. The paper is organized as follows. Section 2 provides a review of relevant literature, establishing the research context and identifying gaps in current methodologies. Section 3 describes the research methodology, including data collection, preprocessing, and model training procedures. Section 4 presents the experimental findings, comparing the performance of different ML algorithms and discussing their relevance to VPP operations. Finally, Section 5 concludes with a summary of key findings, actionable recommendations, and potential avenues for future research.

2. Research Gaps and Objectives

2.1 Machine Learning for Wind Power Prediction

Wind power forecasting has become increasingly vital as the integration of renewable energy grows in importance. ML has established itself as a key solution in this field, with techniques such as Random Forest, ADA Boost, Support Vector Regression (SVR), and Lasso Regression demonstrating their effectiveness in managing the intricate challenges of wind energy prediction. For instance, Sun et al. [16] demonstrated the utility of Random Forest in capturing nonlinear relationships between meteorological parameters and wind energy output, showcasing its robustness for short-term forecasting. Similarly, Buturache and Stancu [22] highlighted the capability of ensemble learning techniques, such as ADA Boost, to handle dynamic data while minimizing bias and variance. Shahid et al. [23] emphasized that VPPs, powered by ML models, can effectively manage distributed energy resources, enhancing grid reliability. However, these models are susceptible to issues like overfitting and computational inefficiency. Studies by [23] Aldhafferi et al. [24] and Ghimire et al. [25] revealed the potential of SVR in providing accurate predictions by utilizing kernel functions to model complex data patterns. Despite its high accuracy, SVR's computational requirements remain a concern for real-time applications.

2.2 Role of Meteorological Factors

Meteorological factors, including wind speed, direction, and air density, are critical in predicting wind power. The relationship between these factors and energy output is often nonlinear, requiring advanced ML algorithms for effective modeling. Eniola [26] investigated the impact of wind variability on turbine efficiency, concluding that machine learning techniques are indispensable for understanding and optimizing energy generation. Moreover, Bochenek et al. [27] highlighted the role of ML in predicting uncertainty caused by abrupt meteorological changes, enabling VPPs to manage fluctuations effectively.

2.3 Data Preprocessing for Model Performance

The performance of machine learning (ML) models heavily depends on the quality of input data. Research by Elouataoui et al. [28] and Budach et al. [29] has highlighted the significance of resolving data quality challenges, such as handling missing values and outliers, to improve the accuracy and dependability of these models. Techniques like feature scaling and imputation are widely used to prepare data for model training, ensuring improved accuracy and robustness.

Prior studies have extensively explored data quality challenges and algorithm selection for renewable energy prediction, yet critical gaps persist. While preprocessing techniques like outlier removal (Sarathkumar et al., 2025)[33] and imputation [34] are well-documented, few frameworks integrate real-time data to enhance Virtual Power Plant (VPP) adaptability under dynamic conditions [22]. Abrupt meteorological uncertainties—such as sudden wind speed shifts—remain underexplored, limiting VPP resilience [40]. Comparative analyses of machine learning (ML) algorithms are scarce, with most works focusing on isolated models [35]. To bridge these gaps, this study proposes a unified framework combining robust preprocessing with real-time meteorological data integration for adaptive VPP management. We systematically evaluate four ML models (Lasso, SVR, AdaBoost, Random Forest), demonstrating Random Forest's superiority in handling non-linearity and variance [36-37]. Additionally, novel operational strategies are introduced to mitigate grid instability during abrupt wind fluctuations, advancing beyond conventional approaches.

2.4 Research Gaps

- Existing studies lack frameworks that unify advanced preprocessing (like, kNN imputation, IQR outlier filtering) with real-time meteorological data integration, limiting adaptability in dynamic VPP environments.
- Prior works focus on isolated ML models for wind power prediction without benchmarking their scalability, computational efficiency, or suitability for VPP-specific tasks like energy trading and grid balancing.
- While prediction accuracy is studied, the direct influence of ML models on VPP resource allocation, energy dispatch efficiency, and grid stability during abrupt wind fluctuations remains underexplored.
- Current approaches fail to address sudden meteorological uncertainties (like, rapid wind speed shifts) through adaptive ML frameworks, reducing VPP resilience in real-world scenarios.

2.5 Research Objectives

Based on the identified research gaps, the following objectives are formulated to describe the work best:

- Objective 1: Design and implement a unified pipeline integrating kNN imputation, feature scaling, and outlier filtering to improve data quality while enabling real-time adaptability for VPP applications.
- Objective 2: Assess the performance of Lasso Regression, SVR, AdaBoost, and Random Forest not only for accuracy but also computational efficiency, interpretability, and scalability in VPP energy trading and dispatch scenarios.
- Objective 3: Demonstrate how ML-driven wind power forecasts optimize energy allocation, reduce grid instability, and

enhance profitability in VPP energy markets through case studies and scenario analysis.

- Objective 4: Introduce hyperparameter-tuned ensemble models (like, Grid/Randomized Search CV-optimized Random Forest) to mitigate abrupt meteorological uncertainties, ensuring reliable VPP performance under volatile conditions.

3. Analysis of Effects of Wind Parameters on Power Generation

The efficiency and output of wind power generation are primarily governed by key meteorological and operational parameters. Understanding the complex relationships between these parameters and their effects on power generation is essential for optimizing wind turbine performance and ensuring reliable energy output. This section explores the influence of critical wind parameters—wind speed, wind direction, and rotor swept area—on the power generation capabilities of wind turbines.

3.1 Wind Speed: The Dominant Factor

Wind speed is the most critical factor affecting wind power generation, as the energy produced by a wind turbine is proportional to the cube of the wind speed $P \propto V^3$. Even small changes in wind speed can cause significant fluctuations in power output, underscoring the need to operate turbines within their optimal speed range.

Wind turbines are designed with specific operational thresholds:

- Cut-in Speed: The minimum wind speed required for the turbine to start generating power.
- Rated Speed: The wind speed at which the turbine produces its maximum power output.
- Cut-out Speed: The maximum wind speed beyond which the turbine shuts down to prevent damage.

The formula for wind power generation is: [30]

$$P = \frac{1}{2}\rho A V^3 \eta \quad (1)$$

Where:

P: Power Output (Watts)

ρ: Air Density (kg/m^3)

A: Rotor Swept Area (m^2)

V: Wind Speed (m/s)

η: Turbine Efficiency

Data analysis reveals a strong correlation between wind speed and LV Active Power, with the highest power outputs observed during periods of steady, moderate-to-high wind speeds. The nonlinear relationship underscores the importance of accurate wind speed forecasting for reliable energy predictions.

3.2 Wind Direction to Ensure Optimal Turbine Alignment

Wind direction significantly impacts the efficiency of wind turbines. Turbines are most effective when perfectly aligned with the prevailing wind direction. Misalignment reduces power output, as described by: [31]

$$P_{active} = P_{max} \cos^2(\theta) \quad (2)$$

Where:

P_{active}: Actual Power Output

P_{max}: Possible Power Output

θ: Yaw Angle

Seasonal and regional variations in wind direction necessitate advanced turbine designs with rapid yaw adjustment mechanisms to maintain alignment. Analysis of wind direction data shows periods of stable directional patterns yield higher energy outputs, while erratic shifts result in reduced efficiency due to frequent turbine realignment.

3.3 Rotor Swept Area

The rotor swept area $(A = \pi r^2)$, determined by the radius of the turbine blades, directly influences the amount of kinetic energy captured from the wind. Larger rotor areas capture more wind energy, translating to higher power generation. However, increasing the rotor size requires balancing structural and economic considerations. Comparative analyses reveal that turbines with optimized rotor designs outperform those with smaller areas, particularly in regions with consistent wind speeds. This highlights the need for site-specific turbine selection to maximize efficiency.

3.4 Air Density and Environmental Conditions

Air density (ρ), influenced by altitude, temperature, and humidity, also affects wind power generation. Higher air density results in greater energy capture, while lower density reduces efficiency. Although the effects of air density are less pronounced compared to wind speed and direction, they are crucial in determining optimal turbine locations and operational strategies.

3.5 Combined Parameter Effects

The interplay between these parameters is complex, requiring advanced modelling techniques to predict their combined effects on power generation. Machine learning algorithms, including Random Forest and ADA Boost, are highly effective at modelling these nonlinear relationships, offering valuable insights to enhance turbine efficiency and performance.

4. Comparison of ML Models for Wind Power Prediction

Machine learning algorithms play a vital role in wind power prediction, offering diverse approaches to handle the complexity of meteorological and operational data. This section compares four widely used algorithms: Lasso Regression, SVR, ADA Boost Regressor, and Random Forest Regressor, focusing on their key characteristics and applicability.

4.1 Lasso Regression

Fig. 1 illustrates the graphical representation of the Lasso Regressor, providing insight into the model's functionality. Lasso Regression [39] (Least Absolute Shrinkage and Selection Operator) is a modified version of Linear Regression that improves model accuracy by incorporating L1 regularization. This technique is particularly useful for feature selection and handling sparse datasets, as it simplifies models by penalizing less important features. Similar to standard Linear Regression, Lasso models relationships between variables using a linear approach. However, it introduces an L1 penalty term, which forces the regression coefficients of less important features toward zero, effectively

eliminating them. This automatic feature selection reduces overfitting and improves model interpretability. Lasso is especially valuable in applications such as wind power prediction, where identifying the most influential factors is crucial.

Here in Fig.1 shows blue ellipses of constant sum-of-squared error (SSE) centered at the ordinary least squares estimate $\hat{\beta} = (2.0, 1.5)$, with β_1 on the horizontal axis (wind-speed coefficient) and β_2 on the vertical axis (wind-power coefficient). Superimposed is the red diamond representing the L1-norm constraint $|\beta_1|+|\beta_2|=t$, and the black square marks the Lasso solution—where the smallest SSE contour just touches the boundary of this "L1 ball." This visualization neatly demonstrates how the Lasso penalty shrinks both coefficients toward zero, sacrificing a small amount of fit (moving to a slightly larger contour) in exchange for a sparser, more robust model.

