Sustainable Energy Research

RESEARCH Open Access

Wind speed prediction for site selection and reliable operation of wind power plants in coastal regions using machine learning algorithm variants

Tajrian Mollick[1], Galib Hashmi[2] and Saifur Rahman Sabuj[1*]

Abstract

The challenge of predicting wind speeds to facilitate site selection and the consistent operation of wind power plants in coastal regions is a global concern. The output of wind turbines is subject to fluctuations corresponding to changes in wind speed. The unpredictable characteristics of wind patterns introduce vulnerabilities to wind power facilities in wind power plants. To address this unpredictability, an effective strategy involves forecasting wind speeds at specific locations during wind power plant operations. While previous research has explored various machine learning algorithms to tackle these issues, satisfactory results have not been achieved, and Bangladesh faces challenges in this regard, especially in low-wind speed areas. This study aims to identify the most accurate machine learning-based algorithm to forecast the short-term wind speed of two areas (Kutubdia and Cox's Bazar) located on the eastern coast of Bangladesh. Wind speed data for a span of 21.5 years, ranging from January 2001 to June 2022, were sourced from two outlets: the Bangladesh Meteorological Department and the website of NASA. Wind speed has been forecasted using 14 different regression-based machine learning models with a comprehensive overview. The results of the experiment highlight the exceptional predictive performance of a boosting-based ensemble method known as categorical boosting, especially in the context of forecasting wind speed data obtained from NASA. Based on the testing data, the evaluation yields remarkable results, with coefficients of determination measuring 0.8621 and 0.8758 for wind speed in Kutubdia and Cox's Bazar, respectively. The study underscores the critical importance of prioritizing optimal turbine site selection in the context of wind power facilities in Bangladesh. This approach can yield benefits for stakeholders, including engineers and project owners associated with wind projects.

Keywords Machine learning, Regression, Wind power plant, Wind speed forecasting

Introduction

Rapid economic growth and improved lifestyles have increased human energy consumption. However, reliance on conventional fossil fuels like natural gas, coal, and oil results in pollution and contributes to global warming. As these resources are non-renewable and finite, nations increasingly invest in renewable energy sources to meet their present and future needs. Wind energy, being readily available and pollution-free, has emerged as a prominent renewable energy solution (Anjum, 2014; Bharani & Sivaprakasam, 2022). Therefore, wind power plants are rapidly evolving globally to address the growing demand for cleaner and more sustainable power. In the last 20 years, there has been a rapid growth in the installed capacity of wind power, as depicted in Fig. 1, which

*Correspondence:
Saifur Rahman Sabuj
s.r.sabuj@ieee.org
[1] Department of Electrical and Electronic Engineering, Brac University, Dhaka 1212, Bangladesh
[2] Institute of Energy, University of Dhaka, Dhaka 1000, Bangladesh

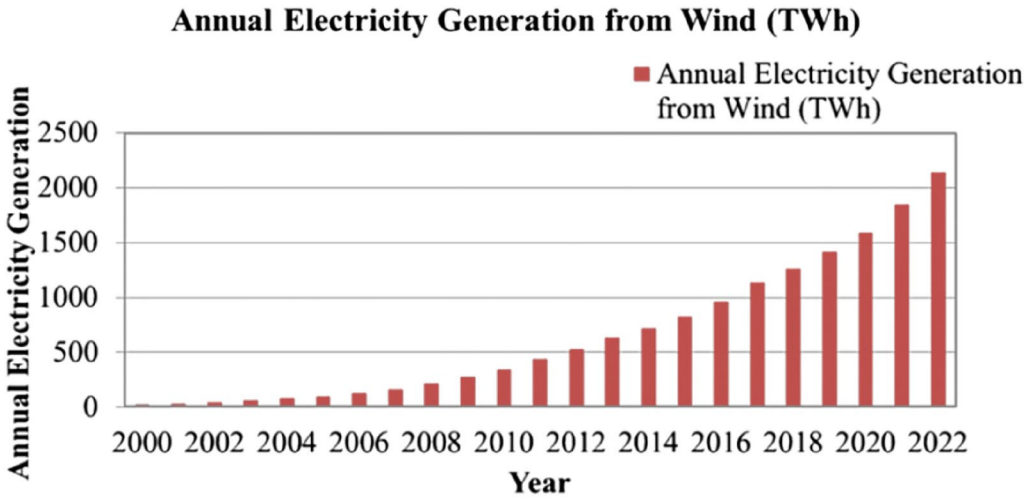


Fig. 1 Annual electricity generation from wind (TWh)

showcases global yearly wind power generation. It is assumed that wind-generated power will top the renewable energy sector by producing around 7932.5 TWh of electricity in 2030 (Iea, 2023). Currently, appropriate actions are being taken in several nations. However, for many countries, like Bangladesh, the contribution of wind power is quite minor.

In 2041, Bangladesh aims to achieve high-income country status, emphasizing the need for sustainable and uninterrupted power supply to drive industrialization. With a forecasted electricity demand of 82,292 MW in 2041, the country faces challenges due to depleting natural gas reserves and dependency on imported fuels. The current energy mix relies heavily on natural gas, and the depletion of reserves by 2028 poses a threat (Babu et al., 2022). Diesel imports for power plants and nuclear power plant limitations further complicate the quest for self-sufficiency. Moreover, Bangladesh, minimizing emissions of greenhouse gases by 21.85% by 2030, faces the dual challenge of increasing energy consumption and decreasing CO_2 emissions to achieve Sustainable Development Goals (SDGs) by 2030 and advanced nation status by 2041 (Das et al., 2020). The current energy mix of Bangladesh is natural gas 64.36%, furnace oil 21%, coal 33.54%, coal 9.52%, solar 0.84%, hydro 1.25%, and wind 0.01% ("Share of primary energy from wind" & Our World in Data, 2023). Embracing renewable energy practices becomes crucial for efficient energy utilization and environmental sustainability. The United States Agency for International Development (USAID), Bangladesh, and the Government of Bangladesh (GoB) collaborated to assist the National Renewable Energy Laboratory (NREL) to conduct a recent national wind resource assessment in Bangladesh. (Babu et al., 2022). According to the evaluation document of NERL, Bangladesh has more than 20,000 km^2 of land with a wind speed of 5.75–7.75 m/s, which leads to a gross wind potential of over 30,000 MW (Siddique et al., 2021). The findings prove that the entire coastline area, e.g., Cox's Bazar, Patenga, Teknaf, Kutubdia, Char Fassion, and Kuakata, falls into the zone that is commercially important for the production of wind power by installing small and medium-scale wind farms.

Therefore, it can be said that if the right laws, programs, and technological innovations are implemented, wind can be included as a key contributor to renewable energies to tackle the energy crisis (Siami-Namini et al., 2018). However, wind energy is an intermittent renewable energy source (IRES) because it cannot be dispatched due to its fluctuating nature. Forecasting the wind speed of a location before constructing a wind power plant may be the answer to the unpredictability of wind speed. Moreover, accurate wind speed predictions during the operation of the wind could aid stakeholders in making vital decisions, such as regarding wind power storage or grid transmission activity (Shi et al., 2022). Thus, to identify optimal sites for wind energy plants and guarantee operational safety, researchers concentrate on developing precise predictions of wind speed (Babu et al., 2022).

A thorough study of the literature shows that there are two basic approaches for wind speed forecasting: the time horizon and modeling theory (as depicted in Fig. 2). Four sorts of wind speed predictions are possible in terms of time horizon, and they are as follows: very short-time (a few seconds), short-time (30 min–6 h), medium-time (6 h–1 day), and long-time (more than 1 day) (Babu et al., 2022). Operational engineers, armed with predictions of wind speed from the short term up to the long term in

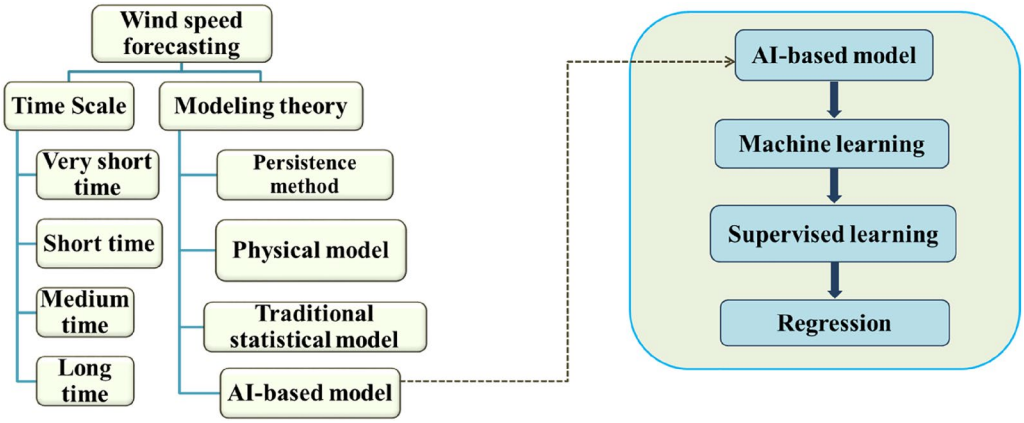


Fig. 2 Wind speed forecasting and the ML algorithm used in this study

advance, can make a variety of decisions to optimize the performance and efficiency of wind energy operations. They can strategically optimize wind turbine operations by altering angles and speeds for maximal energy capture based on wind speed estimates available three hours in advance. They use energy storage based on anticipated wind conditions, distribute resources wisely, and effectively integrate wind energy into the power grid (Yousuf et al., 2019). Anticipated variations in energy production inform financial planning, while safety protocols are implemented in advance of extreme weather. To guarantee the efficient and secure operation of wind energy systems, engineers also plan grid connections and implement environmental impact mitigation strategies during certain wind conditions (Santhosh et al., 2020; Yousuf et al., 2019).

