International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 24(2), November, 2024, pp 137-154
DOI: https://doi.org/10.3329/ijss.v24i2.78000

Optimizing Wind Speed Predictions in Bangladesh: Unveiling the Power of Machine Learning Approaches

Mira Khatun[1], Golam Sabbir Sattar[2], Provash Kumar Karmokar[3] and Md. Rezaul Karim[3*]

[1]Department of Statistics, Pabna University of Science and Technology, Pabna, Bangladesh
[2]Department of Geology and Mining, University of Rajshahi, Rajshahi, Bangladesh
[3]Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh
[*]Correspondence should be addressed to Md. Rezaul Karim
(Email: mrkarim@ru.ac.bd)

[Received September 7, 2024; Accepted October 28, 2024]

Abstract

Wind speed is a naturally occurring phenomenon that arises from the intricate interplay of various atmospheric processes. Wind speed prediction is pivotal in various sectors worldwide, and Bangladesh is no exception. Beyond its impact on agriculture, water management, and disaster preparedness, wind speed also plays a crucial role in urban planning and construction projects. Architects and engineers rely on accurate wind speed forecasts to design buildings and infrastructure that can withstand local wind conditions. Furthermore, the aviation and maritime industries heavily depend on wind speed predictions to ensure the safety of flights and shipping routes. Predicting wind speeds in Bangladesh poses a significant challenge due to the region's susceptibility to frequent seasonal changes influenced by its coastal location and complex, nonlinear climate patterns. To address this important aspect, we leverage Machine Learning (ML) algorithms, including Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), Random Forest (RF), K Nearest Neighbors (K-NN), and Support Vector Machine (SVM) to forecast wind speeds at various weather stations in Bangladesh. We utilized various accuracy metrics, including precision, sensitivity, specificity, F-measure, and overall accuracy, to evaluate the performances of the algorithms. The RF model outperformed the other models with an overall accuracy of 94.73% for predicting wind speed conditions in Bangladesh. On the other hand, the LDA model exhibited the lowest performance, achieving an accuracy of 93.27% in comparison to the other models. It is noticeable that the aforementioned five models showed more than 90% accuracy for windspeed prediction. Additionally, we complement our analysis with visual representations such as box plots, density plots, dot plots, parallel plots, and scatterplot matrix plots. These empirical results also highlighted the RF as the most suitable method for predicting contemporary wind speed patterns in Bangladesh.

Keywords: Wind Speed Prediction, Machine Learning, Random Forest, Linear Discriminant Analysis, Classification and Regression Trees, K Nearest Neighbors, Support Vector Machine, Accuracy Metrics.

AMS Classification: 91B76, 68T09, 62H12, 62R07.

1. Introduction

Accurate and timely prediction of climatic variables is crucial for proactive safety measures. The climate variable wind speed plays a pivotal role in climate assessment. The world is currently grappling with the consequences of climate change, as weather conditions continue to shift and become increasingly extreme with each passing day [1], [2]. To address these challenges, numerous novel techniques and technologies have emerged for modeling and predicting climatic variables. Among these innovative approaches, machine learning (ML) algorithms have demonstrated their effectiveness as robust tools for this important task [31]. Accurate forecasts of wind speed are essential for determining optimal flight paths, departure and arrival times, detecting and tackling deforestation, and shipping schedules. Wind speed and solar radiation serve as the primary factors of renewable energy sources [3]. Renewable energy sources such as solar, wind, and hydropower are crucial for the modern energy sector (Shawon et al., 2021). As the energy crisis intensifies, the utilization of wind turbines to harness wind energy, a widely acknowledged and promising renewable resource, is becoming increasingly prevalent [5].

Consequently, the precise prediction of wind energy, especially in terms of wind speed forecasting, assumes an indispensable role in the effective management of wind energy resources [6]. Inaccurate wind speed predictions can lead to delays, increased fuel consumption, and safety risks. Additionally, wind speed forecasts are instrumental in weather forecasting, enabling meteorologists to provide more precise and timely weather predictions [7]. This, in turn, benefits the general public by providing notice in advance of severe weather events such as storms, cyclones, and hurricanes. These aspects highlight the multifaceted importance of wind speed prediction beyond the sectors previously mentioned, making it a critical area of study and research in Bangladesh and globally. Models for predicting wind speed can aid in preserving people's lives and property, which in turn helps the economy of the country. The direction and speed of wind influence the rate of evaporation, which in turn influences the water level. It results from air shifting from high to low pressure, frequently as a result of temperature variations. Numerous other things, such as construction projects, plant growth-metabolic rates, aviation-maritime activities, and weather forecasting, are impacted by wind speed [8].

Wind speed forecasting models have witnessed a significant transformation with the integration of advanced ML-based approaches [9]. Bangladesh experiences rapid shifts in weather and climate, making it essential to employ models that can continuously learn and adjust their predictions. In this context, where weather conditions exhibit considerable variability, the utilization of ML techniques for wind speed prediction holds promise. These models leverage historical weather data, atmospheric conditions, and other relevant factors to generate forecasts with enhanced accuracy. One notable advantage of ML-based wind speed prediction models is their ability to adapt to changing climate patterns (Karaman, 2023). By enhancing the precision of wind speed predictions in Bangladesh, ML models can contribute to increased energy efficiency and sustainability [10], [11].

From the explanation above, it is clear that this research reflects the ongoing efforts of many scientists who are continuously trying to identify the most effective predictive models for precise wind speed forecasts. These models often involve the optimization and integration of diverse data mining techniques across different weather stations.

The overall sequences of this paper are organized as follows: Section 2 focuses on the literature review. Section 3 encompasses the materials and methodology, including data sources, dataset descriptions, classification algorithms, and an overview of model evaluation criteria. In Section 4, we present the findings and conduct model comparisons using graphical techniques and benchmarks with similar existing studies. Lastly, Section 5 provides a summary of concluding remarks and outlines directions for future research.

2. Review of Literature

Numerous researchers have been diligently employing data mining methods to achieve highly precise wind speed predictions. Researchers [12] [13] undertook a study within the Nigerian context to assess the efficiency of the Artificial Neural Network (ANN) when compared to the Decision Tree (DT) algorithm in forecasting variables such as rainfall, maximum temperature, wind speed, and evaporation. Their investigation revealed that ANN exhibited better predictive capabilities than the Decision Tree algorithm in these specific scenarios.

[14] addressed the challenge of predicting wind speed at various heights, which is crucial for investment decisions in wind farms, especially in Turkey's evolving renewable energy sector. The comparison of seven different machine learning methods highlights SVM as particularly effective for accurate wind speed prediction across different heights, offering a valuable tool for more efficient and informed investment in wind energy. [15] applied the Support Vector Regression (SVR) and Artificial Neural Network (ANN) to forecast wind speed. Their results demonstrated that both SVR and ANN models achieved an accuracy rate of over 99% in short-term wind speed prediction, but ANN models can be slightly more useful than the SVR models.