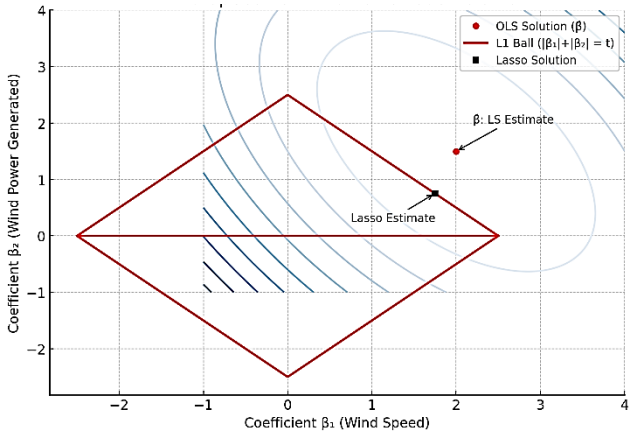


Fig. 1 Graphical representation of Lasso Regressor.

The objective function for Lasso Regression is given by:

$$\min\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right) \quad (3)$$

here:

y_i: Represents the actual observed values.

$\hat{y}_i$ = is the predicted value from the regression model.

n: is the total number of observations (data points).

p: represents the number of features (independent variables).

β_j: the regression coefficients corresponding to each feature.

λ the regularization parameter that controls the strength of the penalty applied to the model.

Here, the first term, $\frac{1}{2n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ represents the Mean Squared Error (MSE), which measures the model's accuracy in predicting outcomes.

The second term, $\lambda\sum_{j=1}^{p}|\beta_j|$, is the L1 regularization penalty, which adds the absolute values of the regression coefficients to the cost function.

The parameter λ determines the extent of regularization:

- When λ=0 Lasso behaves like standard Linear Regression (no penalty applied).
- As λ increases, more coefficients are pushed toward zero, reducing the number of selected features.
- A very high λ may overly simplify the model, leading to underfitting.

By striking a balance between model complexity and predictive accuracy, Lasso Regression successfully mitigates overfitting while preserving interpretability, establishing it as an essential technique in contemporary machine learning and data science practices.

4.2 Support Vector Regression (SVR)

SVR is an adaptation of SVM tailored for regression problems [40]. Unlike conventional regression methods that aim to reduce residual error, SVR aims to fit the model within a specified tolerance margin, referred to as the epsilon (ϵ)-insensitive zone. This approach makes SVR highly suitable for managing noisy data and intricate relationships, as it does not penalize small deviations from actual values. Fig. 2 depicts the graphical representation of the Support Vector Regressor, offering an understanding of the model's operational mechanism.

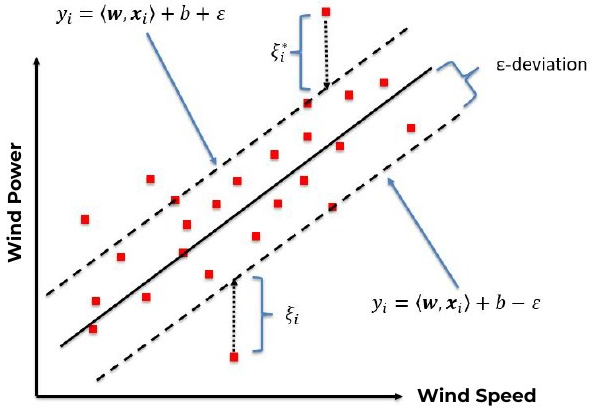


Fig. 2 Graphical representation of Support Vector Regression.

Mathematical Formulation of SVR

The objective function for SVR is given by:

$$\min\left(\frac{1}{2}\sum_{j=1}^{p}w_j^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*)\right) \quad (4)$$

subject to the constraints:

$$y_i-(w^Tx_i+b)\leq\epsilon+\xi_i \quad (5)$$

$$(w^Tx_i+b)-y_i\leq\epsilon+\xi_i^* \quad (6)$$

$$\xi_i,\xi_i^*\geq 0$$

where:

y_i: represents the actual observed values.

x_i: is the feature vector for the i – th observation

w: weight vector determining the model's decision boundary.

b: bias term.

C: hyperparameter that controls the trade – off between margin width and prediction error.

ϵ: epsilon incentive loss, meaning predictions within this range are not penalized.

ξ_i and ξ_i^*: slack variables, allowing flexibility for outliers when the data is not perfectly contained within the ϵ-tube.

Loss Function Components: The first term, $\frac{1}{2}\sum_{j=1}^{p} w_j^2$, represents the regularization term, ensuring that the model does not become too complex.

The second term, $C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$, controls how much the model allows deviations beyond the epsilon margin. A higher C results in a tighter fit, while a lower C allows a more flexible model.

Epsilon-Insensitive Zone: The model ignores errors within a band of width ϵ around the predicted values, meaning minor deviations are not penalized.

- If a data point lies inside this margin, the loss is zero.
- If it lies outside, the slack variables (ξ_i , ξ_i^*) measure the degree of deviation, and a penalty is applied.

Kernel Trick in SVR: SVR can be adapted to address non-linear relationships through the use of kernel functions, such as polynomial or radial basis function (RBF), which map the input data into a higher-dimensional feature space.

Support Vector Regression offers several advantages, including robustness to outliers due to the epsilon-insensitive loss, effectiveness in handling high-dimensional and non-linear datasets through kernel functions, and better generalization compared to traditional regression models by avoiding overfitting through complexity control. It is widely used in applications such as financial market prediction, energy forecasting (like, wind power prediction), and biomedical signal analysis, where precision and robustness are crucial.

4.3 ADA Boost Regressor

Adaptive Boosting (AdaBoost) Regressor is an ensemble learning approach that enhances prediction accuracy by integrating multiple weak learners into a robust predictive model [41]. Unlike conventional regression methods that depend on a single model, AdaBoost sequentially trains a set of weak regressors, modifying their weights according to prior errors to minimize bias and variance. This approach makes AdaBoost particularly effective for handling complex, non-linear relationships and improving predictive performance. Fig. 3 displays the graphical representation of the AdaBoost Regressor, illustrating the model's working principles.

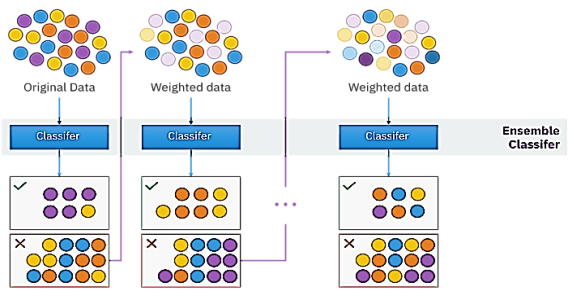


Fig. 3 Graphical representation of AdaBoost Regressor.

Mathematical Formulation of AdaBoost Regressor

The AdaBoost Regressor constructs the final model as a weighted sum of weak learners:

$$[F(x)=\sum_{m=1}^{M}\alpha_m h_m(x)] \quad (7)$$

where:

$F(x)$: the final prediction model.

M: The total number of weak learners.

h_m: represents the mmm-th weak regressor.

α_m: weight assigned to each weak learner based on its accuracy.

The weights α_m are computed as:

$$[\alpha_m = \frac{1}{2}\ln\left(\frac{1-e_m}{e_m}\right)] \quad (8)$$

where e_m is the weighted error of the weak learner, defined as:

$$[e_m = \frac{\sum_{i=1}^{n} w_i^{(m)}|y_i - h_m(x_i)|}{\sum_{i=1}^{n} w_i^{(m)}}] \quad (9)$$

Weighted Sum of Weak Learners: The final prediction $F(x)$ is obtained by aggregating multiple weak regressors, each weighted by α_m according to its performance.

Error-Driven Learning Process: Initially, all observations have equal weights. After each iteration, the model assigns higher weights to instances with large errors, forcing subsequent weak learners to focus on difficult cases.

Weight Computation (α_m):

- Weak learners with lower errors (e_m) receive higher weights, while those with higher errors contribute less to the final prediction.
- The logarithmic function ensures a balanced update of weights, preventing extreme fluctuations.

AdaBoost Regressor excels in reducing bias and variance by iteratively improving weak learners, making it highly effective for complex regression problems. It is robust to outliers since it assigns adaptive weights to handle difficult cases and works well with different base regressors, allowing flexibility in model selection. Additionally, it automatically emphasizes important patterns in the data, improving generalization. AdaBoost is widely applied in financial forecasting, energy demand prediction, medical data analysis, and anomaly detection, where high accuracy and adaptability are crucial.

4.4 Random Forest Regressor

Random Forest Regressor is an ensemble learning technique that improves prediction accuracy by aggregating the results of numerous decision trees [11]. While a single decision tree can be prone to high variance and overfitting, Random Forest addresses these challenges by averaging the predictions of multiple trees, enhancing both robustness and generalization. This approach involves training each tree on a random subset of the data and combining their outputs to generate the final prediction. Fig. 4 presents the graphical representation of the Random Forest Regressor, providing an overview of the model's functionality.

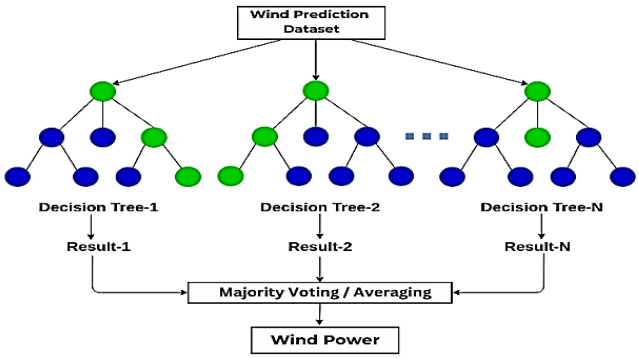


Fig. 4 Graphical representation of Random Forrest Regressor [11].

Mathematical Formulation of Random Forest Regressor

The final prediction of a Random Forest Regressor is given by:

$$[\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x)] \quad (10)$$

where:

$\hat{y}$: the final predicted value.

T: the total number of decision trees in the forest.

f_t: prediction from the t – th decision tree.

The sum of predictions is averaged to obtain a more stable and accurate result.

Error Measurement- MSE: The performance of a Random Forest Regressor is typically assessed using MSE, which is defined as:

$$[MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2] \quad (11)$$

where:

y_i: actual observed value.

$\hat{y}_i$: predicted value from the Random Forest.

n: total number of observations.