Similar to the time horizon, modeling theory is classified into four types of forecasting models: persistence methods, physical models, conventional statistical models, and models based on artificial intelligence (AI) (Chang, 2014). The persistence method seems to be more accurate than other wind forecasting techniques in very short-term forecasting. However, as the prediction horizon expands, the persistence method's accuracy will rapidly decline. Physical models are good for long-term forecasting, but they are time-consuming due to the numerous computations required. Statistical models are used to ascertain the mathematical relationship between inputs and outputs under the assumption of linear correlations. Despite their extensive use in the research, their effectiveness fell short of expectations because they were ineffective in identifying nonlinear interactions (Chang, 2014). A large subset of AI is machine learning (ML), which aims to train the computer to comprehend situations and perform actions that are both advantageous and beneficial to the environment after training it on a previously stated dataset (Jagdale et al., 2022). An examination of existing literature reveals that ML algorithms can be categorized into supervised, unsupervised, semi-supervised, and reinforcement learning categories (Sarker, 2021). A supervised learning algorithm determines a mapping function to map the input variable to the output variable. If a hidden layer is used by the mapping function, then it becomes deep learning (DL), a subclass of ML that can intelligently evaluate data on a large scale (Babu et al., 2022). ML and DL have been widely employed in the field of prediction because of their superior prediction capability over conventional prediction models (Tarek et al., 2023).

Wind speed forecasting can be performed using the following ML algorithms following a detailed investigation of the literature: multiple linear regression (MLR), support vector regression (SVR), lasso regression, ridge regression, random forest (RF), light gradient boosting machine (LightGBM), extreme gradient boost (XGBoost), and long short-term memory networks (LSTM) (Elsaraiti & Merabet, 2021; Hanoon et al., 2022; Krishnaveni et al., 2021; Malakouti, 2023; Mohsin et al., 2021; Salah et al., 2022; Senthil Kumar P, 2019; Shawon et al., 2021; Xie et al., 2021). Air pressure, temperature, humidity, and wind speed were implemented as input variables in the proposed models. Numerous studies pointed out that multi-variable long short-term memory network model (MV-LSTM) methodology is more effective than techniques like autoregressive moving average (ARMA) and single-variable LSTM (Elsaraiti & Merabet., 2021). Additionally, different ML techniques, including bagged regression trees (BTs), SVR, and Gaussian process regression (GPR) were adapted by many reviewers in terms of the weekly prediction of wind speed (Hanoon et al., 2022). A variety of ML methods, such as MLR,

ridge, lasso, RF, SVR, and LSTM, were applied in a different study to predict wind speed for a specific weather station. These models incorporated wind direction, temperature, pressure, timestamp, and other variables for precise estimation. Notably, the RF and LSTM-RNN models outperformed other approaches for accurately wind speed forecasting (Salah et al., 2022). To anticipate short-term wind speed at certain ground observation stations, a MV-LSTM was also evolved in a different study (Xie et al., 2021). ML models were also deployed in the study to forecast wind speed and electricity generation in a SCADA system. Six techniques, including adaptive boosting (AdaBoost) and LightGBM, were applied in Malakouti (2023). Outcomes achieved from the ensemble technique with cross-validation were promising: the wind power and wind speed predictions had root mean square errors (RMSEs) of 11.78 and 0.2080, respectively. Although several studies have successfully used a variety of ML models to anticipate wind speed, there is still unexposed potential to attain the best outcomes considering wind's inconstancy. Previous studies have mostly concentrated on certain areas and used a limited set of ML models. There is a distinct need for more research to better encompass how these models perform in a wider range of geographic contexts, such as areas with changing climates or opposing weather patterns. Moreover, earlier examinations were inconclusive in considering important factors associated with site and turbine selection, leaving a substantial gap in addressing this crucial part of the study's goals.

The wind energy initiatives in Bangladesh have been predominantly concentrated in specific regions, leaving a significant portion of the country unexplored in terms of wind energy projects. Limited studies on short-term wind speed forecasts have been conducted in Bangladesh, hindering the effective communication of mitigation and adaptation strategies to project stakeholders. This research addresses the knowledge gap by focusing on Kutubdia and Cox's Bazar, situated in the southeastern region of Bangladesh, known for their favorable wind potential. In this study, the research employs fourteen well-established ML models to forecast 3-h interval wind speed, utilizing a 21.5-year weather dataset from Bangladesh Meteorological Department (BMD) and NASA's website. The urban environment of BMD suggests a relatively low wind potential, prompting the utilization of the NASA's dataset, which reveals preferable wind energy availability. The application of diverse ML models enhances the accuracy of wind speed predictions, offering valuable insights for site and turbine selection, operational safety measures, and the uninterrupted performance of wind power systems.

System model

The comprehensive methodology employed for this research is fully depicted in Fig. 3. The six fundamental stages of this procedure are succinctly outlined below:

- **Step 1. Data collection and formatting:** Initially, the observed data (wind speed, wind direction, temperature, humidity, and pressure) of the two coastal areas with a 3-h interval from January 1, 2001, to June 30, 2022 (62,808 data samples of 21.5 years) have been collected from two sources: i) BMD (Kutubdia and Cox's Bazar weather stations) and ii) the website of NASA (Data Access Viewer) ("POWER | Data Access Viewer", 2023). Data formatting is done by removing irrelevant data and rearranging the required parts.
- **Step 2. Exploratory data analysis:** Conducting exploratory data analysis aids in gaining a deeper understanding of the data's underlying patterns. It is fundamental to the structure of any machine-learning algorithm. In this part, descriptive statistics are analyzed to extract knowledge from the formatted data.
- **Step 3. Data preprocessing:** Before applying the ML models, data preprocessing is an essential stage in shaping an optimal data structure. In contrast, the absence of well-preprocessed data can compromise the efficiency and performance of machine-learning models, resulting in suboptimal outcomes. This preprocessing phase covers tasks such as handling missing values, extracting and selecting features, and normalizing data.
- **Step 4. Train-test splitting:** Following preprocessing, the dataset is partitioned into three subsets: i) the training set (70%), ii) the validation set (15%), and iii) the test set (15%).
- **Step 5. Model optimization and training:** In this stage, 14 distinct regression-based ML methods, including MLR, Lasso, Ridge, Elastic Net, KNN, DT, GBR, RF, XGBoost, LightGBM, CatBoost, LSTM, and GRU, are deployed to predict the wind speed three hours ahead. For model optimization, k-fold cross-validation is implemented with and without parameter tuning.
- **Step 6. Forecasting and performance evaluation:** The model, which has been trained on the validation dataset, is assessed, and its performance is contrasted with that of the initial model trained on the training dataset. If the disparity is minimal, the forecasting performance using the test dataset is cross-checked with the observed data to ascertain the system's accuracy in construction. A comprehensive assessment of wind resources has been carried out using both

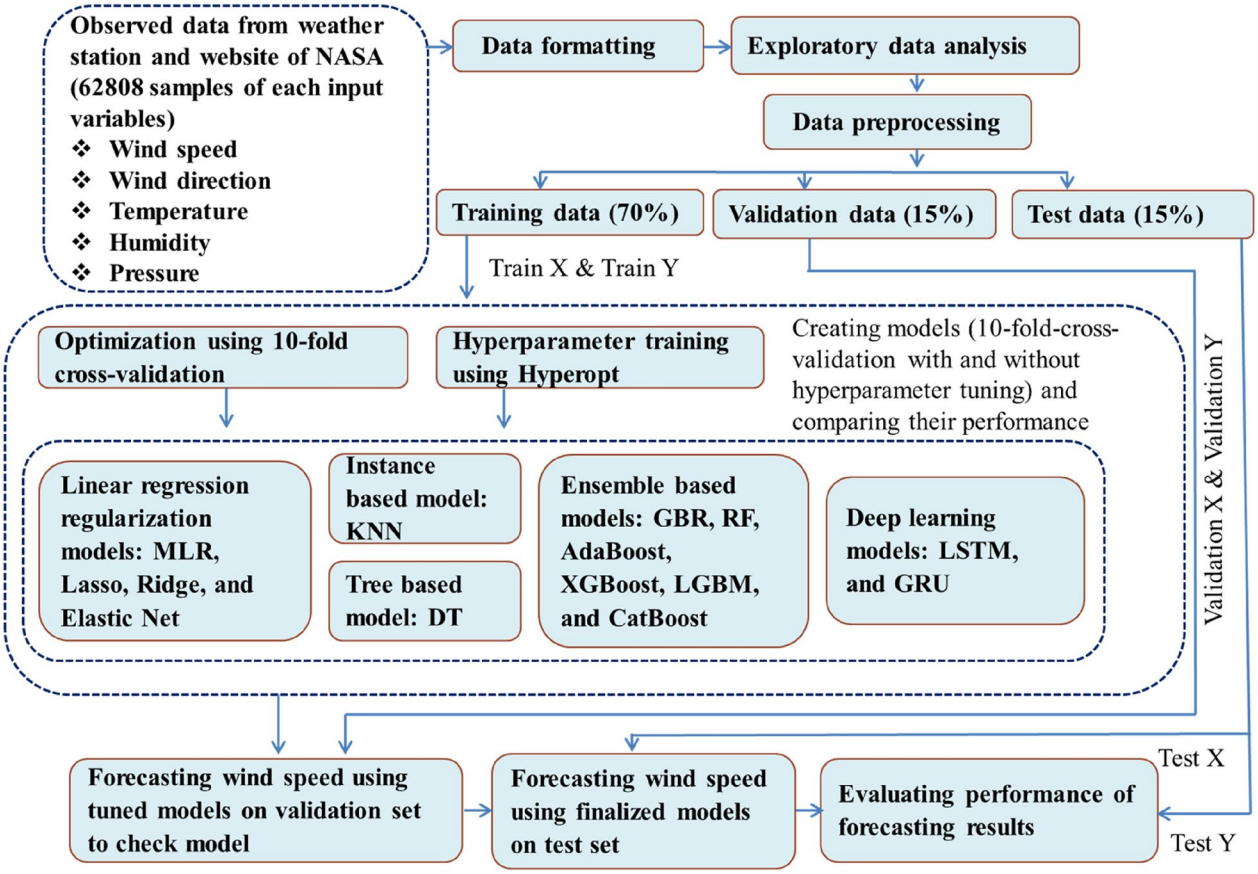


Fig. 3 A detailed framework of the regression models for wind speed forecasting

observed and predicted wind speeds, demonstrating the detailed advantages of forecasting.