Some researchers [16] [17] emphasized the importance of accurate and timely wind speed forecasting. They highlighted that such forecasts play a crucial role in enabling policymakers to proactively plan for various activities, including agriculture, flight operations, and upcoming construction projects. In recent years, several attempts have been made to construct an appropriate model in the time series area based on wind speed data. The objectives, tests, and the ability of the models to represent climate behavior suggest that the wind speed dataset may be widely accepted as a reference dataset in climate research. [18]presented a comprehensive analysis of seven numerical methods for estimating Weibull parameters, which are crucial for wind energy applications. The study reveals significant variations in the performance of the methods, highlighting the importance of selecting the appropriate technique for accurate wind speed predictions. [8] mentioned that the Mean Wind Speed (MWS) is one important characteristic that influences the climate, so they predicted the mean wind speed using ML algorithms: linear regression, random forest, and deep neural network. (Rahman et al., 2021) aimed to predict rainfall data specifically for the Rajshahi district in Bangladesh. They employed various ML algorithms and found that the K-Nearest Neighbors (KNN) method yielded the most accurate predictions. [20] investigated the application of extreme gradient boosting (EGB) in flood susceptibility modeling. (Shawon et al., 2021) conducted a study of short-term wind speed forecasting using one year of historical data. It compares conventional time series methods (ARMA and ARIMA) with ML methods, and ML methods were found to outperform conventional time series methods in predicting short-term wind speed.

[22] employed the group method of data handling (GMDH), multilinear regression (MLR), and artificial neural network (ANN) models to predict wind speed. The paper concludes that the ANN model provides superior accuracy compared to other models. [23] found the ANN with one hidden layer and five neurons is the best-fitted model for analyzing wind speed data from 12 stations in Pakistan. They proposed extending the research by incorporating advanced algorithms like random forests and optimization techniques. (Karaman, 2023) applied machine learning techniques, including artificial neural networks (ANNs), recurrent neural networks (RNNs), convolutional

neural networks (CNNs), and long short-term memory (LSTM) networks. The study highlights LSTM as the most effective method for wind power forecasting. [25] explored the challenge of wind energy's intermittency by comparing the effectiveness of two forecasting models: SARIMA and LSTM. By using statistical error metrics, the results indicate that the LSTM model outperforms SARIMA, making it a more reliable tool for wind speed prediction. [26] compared artificial neural networks (ANN) and multiple linear regression (MLR) for predicting wind speed in Chhattisgarh, India. The paper suggested that the ANN approach outperforms MLR in accuracy and prediction reliability, demonstrating superior performance in capturing complex wind patterns. [27] used meteorological data from 2014-2021 to estimate wind speed in Gokceada, Turkey, using WEKA software. Among various algorithms, Kstar yields the best results for the temperature-pressure-humidity group based on the values of the correlation coefficient and RMSE. The findings highlight the effectiveness of certain algorithms in predicting wind speed based on specific input variables. The study of [28] introduced a machine learning-based method that leverages both ANN and SVM to enhance the prediction accuracy of wind speed. By utilizing data from the Nigeria Meteorological Agency for 2016, the study presented a well-structured forecasting model that offers reliable predictions. [29] conducted a study in which they employed ML and deep learning models to predict PM2.5 concentration in Delhi. They utilized three ML models (multi-layer feed-forward neural network (MLFFNN), SVM, RF) and one deep learning model long short-term memory networks (LSTM) and considered various air pollutants and meteorological parameters as inputs. The results indicated that the LSTM model performed the best with high accuracy metrics for PM2.5 prediction.

[30] evaluated 14 regression-based machine learning models to forecast short-term wind speeds in Kutubdia and Cox’s Bazar, using data from the Bangladesh Meteorological Department and NASA. Among the models, categorical boosting demonstrated superior predictive accuracy. The findings emphasize the importance of accurate wind forecasting for effective turbine site selection, benefiting engineers and project stakeholders in Bangladesh's wind energy sector. [31] explored wind speed prediction in Jerusalem using the Adaptive Neuro-Fuzzy Inference System (ANFIS) and K-Nearest Neighbors Regression (KNNR) algorithms. ANFIS offers high accuracy with a low RMSE and minimal bias, though it struggles with variability. KNNR, while better fitting the data (higher R^2), has a higher RMSE. [32] presented a comparative analysis of various machine learning algorithms for predicting wind speed in Dhaka, Bangladesh. It explored the efficacy of different models such as linear regression, decision trees, random forests, and neural networks. The findings indicated that random forests and neural networks outperform traditional methods, providing robust and reliable predictions. Many researchers [7], [14], [31], [33], [34], have worked on prediction windspeed, yet it remains a subject of research to improve the prediction accuracy.

The country exhibits a wide range of climatic variables, making it one of the most diverse in the world. Although the weather in this country is among the most varied in the world, relatively limited research has been conducted in this area. Particularly, the prediction of wind speed using various data mining techniques across all regions of Bangladesh is a rarely explored topic. Though there is some literature on windspeed prediction, the studies were confined to the traditional method, some compared with time series and ML. This investigation is crucial because wind speed directly impacts wind power generation, which plays a vital role in the country's energy landscape. The findings from this research have the potential to provide valuable information in advance related to rainfall, floods, and droughts, as well as aircraft flights. Such information can help save lives and protect property while also contributing to the overall economic development of the nation.

3. Materials and Methodology

3.1. Dataset Description

A well-known wind speed dataset from the Bangladesh Agricultural Research Centre (BARC) has been considered for the period of 1950 to 2018 in this research. Humidity, maximum temperature, minimum temperature, wind direction, rainfall, and cloud coverage are key factors that influence wind speed. To ensure comparability among these variables, which have different units, we employed a unit-free approach. The wind speed variable takes the numerical value. For target-specific tasks, we converted the wind speed variable into categorical values. By fixing up two conditions, if it gives any value other than 0, then it may be termed as the 'Yes' category; otherwise, the 'No' category. When the 'Yes' represents the windy status of the weather, and the 'No' expresses that there is no excessive flow of wind. To handle the anomalies in the dataset, we employ preprocessing techniques for enhanced prediction, encompassing data cleaning and normalization. During the data preprocessing phase, thorough scrutiny was conducted for missing values and outliers. The dataset is so big, so we didn't impute the missing value rather we omit the values. Additionally, to ensure consistent scaling, numeric variables are standardized through a process of subtracting the mean and dividing by the standard deviation. Table 1 provides the details of the data set description.