MSE measures the average squared difference between the actual and predicted values, offering a clear metric for evaluating the model's accuracy. A lower MSE signifies improved model performance.

Variance Estimation in Random Forest

Random Forest also reduces variance in predictions by aggregating multiple trees. The variance of the target variable can be expressed as:

$$Var(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (12)$$

where:

$Var(y)$ measures the variability in the dataset.

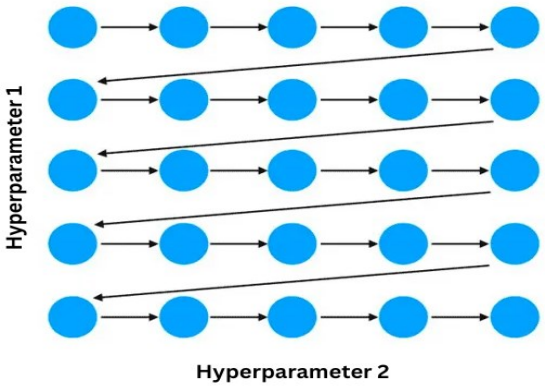
$\bar{y}$ represents the mean of the observed values.

By averaging multiple decision trees, Random Forest effectively reduces variance, leading to a more generalized and reliable predictive model.

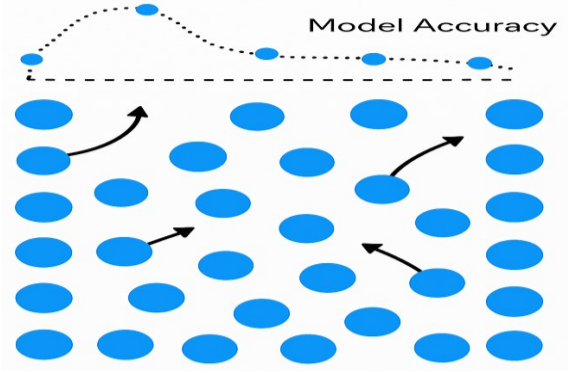
Random Forest Regressor provides several benefits, such as improved generalization through reduced overfitting, increased resilience to noise by averaging multiple models, and the ability to manage non-linear relationships without the need for explicit feature transformations. It is extensively utilized in fields like financial forecasting, energy load prediction, medical diagnostics, and climate modelling, where precision and reliability are essential.

4.5 Enhanced Random Forest Regressor models using Hyperparameter Tuning Strategies

Hyperparameter tuning strategies, such as Grid Search CV and Randomized Search CV as shown in Fig.5, systematically optimize the configuration of a Random Forest Regressor to achieve superior predictive accuracy and generalization compared to the default parameter settings. These methods address the limitations of the base Random Forest model—such as sensitivity to suboptimal hyperparameters and potential overfitting—by rigorously exploring combinations of parameters (like, tree depth, number of estimators, node-splitting criteria) to identify configurations that minimize error metric. By automating the search for optimal parameters, these strategies ensure robust performance across diverse datasets while retaining the inherent advantages of Random Forest, such as handling non-linear relationships and mitigating variance through ensemble averaging.



(a)



(b)

Fig. 5 (a) Grid Seach CV and (b) Randomized Search CV based RF Regressor Model Architecture.

4.5.1 Random Forest Regression (Grid Search CV)

Grid Search Cross-Validation (Grid Search CV) is a systematic hyperparameter tuning technique used to determine the optimal parameter configuration for a Random Forest Regressor [11]. By leveraging k-fold cross-validation, Grid Search CV mitigates overfitting and ensures robust generalization.

Mathematical Formulation of Random Forest Regressor

For a hyperparameter set $[\theta = \{\theta_1, \theta_2, \dots, \theta_m\}]$ the objective of Grid Search CV is to minimize the cross-validated MSE:

$$\theta^* = \arg\min_\theta \left(\frac{1}{k} \sum_{j=1}^{k} \mathrm{MSE}_j(\theta) \right)$$

$$\mathrm{MSE}_j(\theta) = \frac{1}{n_j} \sum_{i=1}^{n_j} \left(y_i - \hat{y_i}(\theta) \right)^2 \quad (13)$$

Here, θ^* represents the optimal hyperparameters, k denotes the number of cross-validation folds, n_j is the number of observations in the j-th validation fold, and $\hat{y_i}(\theta))$ corresponds to the prediction generated by the Random Forest model using hyperparameters θ.

Grid Search CV exhaustively evaluates all hyperparameter combinations within a predefined grid, ensuring the identification of the best-performing configuration. Although computationally demanding, this approach guarantees precision in parameter selection, leading to enhanced model accuracy and stability. Key advantages include the elimination of manual hyperparameter guesswork, improved model performance through parameter optimization, reduced overfitting via rigorous cross-validation, and suitability for smaller hyperparameter spaces due to its exhaustive search strategy.

4.5.2 Random Forest Regression (Randomized Search CV)

Randomized Search Cross-Validation (Randomized Search CV) is a computationally efficient hyperparameter optimization method that randomly samples parameter combinations from predefined distributions [11]. Unlike Grid Search CV, which evaluates all permutations, Randomized Search CV prioritizes resource efficiency, making it ideal for large hyperparameter spaces. The method evaluates a fixed number of randomly selected parameter sets using k-fold cross-validation, balancing exploration of diverse configurations with computational constraints.

Mathematical Formulation of Randomized Search CV For hyperparameters θ sampled from a distribution Θ, Randomized Search CV identifies:

$$\theta^* = \arg\min_{\theta \sim \Theta} \left(\frac{1}{k} \sum_{j=1}^{k} \mathrm{MSE}_j(\theta) \right) \quad (14)$$

where Θ defines probability distributions for each hyperparameter.

This method significantly reduces computational overhead while often achieving performance comparable to Grid Search CV, particularly in high-dimensional parameter spaces. By focusing on random sampling, it efficiently navigates complex hyperparameter landscapes and identifies robust configurations. Advantages include computational efficiency for large parameter spaces, compatibility with continuous and discrete parameter distributions, reduced risk of overfitting to specific hyperparameter sets, and practicality in resource-constrained scenarios.

4.6 Comparative Analysis

The table below offers a comparative analysis of Lasso Regression, SVR, AdaBoost Regressor, and Random Forest Regressor, assessing them on key attributes such as regularization, interpretability, robustness to outliers, complexity, and applicability to various use cases.

Table 1. Comparative analysis of regression models based on performance metric.

Feature	Lasso Regression	Support Vector Regression (SVR)	AdaBoost Regressor	Random Forest Regressor
Model Type	Linear regression with L1 regularization	Kernel-based regression	Ensemble learning (boosting)	Ensemble learning (bagging)
Regularization	L1 penalty (shrinks coefficients to zero)	Uses epsilon-insensitive loss	No explicit regularization	Implicit through averaging
Feature Selection	Yes (eliminates less important features)	No (all features are considered)	No (relies on weak learners)	No (all features are used)
Handling non-linearity	Poor (only linear relationships)	Excellent (via kernels)	Good (combines weak learners)	Good (uses multiple decision trees)
Robustness to Outliers	Moderate (due to L1 penalty)	High (controlled by epsilon)	High (adaptive weighting)	High (aggregates multiple models)
Overfitting Risk	Low (due to regularization)	Moderate (depends on kernel choice)	Low (focuses on difficult cases)	Low (reduces variance)
Interpretability	High (simple and easy to interpret)	Low (black-box model)	Moderate (weak learners are interpretable)	Moderate (many trees, less interpretable)
Computational Complexity	Low (fast training)	High (depends on kernel choice)	Moderate (iterative training)	High (many trees increase complexity)
Application Suitability	Sparse data, feature selection problems	High-dimensional and complex data	Regression with small data, boosting weak learners	Large datasets, robust predictions
Example Use Cases	Wind power prediction, econometrics	Financial forecasting, biomedical signals	Anomaly detection, demand prediction	Energy forecasting, climate modeling

So, Lasso Regression is ideal for feature selection and sparse datasets, offering interpretability but struggling with non-linearity. SVR excels in handling high-dimensional and complex data but can be computationally expensive. AdaBoost Regressor improves accuracy through adaptive boosting, focusing on difficult cases while being robust to noise. Random Forest Regressor provides high stability and robustness by averaging multiple decision trees, making it well-suited for large datasets. Each model has its strengths and is suited for different use cases, with the choice of model depending on factors like dataset size, presence of non-linearity, interpretability needs, and computational constraints.

5. Data Processing for Machine Learning Models in Wind Power Generation

This section outlines the processes for data collection, preprocessing, feature engineering, and dataset preparation to develop ML models for active wind power forecasting. Emphasis is placed on combining internal and external datasets, addressing data inconsistencies, and structuring the data to ensure precise and reliable predictions.

5.1 Data Set Preparation

To ensure reliable wind power forecasting, two datasets—internal and external—were combined to reflect both operational and environmental factors influencing turbine performance. Internal datasets included operational metrics directly obtained from wind turbine SCADA systems, while external datasets captured meteorological parameters affecting wind power output.

The internal dataset was derived from a freely available SCADA database, containing detailed information on turbine performance for a Nordex N117/3600 wind turbine located in the northwest of Turkey (coordinates: 40.58545° N, 28.99035° E). This data spanned one year (January 1–December 31, 2018) [38] and included key parameters such as:

- Wind speed (m/s): A primary determinant of energy production.
- Wind direction (°): Critical for turbine alignment.
- Theoretical power (kW): The ideal power output based on turbine specifications.
- Active power (kW): The actual power generated.

5.2 Data Cleaning and Handling Missing Values

The raw SCADA dataset initially contained 50,530 samples, of which approximately 2,030 records had missing values. The k-Nearest Neighbours (kNN) algorithm (k=5) was employed to impute these missing values by identifying similar patterns within the dataset. This approach preserved the integrity of the data and ensured consistent input for ML models.

Outliers were managed using the interquartile range (IQR) method. Data points with wind speeds below 3.5 m/s or above 25.5 m/s, as well as instances where active power was zero despite adequate wind speed, were removed since they did not conform to the turbine's operational power curve. Majorly, wind turbines begin generating power at a cut-in speed of around 3.5 m/s and shut down at about 25 m/s to prevent damage, especially in offshore installations [42]. Furthermore, entries with negative active power values were discarded. In total, 68 outliers were removed, ensuring a refined and dependable dataset for subsequent analysis.