Training models: renowned predictive algorithms

Predictive ML models based on regression present a versatile range of techniques for forecasting wind speed, each possessing unique strengths and suitability for different contexts. The choice of a model involves different aspects, including the properties of the data, the computing capacity, and the specific requirements of the forecasting goal. A brief synopsis of the models used in this study is provided in Table 1.

Optimizing and fine-tuning models: K-fold cross-validation and Hyperopt

Optimization and hyperparameter adjustments can greatly enhance the performance of regression models and their ability to generalize to new data. One common method for determining how well a ML model performs is k-fold cross-validation. The training data are split into k-folds, or subsets, to apply this technique. The model is then repeatedly trained and evaluated on these folds. Each fold served as the training set while the others provided the validation set in turn. Still, the open-source software Hyperopt finds the optimal values for these parameters using a Bayesian approach. It defines a hyperparameter search space and effectively navigates it with optimization techniques. Different hyperparameter combinations are explored and model performance is evaluated (Hutter et al., 2019). This research applied Hyperopt to fine-tune the hyperparameters that yield superior performance on a training dataset.

Comparing model performance: different evaluation metrics

To evaluate each model's efficacy, mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) are determined. The average squared difference between observed and predicted values is estimated by MSE. It is employed to assess the degree of inaccuracy in statistical models. The lower the MSE the better model fits a dataset. The average absolute difference between the observed and

Table 1 Details of regression models utilized in this research

Regression model	Full form	Main features
MLR	Multiple linear regression	- Explores linear correlations between input variables (Salah et al., 2022)
Lasso	Least absolute shrinkage and selection operator	- Assigns one of the correlated predictors an elevated weight while minimizing the other correlated predictors to almost zero - Imposes a penalty on the total absolute values of the coefficients, named L1 penalty (Salah et al., 2022)
Ridge	Regularized inverse depth generating estimators	- Assigns weights that are similar to correlated predictors - Imposes an L2 penalty to the total squared values of the coefficients (Salah et al., 2022)
Elastic Net	Elastic net regression	- Handles collinear data and prevent overfitting - Combines components of both Lasso (L1) and Ridge (L2) regularization approaches (Malakouti, 2023)
KNN	K-nearest neighbors	- Makes predictions based on the average of the k-nearest neighbors' majority vote for a given data point - Applies a distance metric (Euclidean distance) which defines "nearest" (Tarek et al., 2023)
DT	Decision tree	- Identifies the greatest feature and split point at each node using mean squared error (MSE) - Makes judgments by recursively separating the data based on features (Talekar, 2020)
RF	Random forest	- Combines multiple decision trees to improve prediction accuracy as an ensemble learning method - Entails averaging the predictions made by several trees after they have been trained on arbitrary subsets of the data (Talekar, 2020)
GBR	Gradient boosting regression	- Reduces the loss function using gradient descent optimization - Utilizes the decision trees which are shallow and have little depth, as weak learners (Tarek et al., 2023)
AdaBoost	Adaptive boosting	- Creates a series of weak learners, which are usually shallow decision trees, and evaluates their performance using an exponential loss function (Jasman et al., 2022)
XGBoost	Extreme gradient boosting	- Well-known for its rapidity and effectiveness as a powerful gradient boosting algorithm - Creates a strong regression model by building a sequence of decision trees one after the other, each one fixing the mistakes of the previous one ("POWER | Data Access Viewer", 2023)
LightGBM	Light gradient boosting machine	- Well-recognized for its exceptional performance as a gradient boosting framework - Especially effective with the large datasets and provides quicker training times without sacrificing its remarkable accuracy in regression tasks (Malakouti, 2023)
CatBoost	Categorical boosting	- Combines dynamic learning rates, ordered boosting, and oblivious trees as an advanced gradient boost technique (Jasman et al., 2022) - Gains popularity, especially when working with complex datasets that contain categorical categories
LSTM	Hyperopt	- Understands relationships in sequential data as a type of recurrent neural network (RNN) - Excels at modeling sequences where long-term context is crucial because of its ability to store and propagate information over extended periods of time (Elsaraiti & Merabet, 2021)
GRU	Gated recurrent unit	- Develops relationships in sequential data as a type of recurrent neural network (RNN) - Detects long-term dependencies in sequential data while mitigating the problem of vanishing gradients that conventional RNNs encounter (Tao et al., 2022)

predicted values is called the MAE, or MAE, and it is used to evaluate a regression model's performance (Salah et al., 2022). Conversely, the average squared variation between the forecasted and observed values is measured by the MSE. A reduced MSE indicates an improved model-to-dataset fit. R^2 quantifies how well the prediction model captures their patterns. In a perfect prediction model, R^2 is extremely close to 1. The

Table 2 Statistical performance metrics commonly used in regression analysis

Metric	Full form	Equation
MSE	Mean squared error	$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$
MAE	Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\left\|y_i - \hat{y}_i\right\|$
R^2	Coefficient of determination	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

statistical metrics utilized in this investigation are listed in Table 2. The forecasted value, observed value, and mean value are denoted by $\hat{y}_i$, y_i, and $\bar{y}$, respectively, while n stands for the total number of observations used.

Determining wind energy potential: key factors

The evaluation of performance includes standard analyses of site and turbine selection, offering a comprehensive understanding of the model's efficacy. In assessing the wind energy potential of a given location, several key factors play a crucial role. Understanding and considering these factors are essential for accurately gauging the feasibility and viability of harnessing wind power in a specific area (Baloch et al., 2017; Hulio, 2021; Jiang et al., 2017). Table 3 represents the theoretical terms for determining the wind energy potential of a specific location. Based on the wind power density, a specific location and a specific wind power class are assigned, which leads to the realization of different scales of energy generation, as outlined in Table 4 (Baloch et al., 2017).

Experimental procedure

All simulation-based experiments have been performed with Google Colab. The Python script has been run, employing the following libraries: scikit-learn, keras, seaborn, and matplotlib.

Site selection and data collection

Kutubdia (Upazila) and Cox's Bazar (Sadar Upazila), of the district of Cox's Bazar, Chittagong, Southeast Bangladesh, have been chosen as the study sites. Kutubdia, a coastal island in Bangladesh, poses a unique challenge for wind speed prediction due to its intricate topography and proximity to the Bay of Bengal. Cox's Bazar, known for its extensive beachfront, also demands a specialized approach to wind speed forecasting, considering its distinct geographical features and potential impact on wind project initiatives. The datasets utilized in the experiment were sourced from the BMD weather station and the NASA website. BMD employs Casella cup anemometers for manual wind speed measurements. Though BMD's ground-based measurements are location-specific, the data, recorded in round figures in knot units, may have some reliability limitations (Khadem & Hussain, 2006). Conversely, satellite data cover extensive areas, provide a broader insight into wind patterns, and are consistently standardized. The geometric and other details are provided in Table 5. Figure 4 displays the geographical positions of the stations, where the red circle indicates Kutubdia station and the green one indicates Cox's Bazar station.

Data formatting

The gathered dataset comprises wind speed values recorded at a 3-h interval from January 1, 2001, to June 30, 2022. Each dataset consists of the following variables: i) wind speed, ii) wind direction, ii) temperature; iii) relative humidity; and iv) pressure. Extraneous columns, rows, and elements (e.g., station ID, station name, and details of parameters) have been eliminated from the original datasets. Then the datasets have been prepared in a convenient format for data analysis and preprocessing. The datasets from the BMD stations have been labeled as Dataset 1. Conversely, the datasets sourced from the website of NASA are named Dataset 2.

Exploratory data analysis

Exploratory data analysis (EDA) is essential at the beginning of the data analysis process. To gain a greater knowledge of the dataset's characteristics, make-up, and prospective patterns, it requires closely examining and graphically portraying the dataset. Here, EDA techniques include summarizing key statistics, generating visual plots, and identifying missing values. This process helps in understanding the nature of the data, uncovering the relationships between variables, and guiding subsequent analysis. It plays a crucial role in ensuring the availability of wind speed and generation scale for a specific site.