Table 1: Variables used in the paper

Name of Variables	Type	Measurement
Wind speed	Numerical	Meter per second (ms^{-1})
Wind speed Status	Categorical	Yes / No
Humidity	Numerical	Percentage (%)
Maximum Temperature	Numerical	Degrees Celsius (°C)
Minimum Temperature	Numerical	Degrees Celsius (°C)
Cloud Coverage	Numerical	eighths of the sky covered by clouds (1/8)
Rainfall	Numerical	Millimeters per day (mm/day)

3.2 Proposed Methodology

First, the data preprocessing is conducted in the above windspeed dataset. Since it is a big dataset with a sample size of 22475, and almost 10% of data are missing, for the preprocessing phase, we omit the missing values rather than the imputation. Then the data is transformed by the normalization filtering process. Next, the dataset is split into two parts: one is training data, and the other one is testing data. There are many software packages to do this task. We have used the R Programming Software package version 4.1.3 for implementing the algorithms, and for the analytic part, some hyperparameters are considered, e.g., training data 75% and test data 25%, number of resampling iterations 10, and the amount of granularity in the tuning parameter grid is 3. The training data is used to build and train the model, and the test data is used to check its accuracy. The final forecasted best model is identified by its accuracy. Thereafter the researchers tried to show the model's performance by some graphical representations. The utilization of data mining techniques proves to be highly effective in forecasting various environmental attributes by revealing previously undiscovered relationships within historical data. This process commences with a data pre-processing stage, followed by the assessment of the predictive capabilities of different data mining methods. ML has grown to be the facilitator of the field of Data Science and Data Mining, which is, in turn, the facilitator of Big Data [35], [36].

3.3 Classification Algorithms

Classification Algorithms for Wind Speed Prediction

Predicting wind speed is essential for applications in renewable energy, agriculture, and disaster management. This section outlines key machine learning (ML) algorithms used in this study for wind speed prediction from meteorological stations in Bangladesh. The selected algorithms Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF) were chosen for their ability to handle complex weather data effectively.

Classification and Regression Trees (CART)

CART is a decision tree algorithm that partitions the input data into subsets based on decision rules learned from the features. For a given input xxx, the algorithm recursively splits the dataset based on thresholds for variables such as wind speed and rainfall, resulting in a tree where the terminal nodes represent the predicted class labels. The decision tree is represented mathematically as: $T(x) = \sum_{j=1}^{J} I(x \in R_j)c_j$; where $I(\cdot)$ is an indicator function that assigns x to region R_j, and c_j is the predicted class label ("Windy" or "Non-Windy"). For example, if the wind speed is greater than 0 m/s, the observation is classified as "Windy"; otherwise, additional features like rainfall are used to refine the classification.

Random Forest (RF)

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrapped samples from the training data. The final prediction is obtained by aggregating the predictions from all trees (via majority voting for classification or averaging for regression). Formally, the RF prediction for a given input xxx is: $\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$; where $\boldsymbol{T_b(x)}$ is the prediction from the b-th tree, and $\boldsymbol{B}$ is the total number of trees. The advantage of RF lies in its ability to reduce variance and prevent overfitting by combining the output of several weak learners. For wind speed prediction, RF is particularly effective in capturing complex, non-linear interactions between variables like wind speed, temperature, and rainfall.

K-Nearest Neighbors (KNN)

KNN is a non-parametric method where the classification of a new data point is based on the majority class of its K-nearest neighbors in the training set. The distance between the data points is typically measured using Euclidean distance: $d(x, x_i) = \sqrt{\sum_{j=1}^{p}(x_j - x_{ij})^2}$; where x is the new observation, and x_i are the neighbors in the feature space. The predicted class label for x is determined by the majority vote among the K-nearest data points: $\hat{\boldsymbol{y}}(x) = mode(y_{i1}, y_{i2}, \dots, y_{ik})$; where y_{ik} is the class label of the k-th nearest neighbor. In this study, KNN helps classify wind speed conditions based on historical patterns, assigning the observation to "Windy" or "Non-Windy" based on the majority of its neighbors.

Support Vector Machine (SVM)

SVM is a supervised learning algorithm used for classification tasks. Its goal is to find the hyperplane that best separates the data into two classes. The optimal hyperplane is the one that maximizes the margin between the classes: $\max_{w,b} \frac{2}{\|w\|}$ subject to the constraint: $y_i(w^T x_i + b) \geq 1, \forall i$; where w is the normal vector to the hyperplane, b is the offset, and y_i is the class label

("Windy" or "Non-Windy") for observation *i*. SVM uses a subset of the training data, called support vectors, to define the decision boundary, making it effective in high-dimensional spaces.

Linear Discriminant Analysis (LDA)

LDA is a linear classifier that seeks to find a linear combination of features that best separates two or more classes. The goal of LDA is to maximize the ratio of between-class variance to within-class variance: $J(w) = \frac{w^T S_B w}{w^T S_w w}$; where S_B is the between-class scatter matrix, and S_w is the within-class scatter matrix. The optimal w is the eigenvector corresponding to the largest eigenvalue of the matrix ${S_w}^{-1} S_B$.In this study, LDA reduces the dimensionality of the meteorological data (wind speed, temperature, and rainfall) while maximizing the separation between "Windy" and "Non-Windy" classes.

3.4. Model Evaluation Criteria

Accuracy, Sensitivity, Specificity, Precision, and F-measure are well-recognized statistical metrics used to evaluate and make comparisons regarding the predictive capabilities of different models in the realm of forecasting. For comparative analysis and complete review, data from different meteorological stations in Bangladesh are used, and five ML algorithms are applied to make predictions. In this research, we have tried to find the best algorithm that can be used to predict wind speed.

Confusion Matrix

The confusion matrix, often referred to as the error matrix in the context of ML, is a specialized tabular format used to visualize the effectiveness of ML techniques. This matrix was introduced by [37]. As explained by Powers, each row of the matrix corresponds to instances within a true class, while each column corresponds to instances within a predicted class.

Table 2: Confusion Matrix

		Actual Cases	
		Positive	**Negative**
Predicted	Positive	True Positive (TP)	False Positive (FP)
Cases	Negative	False Negative (FN)	True Negative (TN)

Accuracy

The term "Overall Accuracy" indicates the rate of correct predictions made by the classifier. The highest attainable accuracy score is 1.0, while the lowest is 0.0.

$$\text{Overall Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \tag{1}$$

Precision

Precision is a metric used in the evaluation of classification models and diagnostic tests. In essence, precision assesses how well a model distinguishes true positive cases from instances where it incorrectly predicts positive outcomes. A higher precision score signifies a lower rate of false positives, indicating a more reliable positive prediction capability of the model. The following formula formulates precision-

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

Sensitivity

Sensitivity or true positive rate (TPR) is a statistical measure used in binary classification problems to evaluate the ability of a test or model to correctly identify positive instances (or cases) among all actual positive cases. The following formula quantifies it-

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

Specificity

Specificity or true negative rate (TNR) is a statistical measure that assesses the ability of a diagnostic test or classification model to identify negative cases among all actual negative cases accurately. The best specificity is 1.0, whereas the worst is 0.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

F-Measure

One may wonder which model performs better if it has a high precision value but a low recall value. In this situation, the model cannot be uniquely identified by precision and recall value. When dealing with problems of this nature, we employ the F- measure as

$$\text{F} - \text{Measure} = \frac{2*Precision*Recall}{Precision+Recall} \tag{5}$$

4. Results and Discussion

The majority of existing wind speed forecasting methods rely on data from a single location to develop models and predictions. In light of this, our study aimed to identify the most effective model for wind speed prediction in Bangladesh, employing ML classifiers such as LDA, CART, KNN, SVM, and Random Forest. The best-performed model has been judged through accuracy criteria like the Confusion Matrix, Precision, Sensitivity, Specificity, and F-measures evaluated the performances of these classifiers. From Table 3, we can see that the descriptive statistics of the accuracy of all methods are nearly the same. Random Forest contains the highest maximum value, and it shows the largest mean (0.9469) and median (0.9474), whereas the LDA contains the lowest minimum value and has the smallest mean (0.9327) and median (0.9329).