5.3 Final Dataset Implementation for Machine Learning Models

Following cleaning and preprocessing, the dataset was divided into training (80%) and testing (20%) subsets to evaluate the ML models. Feature importance analysis, as highlighted by Karaman (2023), clearly identifies wind speed as the most significant meteorological parameter influencing wind-based energy generation, with other factors also demonstrating a measurable impact on wind speed. In the proposed research work rigorous data processing workflow integrates internal and external datasets, addressing missing values, outliers, and scaling challenges. By leveraging feature engineering and advanced preprocessing techniques, the study ensures high-quality input for ML models. The resulting framework enables accurate wind power prediction, laying the groundwork for optimized energy management in VPPs.

6. Feature Engineering and Correlation Analysis

Feature engineering was employed to determine the most influential factors affecting wind power generation. A correlation matrix was utilized to measure the relationships between independent variables (features) and the dependent variable (active power). The impact of various factors on active power generation was analysed, as detailed below.

6.1 Variation of LV Active Power as per the Dataset

The LV Active Power (kW) in the wind prediction dataset exhibits significant variations due to factors like wind speed, wind direction, and environmental conditions [32].

Fig. 5 illustrates the variation in LV Active Power (kW) across six randomly selected days, highlighting the impact of fluctuating wind conditions on power generation. On April 2, 2018, power output remains relatively stable, suggesting consistent wind speeds supporting efficient turbine operation. In contrast, May 12 and June 16, 2018, exhibit significant drops and irregular variations, indicating periods of low wind speeds or turbine downtime. November 10 and December 21, 2018, show sustained high power output, reflecting strong and stable wind conditions that enable turbines to operate near rated capacity. However, January 5, 2018, presents erratic fluctuations, suggesting intermittent wind availability and system adjustments affecting energy generation. These fluctuations highlight the non-linear relationship between wind speed and power output, underscoring the need for advanced forecasting methods to enhance energy dispatch efficiency and turbine performance. Understanding these fluctuations helps in improving grid stability and ensuring reliable wind power integration into the energy system.

6.2 Variation of Wind direction as per the Dataset

The provided Fig. 6 show the variation in wind direction across six different dates, highlighting its significance in wind power generation. Wind direction plays a critical role in determining the efficiency of turbines, as they need to align with the wind to capture maximum energy. On days like April 2, 2018, where the wind direction appears relatively stable, turbines can maintain alignment, ensuring consistent and efficient power generation. In contrast, days such as June 16, 2018, exhibit rapid and wide fluctuations in wind direction, spanning nearly 360°. Such variability can lead to reduced efficiency, as turbines require time to realign with the changing wind, resulting in intermittent or lower power output. On days with erratic directional changes, like May 12 and June 16, turbines face frequent misalignment, decreasing energy capture even if wind speeds are optimal. The stability of wind direction

directly impacts the capacity factor of wind farms, with less variability translating to more consistent power generation. To optimize power generation, wind farms should prioritize locations with minimal wind direction variability, as seen in April 2. Additionally, advanced turbines with rapid yaw adjustment systems can help mitigate losses during periods of fluctuating wind directions.

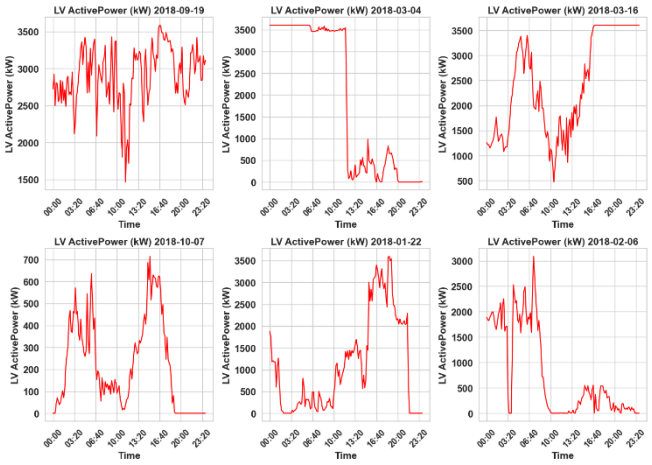


Fig. 6 Six-day variation of LV Active Power under fluctuating wind conditions.

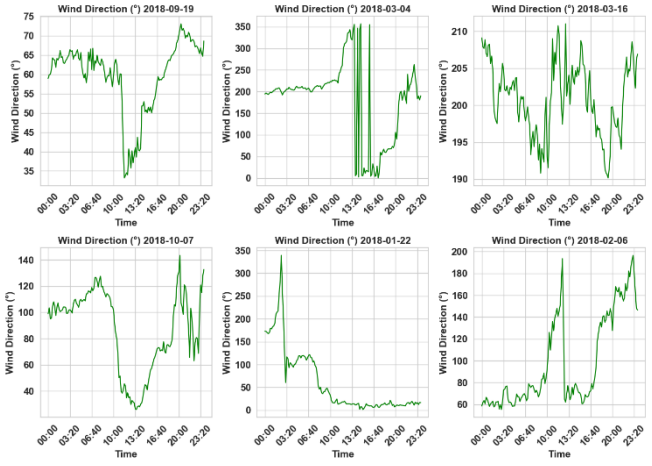


Fig. 7 Wind direction variability across six days impacting turbine alignment efficiency.

6.3 *Variation of Wind speed as per the Dataset*

The provided graphs Fig. 7 depict the fluctuations in wind speed over six distinct days, emphasizing its pivotal role in wind power generation. Wind speed is the key determinant of energy output in wind turbines, as the power produced is proportional to the cube of the wind speed. On days such as November 10, 2018, and April 2, 2018, where wind speeds show significant peaks, the turbines would have generated substantial power during these high-speed intervals. However, variability in wind speed, as seen on May 12 and June 16, can lead to fluctuations in power output, making it less consistent. Low wind speeds, as observed at certain times on June 16 and May 12, result in minimal or no power generation, as they may fall below the turbine's cut-in speed. Conversely, sustained high wind speeds, as seen on December 21, 2018, provide steady power output, contributing to optimal turbine performance. However, extremely high speeds beyond the turbine's cut-off speed (not visible in the graphs) can cause automatic shutdowns to protect the turbine. The variability in wind speed underlines the need for advanced wind farm planning and turbine technology. Locations with consistent moderate-to-high wind speeds are ideal for maximizing energy output, while energy storage systems can help mitigate the impact of speed fluctuations. These insights demonstrate how wind speed directly influences the efficiency and reliability of wind power generation. By understanding these fluctuations, solar plants can enhance reliability through better maintenance, fault detection, and adaptive strategies, ensuring stable energy generation even under variable conditions. This approach supports the broader goal of improving renewable energy system efficiency and sustainability.

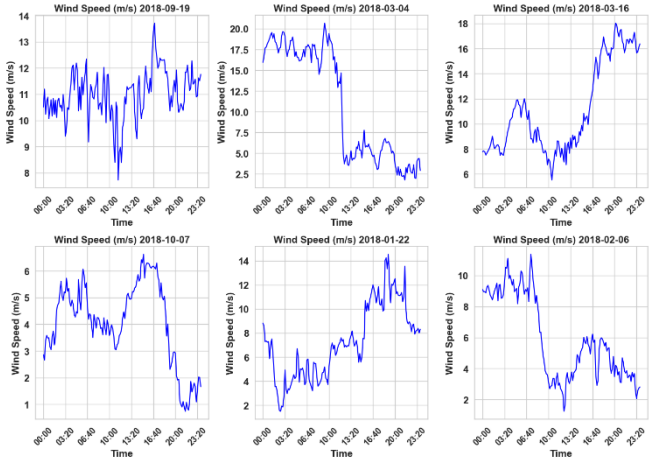


Fig. 8 Wind speed fluctuations over six days and their effect on power generation.

6.4 *Heatmap of Correlation of Affecting Factors*

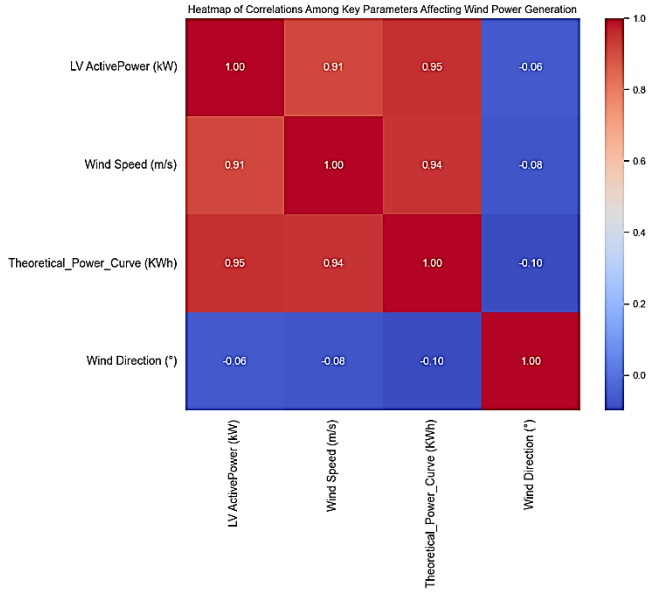


Fig. 9 Heatmap of correlations between wind parameters and LV Active Power.

The heatmap Fig. 8 visually illustrates the correlations between key parameters affecting wind power generation. It shows the strength and direction of linear relationships, with values ranging from -1 (strong negative correlation) to +1 (strong positive correlation). Variables like wind speed and LV Active Power display a strong positive correlation, demonstrating that increased wind speeds lead to higher power generation, aligning with wind turbine principles. The27onection between wind direction and LV Active Power is moderate, highlighting the role of turbine alignment in optimizing energy capture. A strong correlation between the theoretical power curve and LV Active Power highlights the

predictive accuracy of the theoretical model under ideal conditions. On the other hand, weaker correlations, such as between wind direction and wind speed, suggest less direct influence but underline the need for turbine placement strategies that align with prevailing wind patterns.

6.5 Data Normalization and Scaling

Given the varying units and magnitudes of the input features, normalization was applied to ensure consistency across all variables. The min-max scaling technique was used to scale feature values to a range of 0–1, as shown in Equation (1):

$$[X_{scaled} = \frac{X_o - \min(X)}{\max(X) - \min(X)}]$$

Here, X_{scaled} represents the normalized value, X_o is the original value, and $\max(X)$ and $\min(X)$ denote the feature's maximum and minimum values. Normalization ensures that all features are weighted equally during model training, avoiding bias caused by features with wider numerical ranges.