Tables 6, 7 show the descriptive statistics of both datasets with the count, mean, minimum, maximum, and standard deviation of each input variable. It has been seen that each variable is supposed to contain 62,808 samples. Here are some null values in Dataset 1 for each station. For Kutubdia station, wind speed and wind direction each have 241 null values, whereas humidity has 240 null values. For Cox's Bazar station, 22 null values were observed in each of the two variables—wind speed and wind direction. Other variables contain 62,808 records. There are two missing values for each variable in Dataset 2 for both stations. In

Table 3 Theoretical details of wind resource assessment

Considered factors	Definition	Equation
Wind speed distribution model (MLE-Weibull)	• Weibull probability density function (PDF) is commonly used to model the distribution of wind speeds • The Weibull distribution is flexible and can closely approximate the distribution of wind speeds observed in many locations • It is characterized by shape (k) and scale (c) parameters, offers flexibility in capturing wind speed variability • MLE optimally estimates these parameters, providing efficient, consistent, and statistically sound results (Baloch et al., 2017)	$f(v) = \frac{k}{c}\left(\frac{v}{c}\right)^{k-1}\exp[-\left(\frac{v}{c}\right)^{k}], (k > 0, c > 0)$ where • v is the wind speed (m/s), • c is the scale parameter (m/s), and • k is the shape parameter (dimensionless)
Average wind speed (m/s)	• Average wind speed is a measure of the average speed of the wind over a specified period of time at a particular location	$\overline{v} = c\Gamma\left(\frac{1}{k}+1\right)$, where • $\overline{v}$ is the average wind speed, • c is the scale parameter (m/s), and • k is the shape parameter (dimensionless)
Wind power density (W/m^2)	• Wind power density is a measure of the amount of power available in the wind at a particular location and is a crucial parameter in assessing the potential for harnessing wind energy • It is referred as the power per unit area carried by the wind (Baloch et al., 2017; Hulio, 2021; Jiang et al., 2017)	$\overline{w} = \frac{1}{2}\rho c^3\Gamma\left(\frac{3}{k}+1\right)$, where • ρ is air density (kg/m^3), • c is the scale parameter (m/s), and • k is the shape parameter (dimensionless)
Annual average energy output (kWh)	• The annual average energy output refers to the amount of electrical energy generated by a wind turbine over the course of a year It is a key performance metric that provides an indication of the system's overall efficiency and productivity (Baloch et al., 2017; Hulio, 2021; Jiang et al., 2017)	$E_A = T \times \int_0^{\infty} P_A(v) f(v) dv$ $P_A(v) = \begin{cases} P_r, v_r \le v \le v_{out} \\ P\frac{v - v_{in}}{v_r - v_{in}}, v_{in} \le v \le v_r \\ 0, otherwise, \end{cases}$ where • v_{in} is the cut-in wind speed, • v_r is the rated wind speed, the cut-out wind speed, • T is the time period of the wind turbine operates, • P_r is the rated power, and • f(v) is the optimal Weibull PDF
Capacity factor (%)	• The capacity factor (CF) serves as a crucial metric for assessing the performance of a wind turbine, valuable for both end-users and manufacturers • It represents the ration of the real average power produced during a specific timeframe (assuming continuous turbine operation) and the rated peak power, which is the maximum theoretical power (Baloch et al., 2017; Hulio, 2021; Jiang et al., 2017)	$CF = \frac{E_A}{E_R}$ $E_R = T \times P_R$, where • CF is the capacity factor (Jiang et al., 2017)
Logarithmic wind profile law	• Average wind speed deviation with height is a concept that describes how wind speed changes as you move vertically above the Earth's surface This phenomenon is often explained by wind shear, which is the variation in wind speed and direction with altitude (Hulio, 2021)	$v_2 = v_1 \frac{\ln(\frac{h_2}{z_0})}{\ln(\frac{h_1}{z_0})}$, where • v_1 is the wind speed at height h_1, • v_2 is the wind speed at height h_2, and • z_0 is the roughness length of the terrain

Kutubdia and Cox's Bazar, the BMD station calculates the average wind speed over the past 21.5 years to be 1.1 and 1.04 m/s, respectively, at a height of 10 m. In contrast, NASA records their measurements at 3.42 and 3.89 m/s for the same height. As per the standard deviation, the wind speed distribution in Dataset 1 shows values of 1.06 m/s and 1.50 m/s. Similarly, in Dataset 2, the corresponding values are 1.53 and 1.69. These figures represent the least dispersed values among the input variables. It is crucial to note that

Table 4 International standards of wind power generation classification

Height (m)		10		30		50	
Generation scale	**Wind power class**	**Average wind velocity (m/s)**	**Wind power density (W/m^2)**	**Average wind velocity (m/s)**	**Wind power density (W/m^2)**	**Average wind velocity (m/s)**	**Wind power density (W/m^2)**
Poor	1	0 – 4.4	0 – 100	0 – 5.1	0 – 160	0 -5.6	0 - 200
Marginal	2	4.4 – 5.1	100 – 150	5.1 – 5.9	160 – 240	5.6 – 6.4	200 – 300
Moderate	3	5.1 – 5.6	150 – 200	5.9 – 6.5	240 – 320	6.4 – 7	300 – 400
Good	4	5.6 – 6.4	200 – 250	6.5 – 7	320 – 400	7 – 7.5	400 – 500
Excellent	5	6.4 – 7	250 – 300	7 – 7.4	400 – 480	7.5 – 8	500 – 600
Excellent	6	7 – 7.5	300 – 400	7.4 – 8.2	480 – 640	8 – 8.8	600 – 700
Excellent	7	< 7.5	< 400	8.2 – 11	640 – 1600	< 8.8	< 700

Table 5 Information about the selected sites obtained from two different sources

Station name	Station ID	Longitude (°E)	Latitude (°N)	Altitude (m)	Elevation from sea level (m) for BMD data	Elevation from MERRA-2 (m) for NASA data
Kutubdia	41,989	91.85°	21.8167°	10	2.74	25.33
Cox's Bazar	41,992	91.9667°	21.4333°	10	2.10	25.33

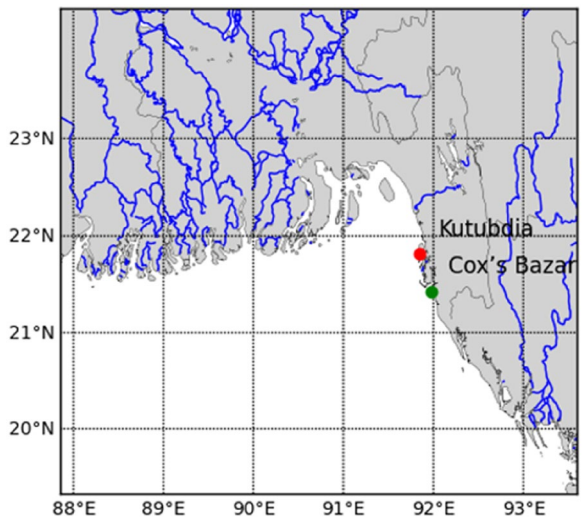


Fig. 4 Geometric location of the studied weather stations

Dataset 1 includes a minimum wind speed value of zero, a characteristic not observed in Dataset 2.

Data preprocessing

Both datasets necessitate preprocessing techniques to meet the requirements of ML algorithms. Addressing null values is a crucial step before initiating the modeling process. Additionally, feature engineering is essential for constructing and training more effective features, ultimately improving the effectiveness of ML models.

Handling missing values: To handle missing values, the following methods are mostly used: forward filling, backward filling, linear interpolation, quadratic interpolation, cubic interpolation, KNN, multiple imputation by chained equations (MICE), and so on (Liu et al., 2021). As the count of missing values is very small for both datasets, the common statistical method of linear interpolation has been adopted to fill up the missing values in the present study.

Zero refining: When employing ML models, certain algorithms may exhibit sensitivity to the existence of zero values. In such instances, it can be advantageous to refine or transform these zero values. To address the excessive presence of zero values in the wind speed data of Dataset 1, zeros have been adjusted to the smallest valid value as verified by BMD (Nurunnahar et al., 2017).

Feature extraction: Date-related features were generated to extract valuable information from the date column of Dataset 1. This led to the augmentation of the dataset with additional columns, including year, month, day, and hour. Dataset 2 already contains these columns relevant to the date. By incorporating time-delayed data from the wind speed time series, we aimed to evaluate the influence of previous values on present observations for both datasets. This approach proves beneficial in accounting for important correlated time lags. In Fig. 5, autocorrelation function (ACF) curves were plotted with 60 lags to determine suitable input lags. The choice

Table 6 Statistic features of Dataset 1

Weather station	Feature	Count	Missing	Minimum	Maximum	Mean	Std. deviation
Kutubdia	Wind speed (m/s)	62,567	241	0	20.58	1.1	1.06
	Wind direction (Degree)	62,567	241	0	990	148.86	113.37
	Temperature (°C)	62,808	240	2.9	38	26.01	4.41
	Humidity (%)	62,568	240	8	100	80.73	12.96
	Pressure (millibar)	62,808	240	9.5	2003.8	1004.58	62.79
Cox's Bazar	Wind speed (m/s)	62,786	22	0	18.52	1.04	1.5
	Wind direction (Degree)	62,786	22	0	900	80.68	117.54
	Temperature (°C)	62,808	0	10.5	39	26.15	4.14
	Humidity (%)	62,808	0	13	100	79.93	15.46
	Pressure (millibar)	62,808	0	100.7	1023.4	1008.02	6.03

Table 7 Statistic features of Dataset 2

Weather station	Feature	Count	Missing	Minimum	Maximum	Mean	Std. deviation
Kutubdia	Wind speed (m/s)	62,806	2	0.03	17.17	3.42	1.53
	Wind direction (Degree)	62,806	2	0	359.91	190.72	99.95
	Temperature (°C)	62,806	2	11.8	35.91	25.65	3.84
	Humidity (%)	62,806	2	21.75	100	80.22	12.92
	Pressure (millibar)	62,806	2	980.5	1019	1005.30	5.03
Cox's Bazar	Wind speed (m/s)	62,806	2	0.02	19.16	3.89	1.69
	Wind direction (Degree)	62,806	2	0	359.93	201.02	103.75
	temperature (°C)	62,806	2	15.47	33.65	26.17	3.04
	Humidity (%)	62,806	2	28.56	100	79.20	10.65
	Pressure (millibar)	62,806	2	985.20	1019.9	1006.54	4.91

of input lag features was determined by a correlation threshold of 0.4 or higher. In this scenario, for Dataset 1, the initial 12 and 8 lags were selected for Kutubdia and Cox's Bazar stations, respectively. In Dataset 2, the first 18 lags were chosen for both stations. A rolling window scheme was employed while taking into account the input lags (Mollick et al., 2023).