Table 3: Summary of accuracy measurement

Model	Min.	1st Quart.	Median	Mean	3rd Quart.	Max.
CART	0.9341	0.9373	0.9389	0.9387	0.9395	0.9442
SVM	0.9412	0.9414	0.9451	0.9445	0.9460	0.9507
KNN	0.9335	0.9341	0.9344	0.9347	0.9357	0.9359
RF	0.9407	0.9457	0.9474	0.9469	0.9482	0.9519
LDA	0.9323	0.9324	0.9329	0.9327	0.9329	0.9329

We have utilized five ML approaches and assessed the effectiveness of this method by comparing its output with pre-existing categorized data. Table 4 shows that the statistical mean values and overall accuracy metrics indicated that the RF model exhibited superior performance, with an overall accuracy of 94.73% when employed to predict wind speed conditions in Bangladesh. In contrast, the LDA model showed the least favorable performance, 93.27%, compared to the other evaluated models.

Table 4: Different model evaluation performance to predict wind status

Model	Class	Sensitivity/ Recall	Specificity	Precision/Positive Predictive Value	F-Measure	Overall Accuracy
RF	Yes	0.9904	0.3492	0.9547	0.9722	94.73
	No	0.3492	0.9904	0.7252	0.9723	
K-NN	Yes	0.9990	0.0794	0.9377	0.9674	93.72
	No	0.0793	0.9905	0.8571	0.9674	
CART	Yes	0.9956	0.1799	0.9439	0.9691	94.07
	No	0.1798	0.9956	0.7472	0.9690	
SVM	Yes	0.9918	0.2381	0.9475	0.9691	94.11
	No	0.2381	0.9917	0.6766	0.9691	
LDA	Yes	1.0000	0.0000	0.9327	0.9651	93.27
	No	0.0000	1.0000	0.0000	0.9652	

4.1. Model Comparison Using Graphical Methods

The importance of graphical views, such as box and whisker plots, dot plots, density plots, parallel plots, and scatter plot matrix plots, cannot be overstated in data analysis and visualization. These visual representations offer unique insights and advantages that complement traditional numerical metrics. Together, these graphical views empower data analysts to uncover hidden insights, make informed decisions, and communicate findings effectively. The importance of using graphical views over just model fitting accuracy lies in the following aspects: Graphical views provide a visual representation of the model performance, allowing for an intuitive understanding of how the model behaves across different scenarios, allowing for easy comparison between multiple models, can help in identifying the trade-offs between precision and recall, or between sensitivity and specificity more clearly, reveal how a model behaves in different regions of its input space, and show the stability and robustness of a model under varying conditions. Box and Whisker Plots provide a valuable visual representation for assessing the distribution of estimated accuracies, identifying the outliers, and grasping the central tendency across various methods and their interrelationships.

Figure 1 illustrates the comparative accuracy of various machine learning models using Box and Whisker Plots, highlighting Random Forest (RF) as the most reliable model for wind speed prediction in Bangladesh. The RF model demonstrates the highest median accuracy and the narrowest interquartile range (IQR), indicating consistent performance with minimal variability across different trials. In contrast, models like Linear Discriminant Analysis (LDA) show lower median accuracy and greater variability, reflected in their wider IQRs and the presence of outliers. The absence of significant outliers in RF further underscores its robustness and stability, making it the most suitable choice for accurately predicting wind speed in the region's complex and variable climate conditions.

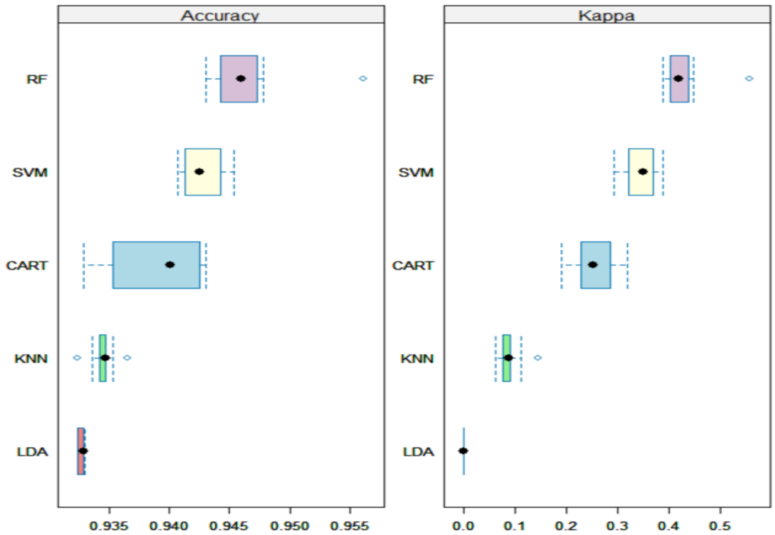

Figure 1: Model accuracy comparison by Box and Whisker Plots

Figure 2 presents a model accuracy comparison using Density Plots, which offers a detailed view of the distribution of accuracy scores for each machine learning model. The Random Forest (RF) model stands out with the highest and most concentrated peak, indicating that it consistently achieves high accuracy across different trials. This concentration highlights RF's ability to handle the complex, non-linear relationships inherent in wind speed data with remarkable precision. In contrast, models like Linear Discriminant Analysis (LDA) exhibit broader and lower peaks, reflecting more variability and less reliable performance. The novelty of this analysis lies in demonstrating RF's superior accuracy and stability in predicting wind speed within Bangladesh's variable climatic conditions, providing a clear statistical advantage over traditional models. The Density Plot's visual representation confirms RF's robustness, making it the preferred model for this critical application.