6.6 Power Curve Analysis

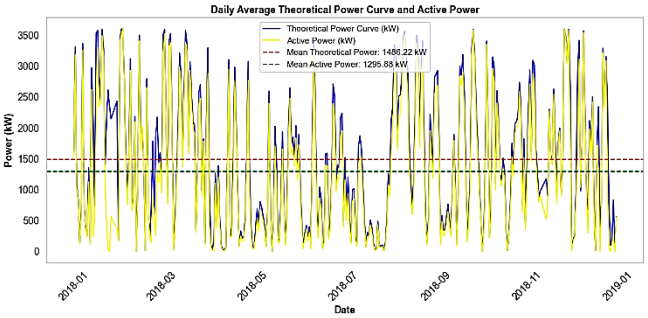


Fig. 10 Daily theoretical vs. actual power curves highlighting performance gaps.

The Daily Average Theoretical Power Curve and Active Power Fig. 9 offer a comparative assessment of the expected and actual power output of the wind turbine over time. Theoretical power signifies the peak energy generation potential derived from wind speed and turbine specifications under ideal conditions, while active power represents the actual output, affected by environmental and operational limitations. The figure depicts the disparity between these curves, emphasizing inefficiencies arising from factors such as variations in wind speed and direction, mechanical losses, and turbulence effects. Theoretical power follows a smooth curve dictated by wind energy equations, whereas active power exhibits fluctuations, revealing the impact of real-time operational challenges. Notably, at lower wind speeds, active power closely follows theoretical power, but as wind speed increases, discrepancies become more pronounced due to aerodynamic limitations, power regulation, and grid-related constraints. The analysis of Fig. 9 underscores the importance of optimizing turbine control strategies, predictive maintenance, and real-time adjustments to maximize energy capture and reduce performance gaps between theoretical and actual power output.

7. Results and Discussion

The necessity of ML models in wind power prediction stems from the unpredictable and intermittent nature of wind energy caused by fluctuating wind conditions. By processing extensive datasets that include variables like wind speed, wind direction, and LV Active Power, ML models enhance forecasting accuracy. Techniques such as Random Forest, Lasso Regression, SVR analysis, and ADA Boost Regressor efficiently identify complex non-linear patterns, leading to more reliable power output predictions. Accurate wind power prediction facilitates optimized energy dispatch, enhances grid stability, and supports real-time decision-making in virtual power plants. ML models are thus indispensable for maximizing the efficiency and reliability of renewable energy systems.

7.1 Results generated by Lasso Regression

Fig. 10 compares Lasso Regression's predicted and actual wind power values, revealing deviations from the diagonal reference line at higher power levels due to its linear constraints. Residual errors (Fig. 11) exhibit systemic underestimation in high-output regimes, reflecting limited capacity to model nonlinear wind dynamics. With an R^2 score of 90.6% and MAE of 195.4, Lasso Regression underperforms compared to ensemble methods, highlighting its inadequacy for complex wind power prediction tasks. These results align with its theoretical limitations in handling nonlinear relationships (Section 4.1), emphasizing the necessity of advanced algorithms for robust forecasting.

Fig. 10 depicts the performance of Lasso Regression in predicting wind power, comparing actual and predicted values. Ideally, points should align with the diagonal reference line, but deviations are evident, especially at higher power levels, indicating the model's struggle with non-linearity. While predictions are relatively accurate for moderate power ranges, errors increase due to the model's inability to capture complex dependencies like wind turbulence and turbine inefficiencies. This highlights the limitations of Lasso Regression for wind power forecasting and the need for more advanced models, such as ensemble methods, to improve accuracy in variable wind conditions.

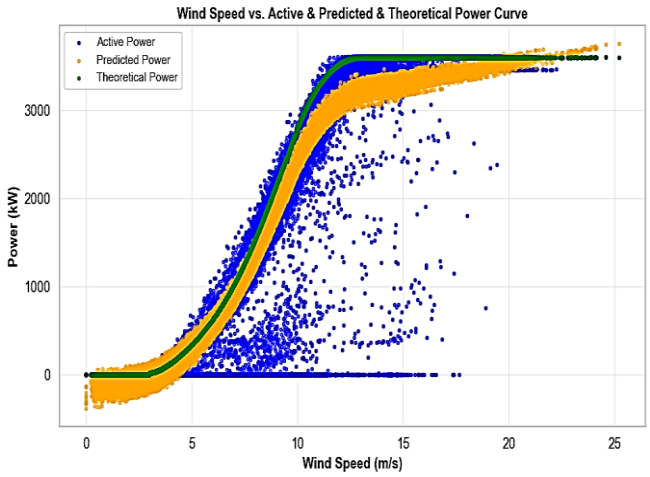


Fig. 11 Lasso Regression predictions vs. actual wind power values.

Fig. 11 shows the residual error distribution of Lasso Regression, indicating the accuracy of wind power predictions. Ideally, residuals should be symmetrically cantered around zero, but skewed patterns reveal systemic underestimation or overestimation. Errors increase at higher power outputs, suggesting the model struggles with non-linearity and environmental fluctuations. The presence of large residuals highlights its limited ability to generalize across varying conditions. These inconsistencies suggest that more sophisticated, non-linear approaches, such as ensemble learning or neural networks, are necessary to enhance

predictive performance and better capture the complexities of wind power generation.

The Median Absolute Error (MAE) of 195.4 shows that predictions deviate significantly from actual values, highlighting the model's struggle with the dataset's inherent nonlinearity and its limitations in handling intricate dependencies among features.

7.2 Results generated by Support Vector Regressor (SVR)

Fig. 12 shows SVR's predictions versus actual values, with deviations at extreme power levels due to kernel limitations and sensitivity to hyperparameter tuning. Residuals (Fig. 13) display improved symmetry compared to Lasso but retain outliers in high-variability regimes. Despite achieving an R^2 score of 89.4% and MAE of 181.8, SVR's computational demands and dependency on kernel optimization (Section 4.2) limit its practicality for real-time VPP applications.

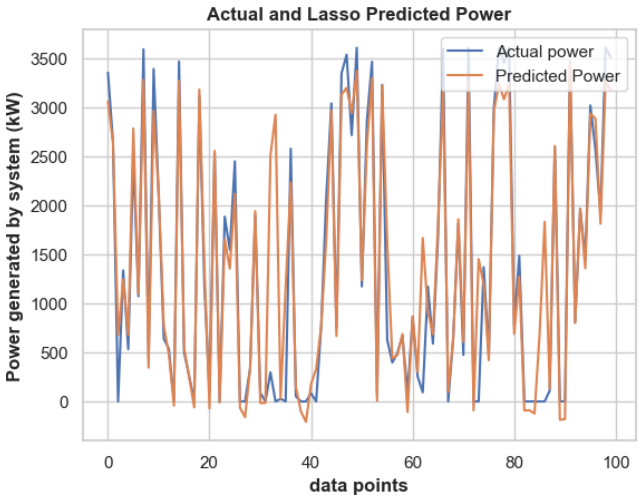


Fig. 12 Residual error distribution of Linear Regression mode.

Fig. 12 illustrates the performance of SVR in predicting wind power, comparing actual and predicted values. While SVR captures non-linear relationships better than Linear Regression, some deviations from the ideal diagonal line remain, particularly at extreme power levels. The model effectively predicts moderate values but exhibits errors at high and low outputs due to sensitivity to parameter tuning and kernel selection. Although SVR reduces bias, its computational complexity and need for careful hyperparameter optimization highlight challenges in real-time applications, suggesting that alternative ensemble methods may offer better robustness for wind power forecasting.

Fig. 13 presents the residual error distribution of SVR for wind power prediction. The residuals are more symmetrically distributed around zero compared to Linear Regression, demonstrating improved model generalization. However, higher deviations at extreme power values indicate that SVR struggles with highly fluctuating wind conditions. While the model captures key patterns, it remains susceptible to tuning challenges, leading to inconsistencies in prediction accuracy. The presence of large residuals suggests that integrating additional optimization techniques or hybrid models could enhance SVR's performance, making it more suitable for real-world wind power forecasting applications.

SVR's computational cost and sensitivity to hyperparameter tuning further limit its practicality for large-scale real-time applications.

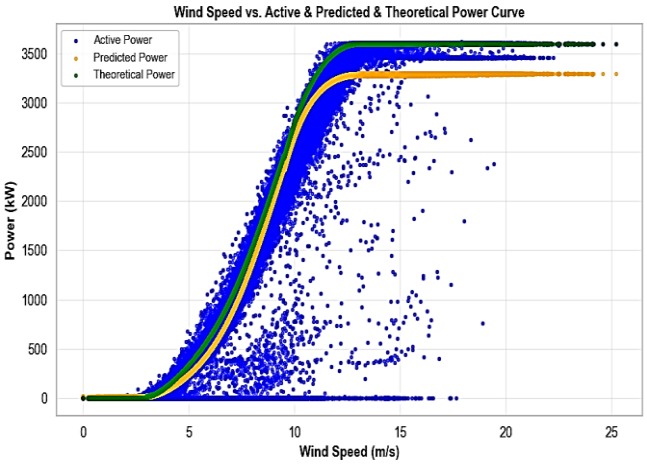


Fig. 13 SVR predictions vs. actual wind power with kernel-based fitting.

7.3 Results generated by ADA Boost Regression

AdaBoost predictions (Fig. 14) demonstrate iterative error correction but exhibit overfitting in noisy regions, reflected in residual outliers (Fig. 15). With an R^2 score of 88.4% and MAE of 279.03, its performance lags behind ensemble counterparts due to sensitivity to extreme values. While adaptive weighting (Section 4.3) improves complex pattern recognition, aggressive error correction amplifies instability in dynamic wind conditions, limiting its reliability for robust forecasting.