Feature selection: Pearson's correlation coefficient (PCC) is employed to quantify the extent of the association between two variables. The correlation may have a positive (+) or negative (-) value for the relationship (Salah et al., 2022). Pearson's correlation matrix, showing the correlation between the features within each dataset, is presented in Fig. 6. It is observed that Dataset 1 shows the lowest connection (r = -0.016) between wind speed and hour for Kutubdia station, while Cox's Bazar station shows the lowest correlation (r = -0.0021) between wind speed and day. In Dataset 2, both stations have the lowest correlation with day. For both datasets, the wind speed is positively correlated with the rolling mean, with the highest value of r. The SelectKBest feature selection method from the sci-kit-learn library, a filter-based approach, has been utilized. This method operates on the PCC between pairs of input variables, aiding in filtering out the most pertinent features. A subset comprising the eight most correlated features was chosen for both datasets (Mollick et al., 2023).

Data normalization: In ML practices, data normalization is a common technique utilized to mitigate the influence of data range variations (Waqas Khan et al., 2020). In this study, robust scaling is adopted for data normalization. This technique employs the median and interquartile range (IQR) to adjust input values. This characteristic of robust scaling ensures its resilience against the detrimental effects of outliers (Zhang et al., 2022).

Data splitting

Each of the considered datasets is divided into three parts: 70% as a training set, 15% as a validation set, and 15% as a test set. A training set is utilized to train the ML model. From this data, the model learns trends, connections, and features. The model has to have its performance assessed after training. This is accomplished by using the validation set. The model is tested on the test

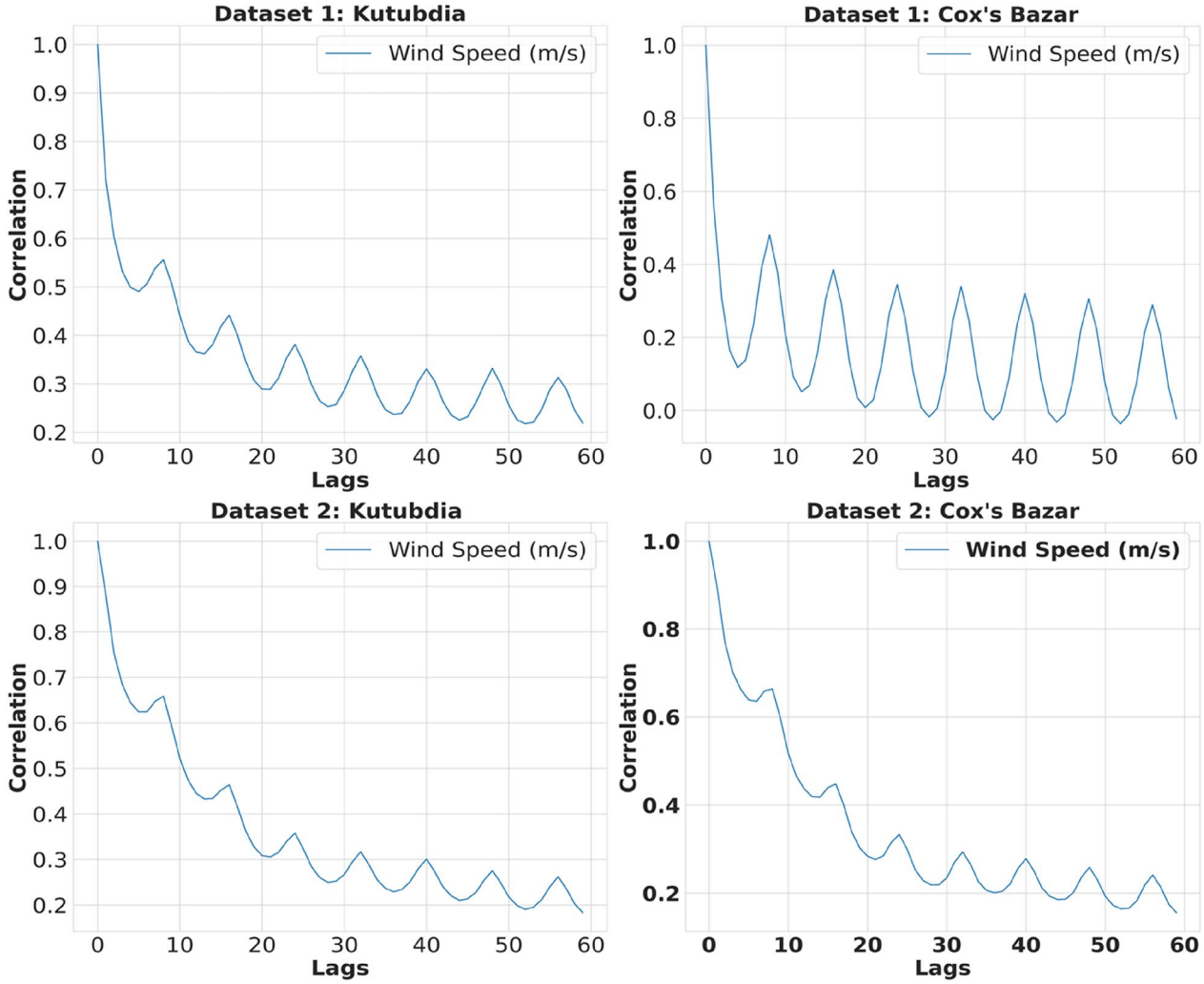


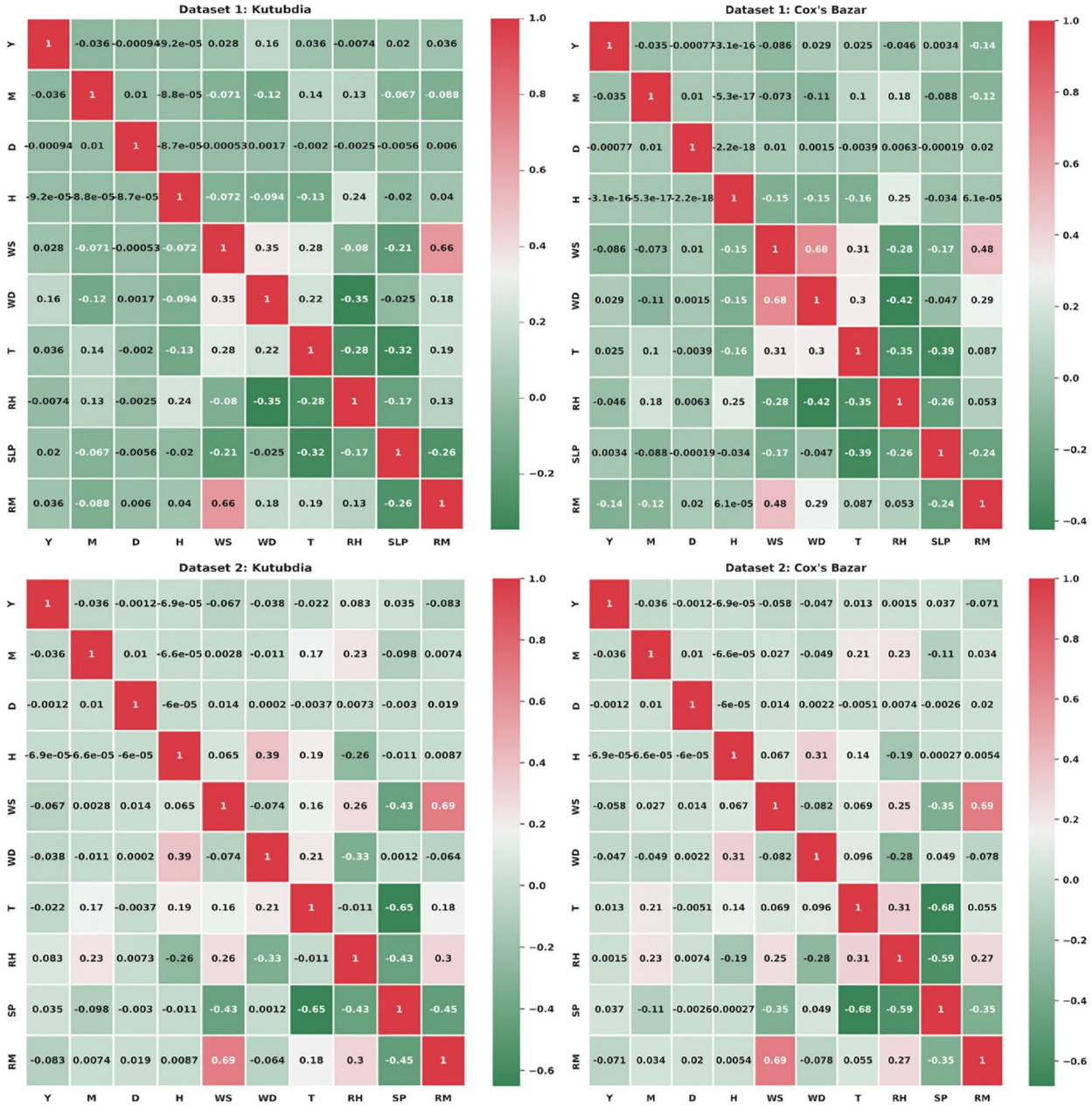
Fig. 5 ACF curves for Dataset 1 and Dataset 2

set after having been trained and validated using the training and validation sets. This set provides an unbiased evaluation of the model's performance since it is not visible to the model during training or tuning.

Model optimization and training

For model training, the study used two independent methods: first, a tenfold cross-validation (tenfold CV), and second, hyperparameter tuning with Hyperopt. Both methods were used to assess the model's performance, and the best model was ultimately picked as a consequence. The validation set was then used to test this improved model. An essential assessment for model generalization is the comparison of CV findings and validation set performance. The model can be used to test the data with confidence if the differences are small, enhancing the final model's robustness.