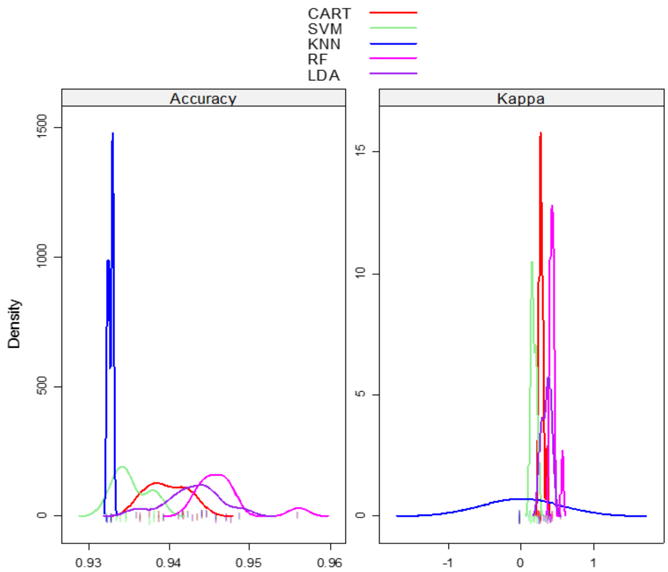

Figure 2: Model accuracy comparison by Density Plots

Figure 3 uses Dot Plots to compare model accuracy, offering a clear visualization of individual accuracy scores across multiple trials for each machine learning model. The Random Forest (RF) model is distinguished by its tight clustering of dots near the upper accuracy range, indicating consistent and high performance. This contrasts with models like Linear Discriminant Analysis (LDA), where the dots are more dispersed, reflecting greater variability and less reliability. The novelty of this analysis lies in the Dot Plot's ability to visually emphasize RF's robustness and superior accuracy in handling the intricate and non-linear patterns of wind speed data in Bangladesh. Statistically, the dense clustering of RF dots underscores its stability, making it the most dependable model for accurate wind speed predictions in the region's complex climate.

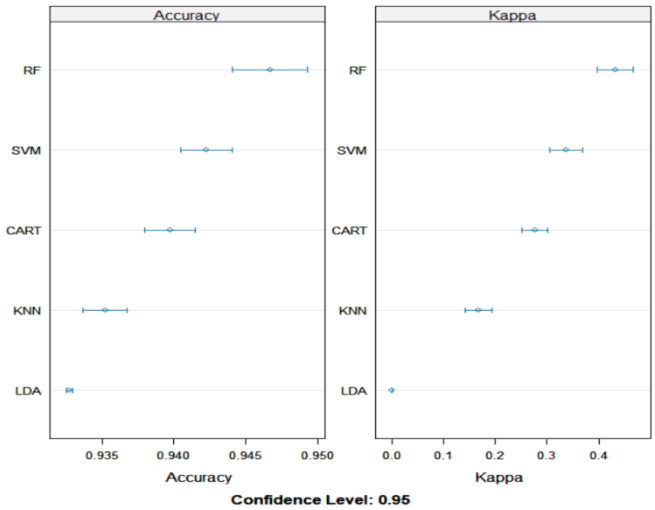

Figure 3: Model accuracy comparison by Dot Plots

In this study, Figure 4 uses Parallel Plots to compare model accuracy, providing a multi-dimensional view of how each machine learning model performs across different trials. The Random Forest (RF) model demonstrates consistent and high accuracy, with its lines closely aligned and relatively stable across the different data subsets, indicating minimal fluctuation in performance. This consistency highlights RF's ability to generalize well across various conditions, a critical factor in the complex and variable climate of Bangladesh. In contrast, other models like Linear Discriminant Analysis (LDA) show more erratic patterns, with lines diverging significantly, reflecting less stable and reliable performance. The novelty of using Parallel Plots in this context lies in their ability to reveal how each model's accuracy varies across multiple trials, statistically confirming RF's superior stability and making it the most reliable model for wind speed prediction in this study.

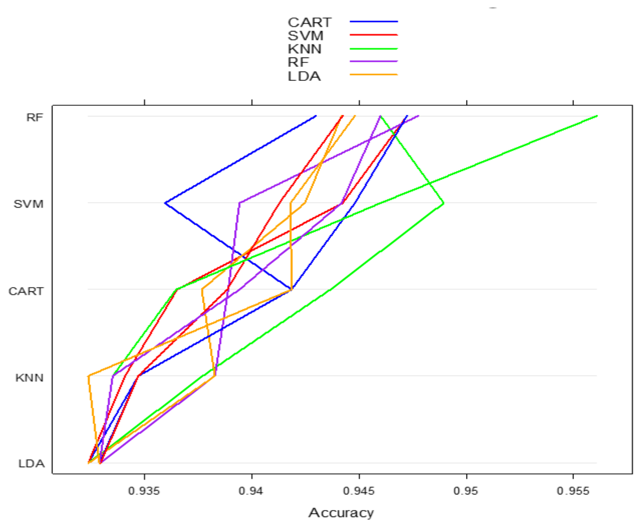

Figure 4: Model accuracy comparison by Parallel Plots.

In the context of the mentioned study, using graphical views for model comparison and evaluation can offer a comprehensive perspective on how the RF algorithm performs in predicting wind speed in Bangladesh. It goes beyond a single accuracy score and provides a richer understanding of the model's strengths and weaknesses in different aspects of forecasting. Therefore, we can infer from Figure 4 that the RF method stands as the preferred choice for forecasting wind speed data in Bangladesh for upcoming time intervals.

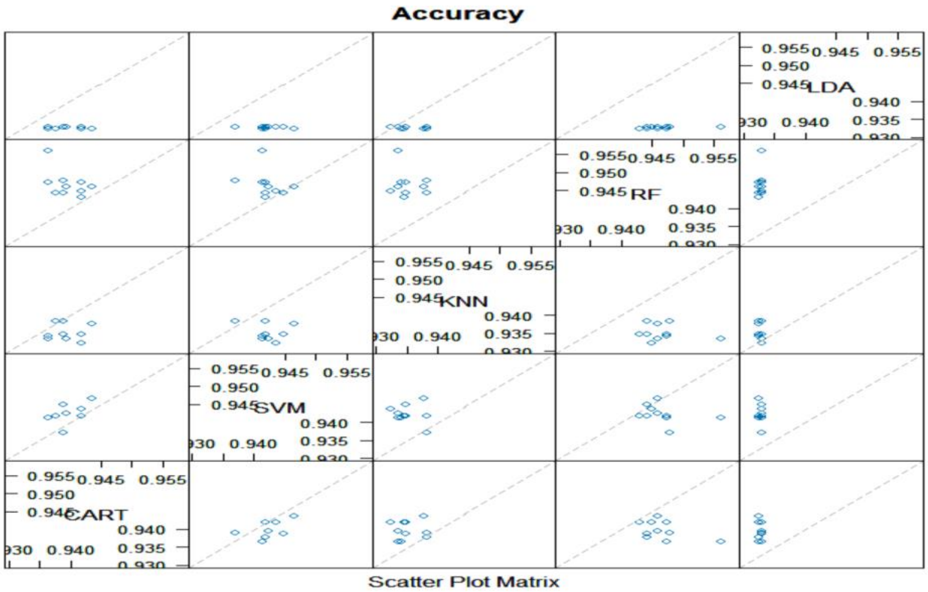

Figure 5: Model accuracy comparison by Scatterplot Matrix Plots

Figure 5 highlights the relationships between predicted and actual wind speed values using a scatterplot matrix across different models, with the Random Forest (RF) model emerging as the superior choice. Statistically, the RF model consistently produces predictions that cluster closely around the ideal prediction line (y = x), reflecting its accuracy across various data subsets. This consistency is vital for reliable wind speed forecasts, particularly in a region like Bangladesh, where weather conditions can vary significantly. The RF model's strength lies in its ability to capture complex, non-linear relationships between variables, as demonstrated by the tight clustering of data points. Furthermore, the RF model effectively reduces overfitting by averaging predictions from multiple decision trees, leading to better generalization of new data. Its robustness in handling noisy data and outliers is evident in the minimal deviations from the ideal prediction line, confirming its reliability. Collectively, these factors underscore the RF model's suitability for wind speed prediction, making it the most reliable choice in this study.