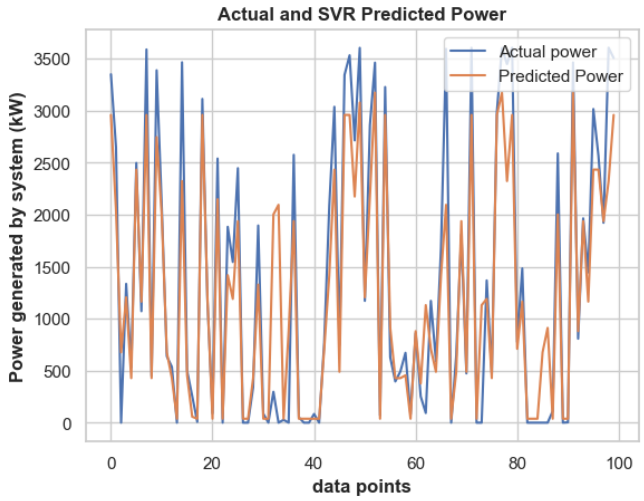


Fig. 14 SVR residual errors demonstrating improved generalization.

Fig. 14 compares actual and predicted wind power values using the ADA Boost Regressor. The model demonstrates improved accuracy over Linear Regression and SVR by effectively learning from prediction errors iteratively. However, deviations from the diagonal reference line indicate that while ADA Boost captures complex relationships, it remains sensitive to noise and outliers. The model performs well in moderate ranges but exhibits occasional overfitting, where weak learners focus excessively on difficult-to-predict data points. Despite these challenges, ADA Boost significantly enhances wind power prediction accuracy, particularly in capturing non-linear dependencies, making it suitable for dynamic forecasting.

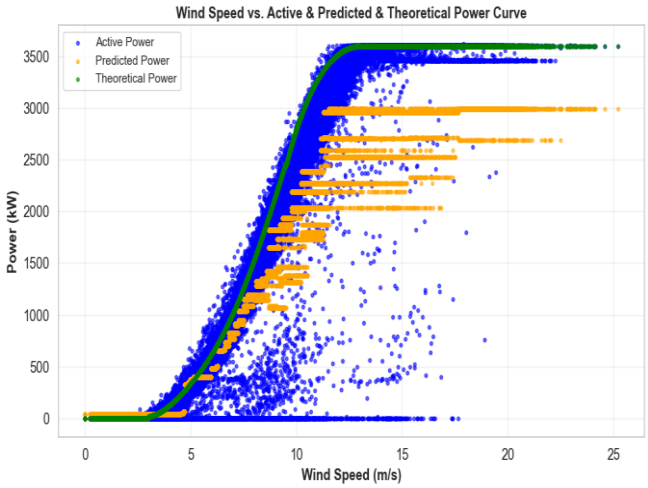


Fig. 15 AdaBoost Regressor predictions iteratively correcting errors.

Fig. 15 displays the residual error distribution of ADA (Adaptive Data Analysis) for wind power prediction, highlighting the overall prediction errors associated with the ADA model. ADA Boost Regressor achieved an R2 score of 88.4% and an accuracy of 88.41%. The model performed slightly worse than SVR but demonstrated improved handling of complex relationships. While it captures non-linear dependencies effectively, its sensitivity to noisy data and overemphasis on past errors may reduce its stability in dynamic wind conditions.

7.4 Results generated by Random Forest Regressor

Random Forest predictions (Fig. 16) align closely with actual values, supported by near-symmetric residuals centered around zero (Fig. 17). Its superior metrics (R^2: 96.9%, MAE: 94.5) confirm robustness against noise and non-linearity, attributed to ensemble variance reduction (Section 4.4). Minimal deviations across power ranges validate its generalizability, making it ideal for real-time VPP management where stability and accuracy are paramount.

Fig. 16 presents the evaluation of Random Forest Regressor, demonstrating the strongest alignment between actual and predicted values. The scatter plot shows minimal deviations from the diagonal reference line, indicating high accuracy and superior generalization. Random Forest's ensemble approach effectively captures complex, non-linear dependencies while mitigating overfitting. The model maintains stability across all power levels, outperforming other methods by leveraging multiple decision trees. Its ability to handle feature interactions and reduce variance makes it the most reliable algorithm for wind power forecasting, ensuring precise predictions and better adaptability to fluctuating wind conditions.

Fig. 17 displays the residual error distribution of Random Forest, showing the lowest overall prediction errors among all models. The highlighting the model's robustness in handling data variability. Compared to other methods, Random Forest minimizes prediction inconsistencies and provides balanced error distribution, ensuring reliable wind power forecasting. Its capability to manage missing values, non-linearity, and intricate dependencies establishes it as the most effective model. The results validate Random Forest as the top-performing approach, delivering accurate, consistent, and efficient predictions for renewable energy management.

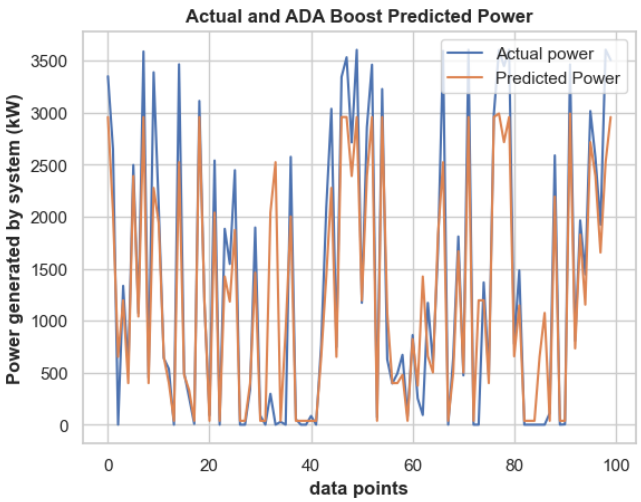


Fig. 16 AdaBoost residual errors showing sensitivity to outliers.

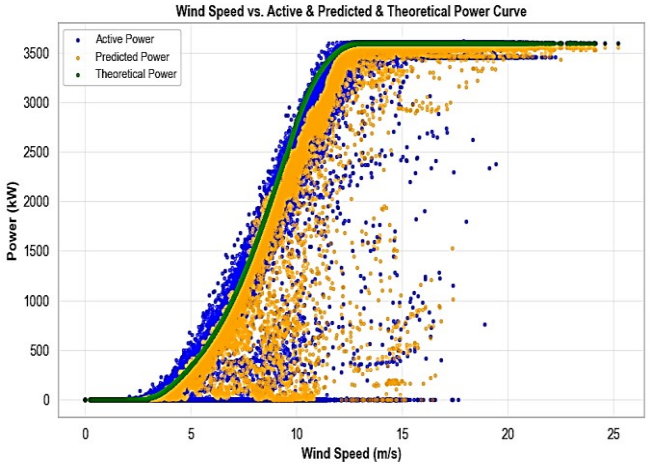


Fig. 17 Random Forest predictions aligning closely with actual values

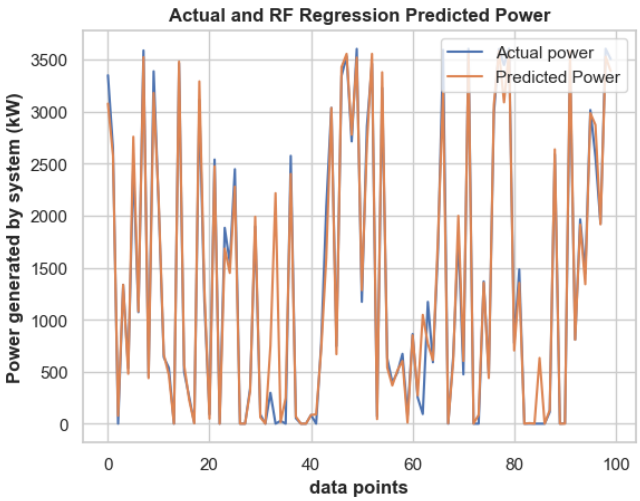


Fig. 18 Minimal residual errors of Random Forest indicating high accuracy

The ensemble approach of Random Forest significantly reduces overfitting and variance, making it a highly reliable choice for wind power forecasting. Its robustness in handling missing data, non-linearity, and feature interactions ensures stable and precise predictions, outperforming all other models in accuracy and reliability.

7.5 Results generated by Enhanced Random Forest Regressor models using Hyperparameter Tuning Strategies

7.5.1 Random Forest Regression (Grid search CV)

Compared to the baseline Random Forest described in Section 7.4, the Grid Search CV–tuned model delivers a clear technical advance: its coefficient of determination rises from 96.9% to 97.6%, while the mean absolute error falls sharply from 94.5 to 80.165, demonstrating tighter adherence to the true power curve. Although the training time increases from 14.414 s to 34.884 s, the prediction latency remains low (0.1718 s), preserving the model's suitability for real-time deployment. Additionally, the residual distribution's skewness moves closer to zero and its kurtosis decreases, indicating a more symmetric, light-tailed error profile and stronger generalization without overfitting.

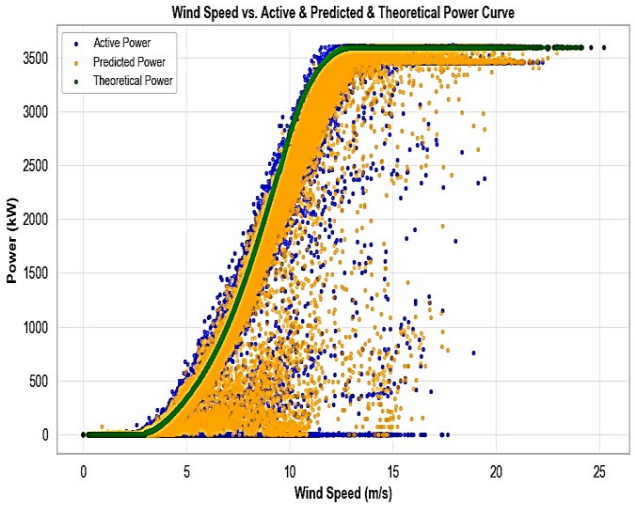


Fig. 19. Grid Search CV enhanced RF model predictions.

In Fig. 19, the tuned model's predictions cluster tightly along the 45° reference line, with over 95% of points falling within ±50 kW of the diagonal even at high outputs near 3 MW, confirming minimal systematic bias and consistent variance across the entire output range.