The efficiency of a ML model was evaluated using a tenfold cross-validation technique. The dataset needs to be split into ten identical folds to perform this. The model is trained on nine of these folds and evaluated on the final or tenth fold. The technique is repeated a total of ten times, with each fold serving as the test set. The final performance metric is generated by averaging the ten individual test scores. Hyperopt has also been conducted to determine the optimal value of the different parameters. We leveraged Hyperopt's fmin function to meticulously search for the optimal hyperparameters, aiming to minimize the negative MSE. The hyperparameter space was meticulously defined, drawing insights from prior research endeavors. With a defined limit of 80 evaluations, the hyperparameter optimization process was meticulously logged and tracked in detailed trials. Subsequently, each model was meticulously



Note: Y: Year; **M:** Month; **D:** Day; **H:** Hour; **WS:** Wind speed; **WD:** Wind direction; **T:** Temperature; **RH:** Relative humidity; **SLP:** Sealevel pressure; **RM:** Rolling Mean.

Fig. 6 Heatmap of the correlation matrix for Dataset 1 and Dataset 2

trained using the best hyperparameters on the designated training dataset. Upon obtaining the optimal hyperparameters through Hyperopt, the models were further fine-tuned on the training data and assessed for performance. Finally, the refined and optimized models were deployed to provide accurate predictions on the test dataset following a validation assessment on the validation set. Table 8 displays details regarding the regression technique, the hyperparameters slated for optimization, and their respective search spaces. Apart from hyperparameter adjustments, all other parameters for each model were maintained at their default settings. The deep learning model, LSTM, is structured as a sequential model, incorporating only one layer

Table 8 Hyperparameter tuning using Hyperopt

Regression method	Hyperparameter	Search space
MLR	-	-
Lasso	alpha	10^{-3}, 10^{-2}, 10^{-1},…..,10^{3}
Ridge	alpha	10^{-3}, 10^{-2}, 10^{-1},…..,10^{3}
Elastic Net	alpha	10^{-3}, 10^{-2}, 10^{-1},…..,10^{3}
KNN	n_neighbors	1, 2, 3,….., 11
DT	max_depth	1, 2, 3, ….., 9
RF	max_depth n_estimators	1, 2, 3, ….., 9 50, 55, 60, ….., 195
GBR	learning_rate max_depth	10^{-5}, 10^{-4}, 10^{-3}, …..,1 1, 2, 3, ….., 9
ADABoost	learning_rate	10^{-5}, 10^{-4}, 10^{-3}, …..,1
XGBoost	learning_rate max_depth	10^{-5}, 10^{-4}, 10^{-3}, …..,1 1, 2, 3, ….., 9
LightGBM	learning_rate max_depth	10^{-5}, 10^{-4}, 10^{-3}, …..,1 1, 2, 3, ….., 9
CatBoost	learning_rate	10^{-5}, 10^{-4}, 10^{-3}, …..,1
LSTM	no. of units learning_rate	10, 20, 30, ….., 100 10^{-5}, 10^{-4}, 10^{-3}, …..,1
GRU	no. of units learning_rate	10, 20, 30, ….., 100 10^{-5}, 10^{-4}, 10^{-3}, …..,1

followed by a dense output layer. This represents a simplified LSTM architecture typically employed for fundamental sequential prediction assignments. The Adam optimizer with a dropout of 0.1 is used, and the MSE is chosen as the loss function. The model is trained for 50 epochs, with training progress printed at each epoch. Similarly, the GRU adheres to the same structure and methodology as LSTM.

Results and discussion

Comparing predictive models using evaluation metrics

After applying the relevant equations of the eight ML models in a Python environment, the prediction results, including RMSE, MAE, MSE, and R^2, are obtained, which are displayed in Tables 9, 10, 11, 12. Overall, all models exhibit similar performance, providing moderate predictions. Notably, CatBoost model outperforms other machine-learning models across various performance metrics for both weather stations.

Utilizing tenfold cross-validation, CatBoost (CATB) demonstrated the best performance across all datasets.

In Dataset 1, the model attained an MSE of 0.3745, MAE of 0.3984, and R^2 of 0.6218 for Kutubdia station. For Cox's Bazar station, the model yielded an MSE of 0.9462, MAE of 0.6164, and R^2 of 0.514. In Dataset 2, the model demonstrated optimal performance with MSE values of 0.3224 and 0.3541, MAE values of 0.4117 and 0.4347, and R^2 values of 0.8618 and 0.8755 for Kutubida and Cox's Bazar, respectively. Post-hyperparameter optimization, the model's performance saw notable improvement on both datasets. However, it is worth noting that, in some scenarios, without any parameter tuning, the LSTM and GRU models exhibit superior performance in the context of tenfold cross-validation. Following hyperparameter tuning, CatBoost emerged as the top-performing model, demonstrating impressive outcomes in Dataset 1. For Kutubdia, it achieved an MSE of 0.3744, MAE of 0.399, and R^2 of 0.6218. Similarly, for Cox's Bazar, it delivered an MSE of 0.9382, MAE of 0.6162, and R^2 of 0.518. Shifting focus to Dataset 2, CatBoost emerged as the top-performing model, with an MSE of 0.3218 and 0.3533, MAE of 0.4117 and 0.4342, and R^2 of 0.8621 and 0.8758 for Kutubdia and Cox's Bazar, respectively.

In the validation phase, CatBoost performed exceptionally, showcasing distinguished results with an MSE of 0.3388, MAE of 0.3912, and R^2 of 0.6409 for Kutubdia station in Dataset 1. Similarly, it attained the best results with an MSE of 0.9328, MAE of 0.6157, and R^2 of 0.5192 for Cox's Bazar station. Turning attention to Dataset 2, the CatBoost again outperformed its counter models with an MSE of 0.3309 and 0.3713, MAE of 0.415, and 0.4398, and R^2 of 0.858 and 0.8714 for Kutubdia and Cox's Bazar, respectively. Following closely, the LGBM model illustrated the second-best performance for all datasets. Moving to the testing phase, in Dataset 1, CatBoost achieved an MSE of 0.3942, a MAE of 0.4042, and an R^2 of 0.6242 for Kutubdia station. Again, CatBoost showcased notable performance with an MSE of 0.9906, MAE of 0.6363, and R^2 of 0.4994 for Cox's Bazar in Dataset 1. In Dataset 2, the dominating performance was achieved by CatBoost, with an MSE of 0.3305, MAE of 0.4164, and R^2 of 0.8552 for Kutubdia. For Cox's Bazar, the model's performance is nearly identical to Kutubdia, with an MSE of 0.3744, MAE of 0.4415, and R^2 of 0.867. For each scenario (validation and testing phase), the LGBM model demonstrated performance closely trailing behind the leading model in all datasets. Conversely, the AdaBoost demonstrated relatively lower performance compared to the other models with the exception of the Cox's Bazar station in Dataset 1. In this case, the Lasso model attained the lowest evaluation metrics.

Apart from the results shown in Tables 9, 10, 11, 12, the difference between the observed wind speed observations and the predicted wind speed based on the best-performing prediction model during the testing phase is also depicted in scatter plots, histograms, and box plots (Figs. 7, 8, and 9). Figures 7 and 8 show the scatter plot

Table 9 Creating and comparing 14 models using tenfold cross-validation and hyperparameter tuning with Hyperopt optimization for Dataset 1 (best results are bolded)

Weather station	Model	10-fold cross-validation			Hyperparameter tuning with Hyperopt		
		MSE	MAE	R^2	MSE	MAE	R^2
Kutubdia	MLR	0.4174	0.4325	0.5782	0.4174	0.4325	0.5782
	Lasso	0.9899	0.7127	-0.0003	0.4310	0.4477	0.5645
	Ridge	0.4174	0.4325	0.5782	0.4174	0.4325	0.5782
	Elastic Net	0.9256	0.6878	0.0648	0.4240	0.4397	0.5716
	KNN	0.4350	0.4291	0.5604	0.4096	0.4172	0.5863
	DT	0.7904	0.5412	0.1976	0.4229	0.4250	0.5727
	RF	0.4021	0.4147	0.5937	0.3919	0.4086	0.6039
	GBR	0.3855	0.4089	0.6105	0.3789	0.4030	0.6174
	AdaBoost	0.6720	0.5978	0.3225	0.4626	0.4598	0.5320
	XGBoost	0.3980	0.4073	0.5976	0.3809	0.4041	0.6152
	LightGBM	0.3798	0.4018	0.6163	0.3789	0.4020	0.6173
	CatBoost	**0.3745**	**0.3984**	**0.6218**	**0.3744**	**0.3990**	**0.6218**
	LSTM	0.3964	0.4173	0.5995	0.4350	0.4501	0.5604
	GRU	0.3984	0.4194	0.5973	0.4050	0.4229	0.5908
Cox's Bazar	MLR	1.1323	0.7412	0.4182	1.1323	0.7412	0.4182
	Lasso	1.9466	1.1068	-0.0002	1.1479	0.7460	0.4102
	Ridge	1.1323	0.7412	0.4182	1.1323	0.7413	0.4182
	Elastic Net	1.8116	1.0660	0.0693	1.1418	0.7431	0.4134
	KNN	1.1381	0.6638	0.4152	1.0675	0.6511	0.4516
	DT	1.9969	0.8122	-0.0265	1.0338	0.6485	0.4691
	RF	0.9779	0.6291	0.4976	1.0116	0.6406	0.4802
	GBR	0.9615	0.6286	0.5061	0.9546	0.6251	0.5095
	AdaBoost	1.0962	0.8532	0.3716	0.98144	0.7329	0.4375
	XGBoost	0.9982	0.6294	0.4872	0.9524	0.6209	0.5107
	LightGBM	0.9472	0.6192	0.5135	0.9468	0.6184	0.5137
	CatBoost	**0.9462**	**0.6164**	**0.5140**	**0.9382**	**0.6162**	**0.5180**
	LSTM	1.0051	0.6588	0.4835	0.9943	0.6464	0.4892
	GRU	1.0067	0.6569	0.4827	1.0042	0.6552	0.4839

and forecasting error histogram plot, respectively, for both datasets during testing phase.