4.2. Benchmarking with Previous Studies

Machine learning techniques demonstrate superior performance in wind speed forecasting and are consequently recommended for use (Rahman et al., 2023). Since the results support all the algorithms and graphically these are proven; so, these findings surpass the achievements of earlier studies. The similarity of conclusion by the RF model has been found for wind speed data as cited in the literature [23], [12], [33] the higher efficiency among the contemporary alternatives and can be considered a feasible method for performing statistical downscaling of temperature data.

(Rahman, 2023) conducted a similar type of research, and their empirical suggestion is that the Random Forest (RF) algorithm emerges as the optimal choice of rainfall prediction for the timeframe specified for Bogura district, Bangladesh. Although they use ML and DL models, they have used the very common classical evaluation metrics such as R^2, RMSE, MSE, and so on to compare the accuracy of the model. They didn't use Accuracy, Sensitivity, Specificity, Precision, and F-measure for accuracy measurement. Actually, the choice of evaluation metrics depends on the nature of the problem. Classification metrics are essential for assessing the performance of models that classify data into different categories, while regression metrics are crucial for evaluating models that predict continuous numeric values. The selection of metrics should align with the specific goals and requirements of the problem at hand. In this investigation, we endeavored to bridge this contemporary gap by conducting a comparative assessment of distinct ML approaches for predicting windspeed conditions in Bangladesh. Table 5 shows the performance of different ML models for different windspeed data. The performance of these models is measured by regression metrics. Table 6 indicates that the best-fitted model is RF by its accuracy, which is 85.66 %. The accuracy of rainfall estimation is above 65% (Zaman, 2018), The accuracy of this study's best-fitted model (RF) is 94.73%. However, the accuracy percentage varies from algorithm to algorithm. The RF method shows the best accuracy for any climatic prediction. Table 6 shows a detailed performance comparison with a study similar to the proposed approach.

Table 5: Performance comparison with similar data

Authors	Models	R^2	MSE	RMSE	MAE
[31]	NARX, NAR, NIO, RNN, CFNN	0.999	$2.6221e^{-6}$	-	-
[22]	RF	-	0.105	0.325	0.251
[23]	MLFFNN, LSTM, RF, SVM-Linear, SVM- Poly	0.968	-	16.358	-
[36]	ARIMA	-	-	3.7647	2.702
[13]	MLP-FF, MLP-GSA, MLP-ABC, MLP- ICA	1.000	0.000	0.007	-

Table 6: Performance comparison with similar existing studies

Author	Data	Model	Overall Accuracy (%)
[12]	Watson Analytics Sales Database	CART, SVM, KNN, RF, NB	55.95
[27]	Malaria Incidence Rate and Climate Data (1990- 2017)	XGBoost, SVM, Naïve Bayes, LR	82.33
[31], [32]	Rainfall data of Bogura district (1971-2015)	CART, LR, SVM, KNN, RF	85.66
Proposed study	Wind speed data (1950- 2018)	CART, SVM, RF, LDA, KNN	94.73

5. Conclusion

The utilization of ML approaches in this research yielded favorable outcomes for both positive and negative wind speed categories, with respect to the evaluated accuracy metrics. As a high-quality accuracy measure, the F-measure, and the mean values of precision and recall have been calculated for the aforementioned models. The empirical findings have demonstrated the incorrigible performance of the RF technique when forecasting wind speed data in the context of Bangladesh. In addition to these model assessment criteria, several other evaluation graphical methods such as Box and Whisker, density, dot, parallel, and scatter plot matrix plots were chosen to assess and compare the model's performance employed in this study. Fortunately, this graphical investigation also supports the RF approach as the most effective and promising method among the five models considered for assessing environmental factors by leveraging insights from historical data. The forecasting performance of the RF model could be good enough for the practitioners of environmental scientists, climatologists, meteorologists, and related researchers on Bangladesh ground from these experiences is the innovation of this study.

Consequently, these outcomes reinforce the validity of our findings. This research will serve as a valuable resource for policymakers, equipping them with important insights to enhance the water management system in specific regions and promote resilient agricultural production. In conclusion, the suggested approach holds the potential to improve predictions of historical wind speeds in nearby areas, thus strengthening forecasts at specific target locations. Moreover, it addresses the limitations of relying solely on data from a single location for wind speed forecasting, opening up a novel avenue for enhancing wind speed predictions.

The present study offers a significant contribution to the field of wind speed prediction by employing machine learning (ML) models, specifically focusing on the climatic conditions of Bangladesh, a region known for its complex and variable weather patterns. The use of Random Forest (RF), which has demonstrated superior accuracy over other models such as Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), underscores the robustness of this approach for predicting wind speed. The novelty of this research lies not only in the application of ML algorithms to a comprehensive dataset spanning from 1950 to 2018 but also in the detailed comparison and validation of model performance using both traditional statistical metrics and advanced graphical methods.

One of the key innovative aspects of this study is the integration of multiple graphical analysis techniques—such as box plots, density plots, dot plots, parallel plots, and scatterplot matrices—which provide a multi-dimensional understanding of model accuracy and robustness. These visual representations go beyond conventional accuracy metrics, offering deeper insights into the stability and reliability of each model in capturing the nonlinear and intricate relationships inherent in wind speed data.

Furthermore, this study addresses a notable gap in the literature by applying contemporary ML techniques to a region that has received limited attention in previous research, thereby contributing valuable knowledge to the field of environmental modeling. The findings not only reinforce the efficacy of the RF model in this context but also provide actionable insights for policymakers, environmental scientists, and other stakeholders concerned with climate resilience and disaster preparedness in Bangladesh.

By successfully demonstrating the applicability of advanced ML models for wind speed prediction in a region characterized by high climatic variability, this study paves the way for future research that could further enhance predictive accuracy. The exploration of deep learning models, as well as the integration of atmospheric image analysis, are promising avenues for future work that could build upon the foundations laid by this research. This study thus offers a novel approach to improving wind speed predictions and contributes to the broader goal of enhancing climate adaptation strategies in vulnerable regions.