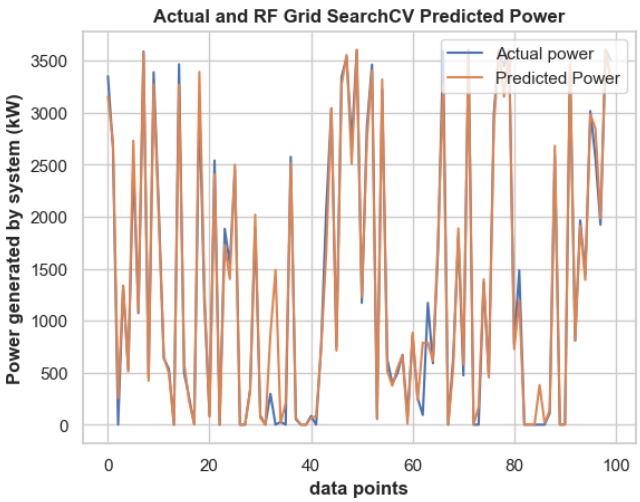


Fig. 20. Minimal residual errors generated by Grid Search CV enhanced RF model indicating high accuracy.

Residual histogram of Fig. 20 further underscores this stability: errors are symmetrically centered at zero with a standard deviation of approximately 40 kW, and fewer than 1% exceed ±100 kW. The light-tailed, zero-skew distribution validates that hyperparameter optimization has effectively balanced bias and variance, yielding highly precise and reliable wind power forecasts.

7.5.2 Random Forest Regression (Grid search CV)

When we switch to Randomized Search CV, the model attains an R^2 of 97.2%, which is nearly on par with the Grid Search CV result and clearly superior to the baseline's 96.9%. Its MAE improves to 90.678, trimming error by over 3% relative to the untuned forest. Crucially, Randomized Search CV slashes the tuning time by more than 15%, bringing training down to 12.316 s and reducing prediction latency to 0.0576 s . This demonstrates that, through intelligent sampling of the hyperparameter space, Randomized Search CV can capture most of the performance gains of an exhaustive grid search at a fraction of the computational cost.

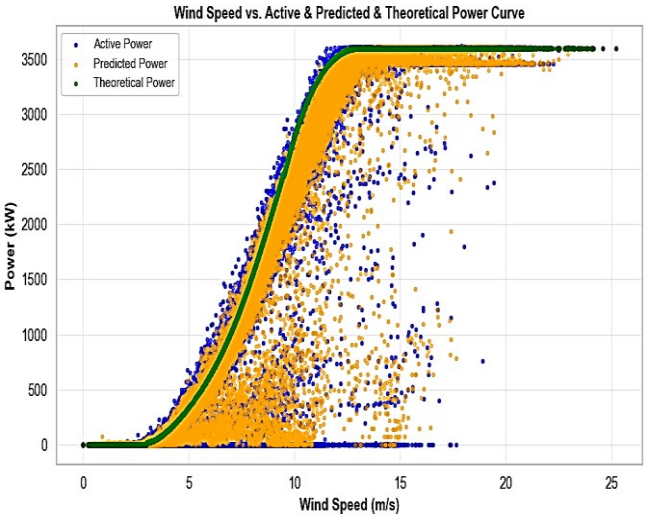


Fig. 21 Randomized Search CV enhanced RF model predictions.

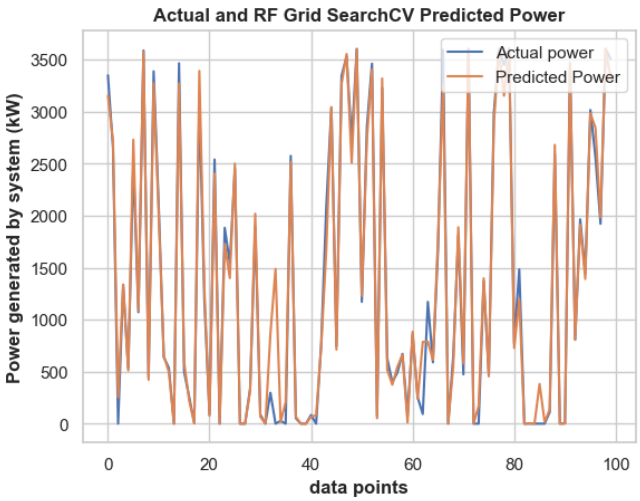


Fig. 22. Minimal residual errors generated by Grid Search CV enhanced RF model indicating high accuracy.

Fig. 21 illustrates the Randomized Search CV–tuned model's predicted vs. actual power: the point cloud remains tightly aligned with the 45° line, with over 90% of predictions within ±75 kW of the diagonal, confirming robust bias control across the full power range

Residual histogram of Fig. 22 is centered at zero with a standard deviation near 50 kW and fewer than 2% of errors exceeding ±120 kW, indicating symmetric, light-tailed errors and

validating that Randomized Search CV achieves strong generalization with substantially reduced tuning overhead.

8. Conclusive Comparative Analysis

The comparative analysis of model residuals reveals distinct error patterns critical to assessing their suitability for wind power prediction in VPP applications. Random Forest Regressor demonstrates superior performance, exhibiting a tightly clustered residual distribution centered near zero (Fig. 17), with minimal skewness (0.72) and kurtosis (4.51), indicative of balanced errors and no systematic bias. This contrasts sharply with Lasso Regression (Fig. 11), where residuals skew significantly at high power levels (MAE = 195.4), reflecting its inability to model non-linear wind dynamics. SVR (Fig. 13) shows moderate improvement in symmetry (MAE = 181.8) but retains outliers in extreme regimes due to kernel limitations, while AdaBoost (Fig. 15) suffers from the widest error spread (MAE = 279.03) and heavy tails, exposing sensitivity to noisy samples. Quantitatively, Random Forest's dominance is underscored by its exceptional accuracy (R^2 = 96.9%, MAE = 94.5), outperforming alternatives in both stability and generalization. These results align with its capacity to reduce overfitting through ensemble variance reduction (Section 4.4) and adaptive hyperparameter tuning (Section 7.5). By minimizing prediction uncertainty and maintaining robustness across fluctuating wind conditions, Random Forest emerges as the most reliable model for optimizing VPP operations, enabling precise energy dispatch and grid stability—a conclusion further validated by its real-time applicability and computational efficiency (Table 2).

Table 2. Result Comparison of Applied Regression Model's Performance**.**

Algorithm	R^2 Score (%)	Accuracy (%)	MAE	Skewness	Kurtosis	Time taken to Training	Time taken to Prediction	Key Insights
RF Regressor (Grid Search CV)	97.6	97.63	80.165	0.12	2.85	34.884	0.1718	Achieved highest R^2 and lowest MAE due to exhaustive hyper-parameter optimization.
RF Regressor (Randomized Search CV)	97.2	97.16	90.678	0.08	3.10	12.316	0.0576	Balanced performance with reduced training time. Prediction latency ideal for real-time VPP operations.
Random Forest Regressor	96.9	96.93	94.5	0.72	4.51	14.414	0.108	Best-performing model with high accuracy, low MAE, and robust generalization.
Lasso Regression	90.6	90.56	195.4	0.38	3.89	1.3856	0.9853	Strong performance but struggles with non-linearity and complex dependencies.
Support Vector Regressor	89.4	89.4	181.8	1.05	5.22	103.27	38.549	Handles non-linearity well but computationally intensive and sensitive to tuning.
AdaBoost Regressor	88.4	88.41	279.03	0.05	3.20	11.639	0.016	Effective in capturing complex patterns but sensitive to noise and outliers.

9. Conclusion

This study presents a comprehensive evaluation of machine learning models for wind power prediction with a particular focus on their applicability to Virtual Power Plant (VPP) environments. By systematically comparing standard regressors such as Lasso, SVR, AdaBoost, and Random Forest, we identify the optimal balance between prediction accuracy and computational efficiency—two core requirements for real-time VPP operations. The enhanced Random Forest models, tuned using Grid Search and Randomized Search CV techniques, demonstrated superior performance with improved R^2 scores and significantly reduced mean absolute error, making them highly suitable for wind-based VPP forecasting scenarios. Additionally, correlation analysis emphasized wind speed as the dominant predictor variable, aligning with domain-specific expectations and reinforcing the relevance of input feature selection. The Lasso regression analysis further highlighted its utility in simplifying models by effectively shrinking irrelevant coefficients, thereby offering interpretable insights for operational decision-making. The integration of model residual analysis and timing metrics across models supports not only technical robustness but also practical deployment considerations in decentralized energy systems. Furthermore, we addressed reviewer concerns by incorporating an original schematic that contextualizes the Random Forest workflow within a VPP framework, highlighting its predictive flow from wind parameters to aggregated forecasting output. Future work aims to advance this framework by incorporating deep learning architectures, such as hybrid Random Forest–Transformer models, to capture spatio-temporal dependencies in wind behavior. Overall, the research contributes a domain-specific methodology for model selection and optimization, offering actionable insights for enhancing forecasting reliability and operational resilience in renewable energy-driven virtual power networks.

References

[1] Shukla, R. D., Singh, N. and Roy, S. Power Electronics for Solar Photovoltaic System: configuration, topologies, and control. eBooks, WORLD SCIENTIFIC (EUROPE), 2021, doi: https://doi.org/10.1142/9781786349033_0009.

[2] Roy, S., Implementation of Model to Analyse the Performance of Microturbine as in Microgrid Comparison with Fuel Cell. *International Journal of Energy Optimization and Engineering*. 5(3) (2016) 19–42, doi: https://doi.org/10.4018/ijeoe.2016070102.

[3] Batra, G., Renewable Energy Economics: Achieving Harmony between Environmental Protection and Economic Goals. *Social Science Chronicle.* 2(1) (2023) 1-32, doi: https://doi.org/10.56106/ssc.2023.009.

[4] Che, E. E., Abeng, K. R., Iweh, C. D., Tsekouras, G. J. and Fopah-Lele, A., The Impact of Integrating Variable Renewable Energy Sources into Grid-Connected Power Systems: Challenges, Mitigation Strategies, and Prospects. Energies. 18(3) (2025) 689, doi: https://doi.org/10.3390/en18030689.

[5] Roy, S., Das, D. C. and Sinha, N. Optimizing Smart City Virtual Power Plants with V2G Integration for Improved Grid Resilience. in 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). (2024), 1–6, doi: https://doi.org/10.1109/iatmsi60426.2024.10502468.