The scatter plot presents the predicted versus the observed wind speed values. Plots evaluate the cause-and-effect relationship between projected and observed wind speed and measure the robustness of the association between these two variables using the coefficient of determination R^2. In terms of R^2 for Kutubdia and Cox's Bazar in Dataset 1, the Catboost model produced the best prediction performance ($R^2 = 0.642$ and 5342, respectively). Similarly, in Dataset 2 the model produced the best results ($R^2 = 0.8552$ and 0.867, respectively) for both stations. Additionally, there is considerably less deviation from the regression line in Dataset 1 for all cluster points compared to Dataset 2. In contrast to Dataset 2, the CatBoost model exhibited robust prediction performance for Dataset 1. In summary, when compared to BMD data, the CatBoost model exhibited the least deviation from the line for all data samples, marking a significant shift in NASA data. This aligns with the accuracy metrics, particularly the R^2 values presented in Tables 9, 10, 11, 12.

The histogram plot graphically interprets the error distribution by displaying the number of error values within a certain range, and it includes the Gaussian kernel density function to guarantee that the error follows a normal distribution. The plots indicate that in Dataset 1, the CatBoost model exhibits the standard deviation (0.6278 and 0.9952 for Kutubdia and Cox's Bazar, respectively), suggesting that the data points cluster closely around the mean. Meanwhile, in Dataset 2, the CatBoost model demonstrates a standard deviation of 0.5749 and 0.6119 for Kutubdia and Cox's Bazar, respectively. The smaller

Table 10 The evaluation metrics for 14 models on both validation and test segment for Dataset 1 (best results are bolded)

Weather station	Model	Validation dataset			Test dataset		
		MSE	MAE	R^2	MSE	MAE	R^2
Kutubdia	MLR	0.3955	0.4272	0.5808	0.4467	0.4371	0.5741
	Lasso	0.3955	0.4272	0.5808	0.4467	0.4371	0.5741
	Ridge	0.4052	0.4424	0.5705	0.4618	0.4526	0.5597
	Elastic Net	0.3989	0.4342	0.5771	0.4551	0.4445	0.5661
	KNN	0.3862	0.4122	0.5906	0.4472	0.4265	0.5737
	DT	0.3892	0.4171	0.5874	0.4448	0.4290	0.5760
	RF	0.3638	0.4022	0.6143	0.4155	0.4127	0.6038
	GBR	0.3477	0.3960	0.6314	0.4073	0.4085	0.6117
	AdaBoost	0.4331	0.4523	0.5409	0.4871	0.4646	0.5357
	XGBoost	0.3485	0.3969	0.6306	0.4031	0.4085	0.6157
	LightGBM	0.3437	0.3953	0.6357	0.4072	0.4079	0.6118
	CatBoost	**0.3388**	**0.3912**	**0.6409**	**0.3942**	**0.4042**	**0.6242**
	LSTM	0.3642	0.4143	0.6139	0.4206	0.4254	0.5990
	GRU	0.3685	0.4136	0.6094	0.4257	0.4242	0.5941
Cox's Bazar	MLR	1.1406	0.7451	0.4121	1.1681	0.7559	0.4097
	Lasso	1.1642	0.7510	0.3999	1.1843	0.7599	0.4015
	Ridge	1.1406	0.7451	0.4121	1.1681	0.7559	0.4097
	Elastic Net	1.1553	0.7476	0.4045	1.1788	0.7571	0.4042
	KNN	1.0651	0.6512	0.4510	1.1283	0.6711	0.4297
	DT	1.0297	0.6496	0.4693	1.1088	0.6716	0.4396
	RF	0.9655	0.6268	0.5023	1.0180	0.6464	0.4854
	GBR	0.9496	0.6249	0.5105	1.0024	0.6444	0.4933
	AdaBoost	0.9684	0.7280	0.4421	0.9536	0.7272	0.4501
	XGBoost	0.9416	0.6195	0.5147	1.0003	0.6416	0.4945
	LightGBM	0.9395	0.6183	0.5158	0.9944	0.6380	0.4974
	CatBoost	**0.9328**	**0.6157**	**0.5192**	**0.9906**	**0.6363**	**0.4994**
	LSTM	0.9895	0.6514	0.4900	1.0166	0.6645	0.4862
	GRU	1.0065	0.6482	0.4812	1.0431	0.6626	0.4728

standard deviation was achieved by the model in case of Kutubdia station for both datasets. This implies that the data points are more tightly grouped around the mean when predicted by this model.

Figure 9 displays boxplots illustrating prediction errors for various models using test datasets. Each graph represents the distribution of residual errors, indicating key statistics like minimum, first quartile, median, third quartile, and maximum values. The bagging and boosting ensemble models, particularly RF, GBR, XGBoost, LightGBM, and CatBoost, showcase similar performance, with noticeable differences in the width of the box across all datasets. Regarding outliers, all models perform in a similar manner.

Quartile percent values, which may indicate additional information about the efficacy of each model individually, are shown in Tables 13 and 14. It is seen that the CatBoost produces a smaller IQR of 0.5408 for Kutubdia station in Dataset 1 than the other models do. For Cox's Bazar station the decision tree (DT) model has the smallest IQR of 0.6845. In Dataset 2, the CatBoost model generates the smallest IQR, measuring 0.6369 for Kutubdia and 0.6730 for Cox's Bazar station. The bagging and boosting ensemble models exhibit lower standard deviations in prediction values for, primarily due to their ensemble learning nature and effective handling of outliers. These models combine multiple weak learners and apply regularization techniques to prevent overfitting, resulting in more stable and consistent predictions. Additionally, their focus on important features contributes to the reduced variability in predictions across different data points.

As stated earlier, in this study, 14 ML techniques, including MLR, Lasso, Ridge, Elastic Net, KNN, DT, RF, GBR, AdaBoost, XGBoost, LightGBM, CatBoost, LSTM,

Table 11 Creating and comparing 14 models using tenfold cross-validation and hyperparameter tuning with Hyperopt optimization for Dataset 2 (best results are bolded)

Weather station	Model	10-fold cross-validation			Hyperparameter tuning with Hyperopt		
		MSE	MAE	R^2	MSE	MAE	R^2
Kutubdia	MLR	0.4976	0.5310	0.7868	0.4976	0.5310	0.7868
	Lasso	1.9781	1.0706	0.1533	0.5115	0.5361	0.7809
	Ridge	0.4976	0.5310	0.7868	0.4976	0.5310	0.7868
	Elastic Net	1.3426	0.8776	0.4252	0.5103	0.5373	0.7814
	KNN	0.4106	0.4725	0.8240	0.3913	0.4607	0.8324
	DT	0.6882	0.6104	0.7049	0.4571	0.4973	0.8041
	RF	0.3473	0.4286	0.8512	0.4057	0.4705	0.8261
	GBR	0.3861	0.4607	0.8345	0.3351	0.4200	0.8564
	AdaBoost	0.6405	0.6127	0.7255	0.5284	0.5457	0.7736
	XGBoost	0.3413	0.4234	0.8538	0.3339	0.4192	0.8569
	LightGBM	0.3348	0.4215	0.8565	0.3332	0.4200	0.8572
	CatBoost	**0.3224**	**0.4117**	**0.8618**	**0.3218**	**0.4117**	**0.8621**
	LSTM	0.4171	0.4825	0.8215	0.4107	0.4793	0.8241
	GRU	0.4237	0.4874	0.8187	0.4223	0.4869	0.8191
Cox's Bazar	MLR	0.5343	0.5504	0.8121	0.5343	0.5504	0.8121
	Lasso	2.1950	1.1540	0.2288	0.5524	0.5602	0.8057
	Ridge	0.5343	0.5504	0.8121	0.5343	0.5505	0.8121
	Elastic Net	1.5971	0.9822	0.4388	0.5513	0.5618	0.8062
	KNN	0.4495	0.4964	0.8419	0.4275	0.4844	0.8497
	DT	0.7535	0.6428	0.7351	0.4889	0.5205	0.8280
	RF	0.3810	0.4521	0.8660	0.4311	0.4887	0.8484
	GBR	0.4111	0.4763	0.8554	0.3692	0.4455	0.8702
	AdaBoost	0.7095	0.6509	0.7505	0.5635	0.5664	0.8019
	XGBoost	0.3801	0.4484	0.8663	0.3688	0.4445	0.8703
	LightGBM	0.3637	0.4420	0.8721	0.3642	0.4419	0.8720
	CatBoost	**0.3541**	**0.4347**	**0.8755**	**0.3533**	**0.4342**	**0.8758**
	LSTM	0.4498	0.5031	0.8420	0.4543	0.5061	0.8403
	GRU	0.4648	0.5124	0.8367	0.4727	0.5183	0.8340

and GRU, have been used to estimate the short– time wind speed forecast. Result shows, the CatBoost model is identified as the most proficient predictor of short-term wind speed forecast based on the conducted estimation procedures, exhibiting the smallest error metric scores and the highest level of accuracy compared to alternative methods. However, the forecasting accuracy for Dataset 2 surpasses that of Dataset 1. Table 15 displays the performance assessment, showcasing the most successful outcome achieved, in contrast to models examined in previous studies.

Generation scale and turbine compatibility

Wind resource assessment is a critical step in evaluating the viability of a location for harnessing wind energy. It involves understanding the wind characteristics unique to a specific site, essential for optimizing the design and performance of wind energy projects. In this study, maximum likelihood estimation (MLE) of the Weibull distribution is used which to aid in modeling the probability distribution of the observed and predicted wind speeds of both stations, providing valuable insights into the expected wind energy potential.