Future Study and Policy Implications

The findings, namely, the best-predicted model, will help policymakers, related researchers, and concerned authorities take necessary steps in advance to protect against climatic hazards and mitigate the impact of high wind speeds and resulting calamities. This research can be extended in the future, considering deep learning. Besides these, atmospheric image analysis can be considered a prospective future research concern.

Abbreviations

Abbreviation	Full Form
ML	Machine Learning
LDA	Linear Discriminant Analysis
CART	Classification and Regression Trees
RF	Random Forest
KNN	K Nearest Neighbors
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
MWS	Mean Wind Speed
ANN	Artificial Neural Network
SVR	Support Vector Regression
MSE	Mean Square Error
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
ARIMA	Auto Regressive Integrated Moving Average
SARIMA	Seasonal Auto Regressive Integrated Moving Average
MLR	Multilinear Regression
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
KNNR	K-Nearest Neighbors Regression
GMDH	Group Method of Data Handling
CFNN	Cascade Forward Neural Network
DL	Deep Learning
MLFFNN	Multi-Layer Feed-Forward Neural Network
ANFIS	Adaptive Neuro-Fuzzy Inference System
SPNN	Smooth Pinball Neural Network
BARC	Bangladesh Agricultural Research Centre

Data Availability

The research was carried out using data from the Bangladesh Agricultural Research Council (BARC). This is secondary data covering the period 1950 to 2018. The data are available on the BARC website at the link: http://barcapps.gov.bd/climate/wind.

Funding: No fund was received for this research.

Conflicts of Interest: The authors declare that they have no conflict of competing interests.

Acknowledgments: The authors thank the editor and the reviewer for providing useful comments and suggestions on the earlier version of the manuscript.

References

[1] Abuayyash, K. Alsamamra, H., Teir, M., and Doufesh, H. (2024). Wind Speed Prediction in Jerusalem Using Machine Learning Algorithms: A Case Study of Using ANFIS and KNNR. American Journal of Modern Energy, 10(2), 25–37. https://doi.org/10.11648/j.ajme.20241002.12

[2] Alsubaie, F., and Aldoukhi, M. (2024). Using machine learning algorithms with improved accuracy to analyze and predict employee attrition. Decision Science Letters, 13(1), 1–18. https://doi.org/10.5267/dsl.2023.12.006

[3] Anas Faskari, S., Faskari Shehu, A., Asi, T., and Belgore, au. (2023). Machine Learning Approach to Wind Speed Prediction using Soft Computing Tools. JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION, 11(2). www.atbuftejoste.net

[4] Anggraini, T. S., Irie, H., Sakti, A. D., and Wikantika, K. (2023). Novel Global Air Quality Index Development Through Integration of Remote Sensing Data, Ground-Based Stations, and Machine Learning Techniques. Environmental Advances, 15(August 2023), 100456. https://doi.org/10.1016/j.envadv.2023.100456

[5] Antor, A. F., and Wollega, E. D. (2020). Comparison of machine learning algorithms for wind speed prediction. Proceedings of the International Conference on Industrial Engineering and Operations Management, August.

[6] Arends, E. L., Watson, S. J., Basu, S., and Cheneka, B. (2020). Probabilistic wind power forecasting combining deep learning architectures. 2020 17th International Conference on the European Energy Market (EEM), 1–6. https://doi.org/10.1109/EEM49802.2020.9221929

[7] Bhuiyan, M. R. A., Mamur, H., and Begum, J. (2021). A brief review on renewable and sustainable energy resources in Bangladesh. Cleaner Engineering and Technology, 4, 100208. https://doi.org/10.1016/j.clet.2021.100208

[8] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[9] Burney, S. M. A., Drakhshan, K., and Karim, S. (2023). Forecasting Wind Speed Using Machine Learning ANN Models at 4 Distinct Heights at Different Potential Locations in Pakistan. WSEAS TRANSACTIONS ON COMPUTERS, 22, 127–141. https://doi.org/10.37394/23205.2023.22.16

[10] Colone, L., Dimitrov, N., and Straub, D. (2019). Predictive repair scheduling of wind turbine drive-train components based on machine learning. Wind Energy, 22(9), 1230–1242. https://doi.org/10.1002/we.2352

[11] Guenoukpati, A., Salami, A. A., Kodjo, M. K., and Napo, K. (2020). Estimating Weibull Parameters for Wind Energy Applications using Seven Numerical Methods: Case studies of three costal sites in West Africa. International Journal of Renewable Energy Development, 9(2), 217–226. https://doi.org/10.14710/ijred.9.2.217-226

[12] Gupta, S., Saluja, K., Goyal, A., Vajpayee, A., and Tiwari, V. (2022). Comparing the performance of machine learning algorithms using estimated accuracy. Measurement: Sensors, 24(September), 100432. https://doi.org/10.1016/j.measen.2022.100432

[13] Hazbeh, O., Aghdam, S. K. ye, Ghorbani, H., Mohamadian, N., Ahmadi Alvar, M., and Moghadasi, J. (2021). Comparison of accuracy and computational performance between the machine learning algorithms for rate of penetration in directional drilling well. Petroleum Research, 6(3), 271–282. https://doi.org/10.1016/j.ptlrs.2021.02.004

[14] Hossain, M. A., Chakrabortty, R. K., Elsawah, S., and Ryan, M. J. (2021). Very short-term forecasting of wind power generation using hybrid deep learning model. Journal of Cleaner Production, 296, 126564. https://doi.org/10.1016/j.jclepro.2021.126564

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

[16] Kadriye Duden, F. (2023). Wind Speed Prediction Using Data Mining Approches: A Case Study of Gokceada. Turkey. Scientific Journal of Mehmet Akif Ersoy University, 6(2), 19–23. https://dergipark.org.tr/pub/sjmakeu

[17] Kara, T., and Şahin, A. D. (2023). Implications of Climate Change on Wind Energy Potential. Sustainability, 15(20), 14822. https://doi.org/10.3390/su152014822

[18] Karaman, Ö. A. (2023b). Prediction of Wind Power with Machine Learning Models. Applied Sciences, 13(20), 11455. https://doi.org/10.3390/app132011455

[19] Lawrence, J., Blackett, P., and Cradock-Henry, N. A. (2020). Cascading climate change impacts and implications. Climate Risk Management, 29, 100234. https://doi.org/10.1016/j.crm.2020.100234

[20] L'Heureux, A., Grolinger, K., Elyamany, H. F., and Capretz, M. A. M. (2017). Machine Learning with Big Data: Challenges and Approaches. In IEEE Access (Vol. 5, pp. 7776–7797). https://doi.org/10.1109/ACCESS.2017.2696365