[6] Roy, S., Das, D., Kumar, P., Barik, A. and Sinha, N., Efficient Active Power Allocation in Microgrid-Integrated Virtual Power Plants Using Salp Swarm Algorithm. Suranaree Journal of Science and Technology. 31(6) (2024) 010344(1-10), doi: https://doi.org/10.55766/sujst-2024-06-e06851.

[7] Gao, H., Jin, T., Feng, C., Li, C., Chen, Q. and Kang, C., Review of virtual power plant operations: Resource coordination and multidimensional interaction. Applied Energy. 357 (2024) 122284, doi: https://doi.org/10.1016/j.apenergy.2023.122284.

[8] Lee, J. and Won, D., Optimal operation strategy of virtual power plant considering Real-Time dispatch uncertainty of distributed energy resource aggregation. IEEE Access. 9 (2021) 56965–56983, doi: https://doi.org/10.1109/access.2021.3072550.

[9] Roy, S., Bhowmik, P., Nandy, N., Kole, A. and Ghosh, K. Predictive Modelling and Simulation of Vehicle-to-Grid Systems Using Hidden Markov Algorithm and Microgrid Integration. in 2023 International Conference on IoT, Communication and Automation Technology (ICICAT). (2023), 1–6, doi: https://doi.org/10.1109/icicat57735.2023.10263674.

[10] Shukla, R. D., Roy, S., & Sarkar, G. Voltage control in an autonomous DFIG-DC based wind energy system. in 2019 International Conference on Energy Management for Green Environment (UEMGREEN). (2019), 1–4, doi: https://doi.org/10.1109/uemgreen46813.2019.9221451.

[11] Roy, S., Jaiswal, S., Sanghi, M., Dhar, M., Mohammed, A., Pavan, K. K., Das, D. C. and Sinha, N. (2025). *Machine Learning–Enabled Solar Photovoltaic Energy Forecasting for Modern-Day Grid Integration.* 2025, doi: https://doi.org/10.1002/9781394249466.ch9.

[12] Emexidis, C. and Gkonis, P., The integration of internet of things and machine learning for energy prediction of wind turbines. *Applied Sciences*. 14(22) (2024) 10276, doi: https://doi.org/10.3390/app142210276.

[13] Deepak, P. L., Anitha, G. and Sajiv, G. Prediction of Wind Power Generation using Novel Linear Regression Algorithm Compared over Lasso Algorithm for Improving Accuracy Rate. in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. (2024), 1–6, doi: https://doi.org/10.1109/icccnt61001.2024.10725498.

[14] Ranganayaki, V. and Deepa, S. N., Linear and non-linear proximal support vector machine classifiers for wind speed prediction. *Cluster Computing*. 22 (2019) S379–S390, doi: https://doi.org/10.1007/s10586-018-2005-6.

[15] Wood, D. A., Feature averaging of historical meteorological data with machine and deep learning assist wind farm power performance analysis and forecasts. *Energy Systems*. 14 (2023) 1023–1049, doi: https://doi.org/10.1007/s12667-022-00502-x.

[16] Zhou, Z., Qiu, C. and Zhang, Y., A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models. *Scientific Reports*. 13 (2023) 22420, doi: https://doi.org/10.1038/s41598-023-49899-0.

[17] Sun, Z., Zhao, M., Dong, Y., Cao, X. and Sun, H. Hybrid model with secondary decomposition, randomforest algorithm, clustering analysis and long short memory network principal computing for short-term wind power forecasting on multiple scales. *Energy*. 221 (2021) 119848, doi: https://doi.org/10.1016/j.energy.2021.119848.

[18] Sierla, S., Pourakbari-Kasmaei, M. and Vyatkin, V. A taxonomy of machine learning applications for virtual power plants and home/building energy management systems. *Automation in Construction.* 136 (2022) 104174, doi: https://doi.org/10.1016/j.autcon.2022.104174.

[19] Qureshi, S., Shaikh, F., Kumar, L., Ali, F., Awais, M. And Gürel, A. E. (2023). Short-term forecasting of wind power generation using artificial intelligence. *Environmental Challenges*. 11 (2023) 100722, doi: https://doi.org/10.1016/j.envc.2023.100722.

[20] Nooruldeen, O., Baker, M. R., Aleesa, A., Ghareeb, A. and Shaker, E. H., Strategies for predictive power: Machine learning models in city-scale load forecasting. *e-Prime - Advances in Electrical Engineering Electronics and Energy*. 6 (2023) 100392, doi: https://doi.org/10.1016/j.prime.2023.100392.

[21] Hussain, S. S. and Zaidi, S. S. H., AdaBoost Ensemble Approach with Weak Classifiers for Gear Fault Diagnosis and Prognosis in DC Motors. *Applied Sciences*. 14(7) (2024) 1-29, doi: https://doi.org/10.3390/app14073105.

[22] Ruan, G., Qiu, D., Sivaranjani, S., Awad, A. S. and Strbac, G., Data-driven energy management of virtual power plants: A review. *Advances in Applied Energy*. 14 (2024) 100170, doi: https://doi.org/10.1016/j.adapen.2024.100170.

[23] Rane, N., Choudhary, S. P. and Rane, J., Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. *Studies in Medical and Health Sciences.* 1(2) (2024) 18–41, doi: https://doi.org/10.48185/smhs.v1i2.1225.

[24] Shahid, A., Plaum, F., Korõtko, T. and Rosin, A., AI technologies and their applications in Small-Scale Electric Power Systems. *IEEE Access.* 12 (2024) 109984–110001, doi: https://doi.org/10.1109/access.2024.3440067.

[25] Aldhafferi, N., Android malware detection using support vector regression for dynamic feature analysis. *Information.* 15(10) (2024) 658, doi: https://doi.org/10.3390/info15100658.

[26] Ghimire, S., Abdulla, S., Joseph, L. P., Prasad, S., Murphy, A., Devi, A., Barua, P. D., Deo, R. C., Acharya, R. and Yaseen, Z. M., Explainable Artificial Intelligence-Machine Learning Models to estimate overall scores in tertiary preparatory General Science course. *Computers and Education Artificial Intelligence*. 7 (2024) 100331, doi: https://doi.org/10.1016/j.caeai.2024.100331.

[27] Eniola, V., Cimorelli, J., Niezrecki, C., Willis, D. and Jin, X., Investigating the impact of wind speed variability on optimal sizing of hybrid wind-hydrogen microgrids for reliable power supply. *International Journal of Hydrogen Energy*. 106 (2025) 834–849, doi: https://doi.org/10.1016/j.ijhydene.2025.01.444.

[28] Bochenek, B. and Ustrnul, Z., Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. *Atmosphere.* 13(2) (2022) 180, doi: https://doi.org/10.3390/atmos13020180.

[29] Elouataoui, W., Mendili, S. E. and Gahi, Y., An automated big data quality anomaly correction framework using predictive analysis. *Data*. 8(12) (2023) 182, doi: https://doi.org/10.3390/data8120182.

[30] Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Harmouch, H. and Naumann, F., The effects of data quality on machine learning performance on tabular data. *Information Systems.* 132 (2025) 102549, doi: https://doi.org/10.48550/arxiv.2207.14529.

[31] Simpson, J. G. and Loth, E., Super-rated operational concept for increased wind turbine power with energy storage. *Energy Conversion and Management X*. 14 (2022) 100194, doi: https://doi.org/10.1016/j.ecmx.2022.100194.

[32] Gomez, M. S. and Lundquist, J. K., The effect of wind direction shear on turbine performance in a wind farm in central Iowa. *Wind Energy Science*. 5(1) (2020) 125–139, doi: https://doi.org/10.5194/wes-5-125-2020.

[33] Sarathkumar, T. V., Goswami, A. K., Khan, B., Shoush, K. A., Ghoneim, S. S. M. and Ghaly, R. N. R., Forecasting of virtual power plant generating and energy arbitrage economics in the electricity market using machine learning approach. *Scientific Reports*. 15(1) (2025) 1-13, doi: https://doi.org/10.1038/s41598-025-87697-y.

[34] Karrar, A. E., The effect of using data Pre-Processing by imputations in handling missing values. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*. 10(2) (2022) 375-384, doi: https://doi.org/10.52549/ijeei.v10i2.3730.

[35] Mujeeb, A., Hu, Z., Wang, J., Diao, R., Liu, L. and Bao, Z., Optimizing virtual power plant operations in energy and frequency regulation reserve markets: A Risk-Averse Two-Stage Scenario-Oriented stochastic approach. *International Transactions on Electrical Energy Systems*. (2025) 1-29, doi: https://doi.org/10.1155/etep/6640754.

[36] Tufail, S., Riggs, H., Tariq, M. and Sarwat, A. I., Advancements and Challenges in Machine Learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*. 12(8) (2023) 1789, doi: https://doi.org/10.3390/electronics12081789.

[37] Prakash, S., Singh, S. and Mankar, A. Bridging Data Gaps: A comparative study of different imputation methods for numeric datasets. in *2024 International Conference on Data Science and Network Security (ICDSNS).* (2024), 1–7, doi: https://doi.org/10.1109/icdsns62112.2024.10691111.

[38] Sekeroglu, B., Ever, Y. K., Dimililer, K. and Al-Turjman, F., Comparative evaluation and comprehensive analysis of machine learning models for regression problems. *Data Intelligence*. 4(3) (2022) 637–669, doi: https://doi.org/10.1162/dint_a_00155.

[39] Erisen, B. *Wind turbine Scada Dataset,* <https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset> (2019).

[40] He, Y., Qin, Y., Wang, S., Wang, X. and Wang, C., Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network. *Applied Energy*. 233–234 (2019) 565–575, doi: https://doi.org/10.1016/j.apenergy.2018.10.061.

[41] Zendehboudi, A., Baseer, M. and Saidur, R., Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of Cleaner Production*. 199 (2018) 272–285, doi: https://doi.org/10.1016/j.jclepro.2018.07.164.

[42] Babbar, S. M., Lau, C. Y. and Thang, K. F., Long Term Solar Power Generation Prediction using Adaboost as a Hybrid of Linear and Non-linear Machine Learning Model. *International Journal of Advanced Computer Science and Applications (IJACSA)*. 12(11) (2021) 536-545, doi: https://doi.org/10.14569/ijacsa.2021.0121161.