[21] Mandal, N., and Sarode, T. (2020). Prediction of Wind Speed using Machine Learning. International Journal of Computer Applications, 176(32), 34–37. https://doi.org/10.5120/ijca2020920370

[22] Masood, A., and Ahmad, K. (2023). Data-driven predictive modeling of PM2.5 concentrations using machine learning and deep learning techniques: a case study of Delhi, India. Environmental Monitoring and Assessment, 195(1). https://doi.org/10.1007/s10661-022-10603-w

[23] Mirzaei, S., Vafakhah, M., Pradhan, B., and Alavi, S. J. (2021). Flood susceptibility assessment using extreme gradient boosting (EGB), Iran. Earth Science Informatics, 14(1), 51–67. https://doi.org/10.1007/s12145-020-00530-0

[24] Mollick, T., Hashmi, G., and Sabuj, S. R. (2024). Wind speed prediction for site selection and reliable operation of wind power plants in coastal regions using machine learning algorithm variants. Sustainable Energy Research, 11(1), 5. https://doi.org/10.1186/s40807-024-00098-z

[25] Nikitas, G., Bhattacharya, S., and Vimalan, N. (2020). Wind energy. Future Energy: Improved, Sustainable and Clean Options for Our Planet, 331–355. https://doi.org/10.1016/B978-0-08-102886-5.00016-5

[26] Nkiruka, O., Prasad, R., and Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. Informatics in Medicine Unlocked, 22, 100508. https://doi.org/10.1016/j.imu.2020.100508

[27] Nurunnahar, S., Talukdar, D. B., Rasel, R. I., and Sultana, N. (2017). A short term wind speed forcasting using SVR and BP-ANN: A comparative analysis. In 20th International Conference of Computer and Information Technology, ICCIT 2017 (Vols 2018-Janua, pp. 1–6). https://doi.org/10.1109/ICCITECHN.2017.8281802

[28] Olaiya, F., and Adeyemo, A. B. (2012). Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. International Journal of Information Engineering and Electronic Business, 4(1), 51–59. https://doi.org/10.5815/ijieeb.2012.01.07

[29] Ortiz-García, E. G., Salcedo-Sanz, S., Pérez-Bellido, Á. M., Gascón-Moreno, J., Portilla-Figueras, J. A., and Prieto, L. (2011). Short-term wind speed prediction in wind farms based on banks of support vector machines. Wind Energy, 14(2), 193–207. https://doi.org/10.1002/we.411

[30] Rahman, M. M., Shakeri, M., Khatun, F., Tiong, S. K., Alkahtani, A. A., Samsudin, N. A., Amin, N., Pasupuleti, J., and Hasan, M. K. (2023). A comprehensive study and performance analysis of deep neural network-based approaches in wind time-series forecasting. Journal of Reliable Intelligent Environments, 9(2), 183–200. https://doi.org/10.1007/s40860-021-00166-x

[31] Rahman, M. S. (2023). Prediction Ranifall Using Machine Learning Algorithms : Empirical Evidence from Bogura and Rangpur District of Bangladesh Md. Rafiqul Islam, Md. Mostafizur Rahman, Afsana Afruz, Md. Abdul Data Mining and Environmental Research

Group, Department of. 23(2), 47–62.
[32] Rahman, Md. M., Khalek, Md. A., and Rahman, M. S. (2021). Performance of Different Data Mining Methods for Predicting Rainfall of Rajshahi District, Bangladesh. Data Science and SDGs, 67–78. https://doi.org/10.1007/978-981-16-1919-9_6
[33] Shawon, S. M. R. H., Saaklayen, M. A., and Liang, X. (2021a). Wind Speed Forecasting by Conventional Statistical Methods and Machine Learning Techniques. 2021 IEEE Electrical Power and Energy Conference, EPEC 2021, 304–309. https://doi.org/10.1109/EPEC52095.2021.9621686
[34] Shawon, S. M. R. H., Saaklayen, M. A., and Liang, X. (2021b). Wind Speed Forecasting by Conventional Statistical Methods and Machine Learning Techniques. In 2021 IEEE Electrical Power and Energy Conference, EPEC 2021 (pp. 304–309). https://doi.org/10.1109/EPEC52095.2021.9621686
[35] Shi, J., Guo, J., and Zheng, S. (2012). Evaluation of hybrid forecasting approaches for wind speed and power generation time series. Renewable and Sustainable Energy Reviews, 16(5), 3471–3480. https://doi.org/10.1016/j.rser.2012.02.044
[36] Sri Preethaa, K. R., Muthuramalingam, A., Natarajan, Y., Wadhwa, G., and Ali, A. A. Y. (2023). A Comprehensive Review on Machine Learning Techniques for Forecasting Wind Flow Pattern. Sustainability (Switzerland), 15(17). https://doi.org/10.3390/su151712914
[37] Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment, 62(1), 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7
[38] Trenberth, K. E., Dai, A., Van Der Schrier, G., Jones, P. D., Barichivich, J., Briffa, K. R., and Sheffield, J. (2014). Global warming and changes in drought. Nature Climate Change, 4(1), 17–22. https://doi.org/10.1038/nclimate2067
[39] Türkan, Y. S., Yumurtacı Aydoğmuş, H., and Erdal, H. (2016). The prediction of the wind speed at different heights by machine learning methods. An International Journal of Optimization and Control: Theories and Applications (IJOCTA), 6(2), 179–187. https://doi.org/10.11121/ijocta.01.2016.00315
[40] Tyass, I., Khalili, T., Rafik, M., Abdelouahed, B., Raihani, A., and Mansouri, K. (2023). Wind Speed Prediction Based on Statistical and Deep Learning Models. International Journal of Renewable Energy Development, 12(2), 288–299. https://doi.org/10.14710/ijred.2023.48672
[41] Vågerö, O., Bråte, A., Wittemann, A., Robinson, J. Y., Zeyringer, M., and Sirotko-Sibirskaya, N. (2024). Machine learning of public sentiments towards wind energy in Norway. Wind Energy. https://doi.org/10.1002/we.2902
[42] Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10(5), 988–999. https://doi.org/10.1109/72.788640
[43] Verma, M., and Ghritlahre, H. K. (2023). Forecasting of Wind Speed by Using Three Different Techniques of Prediction Models. Annals of Data Science, 10(3), 679–711. https://doi.org/10.1007/s40745-021-00333-0
[44] Verma, M., Ghritlahre, H. K., and Chandrakar, G. (2023). Wind Speed Prediction of Central Region of Chhattisgarh (India) Using Artificial Neural Network and Multiple Linear Regression Technique: A Comparative Study. Annals of Data Science, 10(4), 851–873. https://doi.org/10.1007/s40745-021-00332-1
[45] Zhang, Y., Zhang, W., Guo, Z., and Zhang, S. (2022). An effective wind speed prediction model combining secondary decomposition and regularised extreme learning machine optimised by cuckoo search algorithm. Wind Energy, 25(8), 1406–1433. https://doi.org/10.1002/we.2737.