Based on the superior prediction accuracy demonstrated, we have opted to proceed with the satellite data for further investigation, favoring it over BMD data.

In order to correspond with the wind speed measurements commonly recorded by commercial turbines at hub heights of 50 m and 120 m, the wind speed data were transformed from 10 m to those specific heights using the logarithmic law wind formula. The weather station is located in a built-up area, and the roughness value (z_0) in this context falls within the range of 0.1 to 0.4 m. For our analysis, we have adopted the value of 0.3 (Islam et al., 2013). Figure 10 illustrates the probability density

Table 12 The evaluation metrics for 14 models on both validation and test segment for Dataset 2 (best results are bolded)

Weather station	Model	Validation dataset			Test dataset		
		MSE	MAE	R^2	MSE	MAE	R^2
Kutubdia	MLR	0.5070	0.5316	0.7825	0.4912	0.5293	0.7849
	Lasso	0.5213	0.5372	0.7764	0.5007	0.5344	0.7807
	Ridge	0.5070	0.5316	0.7825	0.4912	0.5293	0.7849
	Elastic Net	0.5182	0.5379	0.7777	0.4988	0.5349	0.7815
	KNN	0.3972	0.4631	0.8296	0.3875	0.4604	0.8303
	DT	0.4665	0.5002	0.7999	0.4543	0.4980	0.8010
	RF	0.4208	0.4737	0.8194	0.4138	0.4756	0.8187
	GBR	0.3509	0.4242	0.8495	0.3447	0.4256	0.8490
	AdaBoost	0.5370	0.5475	0.7697	0.5159	0.5464	0.7741
	XGBoost	0.3427	0.4222	0.8530	0.3429	0.4249	0.8498
	LightGBM	0.3411	0.4221	0.8537	0.3447	0.4250	0.8490
	CatBoost	**0.3309**	**0.4150**	**0.8580**	**0.3305**	**0.4164**	**0.8552**
	LSTM	0.3832	0.4554	0.8356	0.3739	0.4549	0.8362
	GRU	0.3860	0.4589	0.8345	0.3709	0.4558	0.8375
Cox's Bazar	MLR	0.5679	0.5610	0.8032	0.5434	0.5538	0.8070
	Lasso	0.5833	0.5707	0.7979	0.5587	0.5623	0.8015
	Ridge	0.5679	0.5610	0.8032	0.5434	0.5538	0.8070
	Elastic Net	0.5803	0.5712	0.7989	0.5582	0.5633	0.80171
	KNN	0.4500	0.4891	0.8441	0.4401	0.4891	0.8436
	DT	0.5235	0.5337	0.8186	0.4973	0.5254	0.8233
	RF	0.4548	0.4988	0.8424	0.4451	0.4943	0.8419
	GBR	0.3913	0.4528	0.8644	0.3833	0.4504	0.8638
	AdaBoost	0.5908	0.5761	0.7953	0.5696	0.5702	0.7977
	XGBoost	0.3949	0.4526	0.8632	0.3843	0.4502	0.8634
	LightGBM	0.3904	0.4497	0.8647	0.3834	0.4470	0.8638
	CatBoost	**0.3713**	**0.4398**	**0.8714**	**0.3744**	**0.4415**	**0.8670**
	LSTM	0.4411	0.4901	0.8472	0.4342	0.4903	0.8457
	GRU	0.4676	0.5087	0.8380	0.4575	0.5089	0.8375

function (PDF) plot of both observed and predicted wind speed data for both stations. The average wind velocity and wind power density have been computed using the Weibull distribution parameters (k and c) detailed in Tables 16 and 17, corresponding to heights of 50 m and 120 m. Wind power class and generation scale have been assigned based on the calculated wind power density. While the parameter values exhibit slight variations from those of the observed data, consistent matching of wind power class and generation scale is observed across all cases, except for Kutubdia station at 120 m height.

When confronted with a location characterized by small and marginal generation-scale wind speeds (e.g., Kutubdia and Cox's Bazar), there are particular factors to take into account when selecting and optimizing turbines. It becomes imperative to opt for turbines specifically engineered to function effectively in such circumstances. For instance, specialized turbines with efficient blades are crucial for capturing energy from slower winds in low-wind conditions. A larger rotor diameter allows for more effective energy extraction at lower wind speeds. Additionally, selecting turbines with lower cut-in speeds ensures power generation starts at lower wind speeds, maximizing overall energy yield. Optimizing pitch control is crucial for maximizing energy extraction from low wind speeds. Fine-tuning the turbine's speed regulation system, including adjusting the generator's speed curve, enhances efficiency in these conditions. Additionally, careful consideration of wake effects and proper spacing between turbines, coupled with advanced wake modeling techniques, plays a pivotal role in optimizing energy production within the

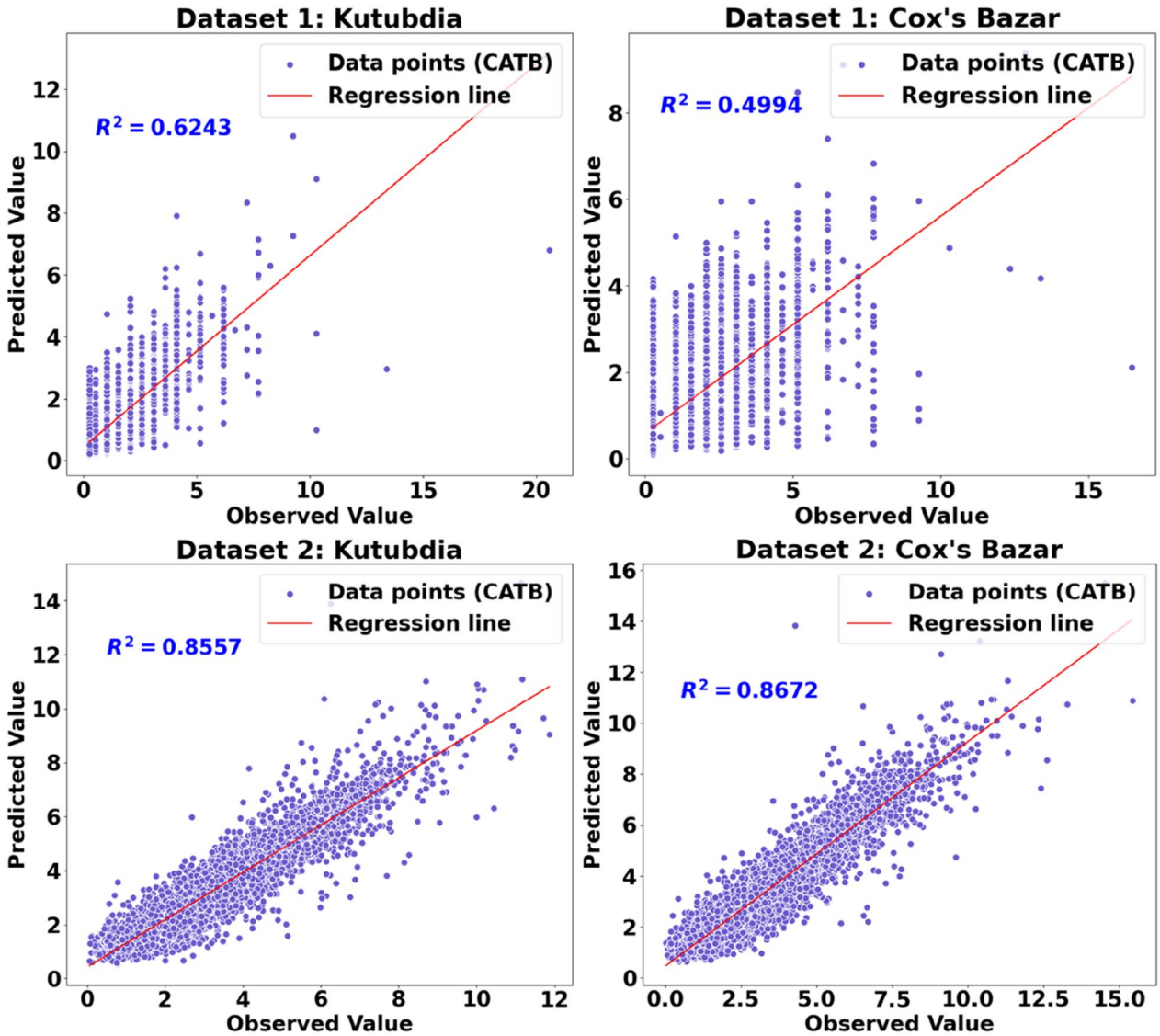


Fig. 7 Scatter plots of wind speed prediction for Dataset 1 and Dataset 2

wind farm. It is noteworthy to mention that the actual turbine specifications may vary based on manufacturers and specific models. It is important to consult the manufacturer's specifications for precise details. Recent and popular models such as Vestas, Siemens Gamesa, General Electric (GE) Renewables, Nordex, Enercon, Senvion, Suzlon, Goldwind, Ming Yang, and Envision Energy are commonly employed for turbines in sites with lower wind speeds. Table 18 displays the attributes of some low-speed wind turbines of different models as observed in recent years (Bauer, 2023). The decision options for turbine selection involve evaluating two key criteria: capacity factor (CF), which is widely utilized as a primary decision factor, and annual average energy output (Darwish et al., 2019).

In this investigation, the capacity factor is considered an evaluation metric for choosing the suitable turbine based on the observed satellite data. Table 19 displays the annual average energy output and capacity factor associated with each turbine type listed in Table 18, based on the observed satellite data for both locations. The findings indicate that among the various turbine models, the Goldwind model exhibits the most favorable performance. Specifically, the turbine GW 171/3850

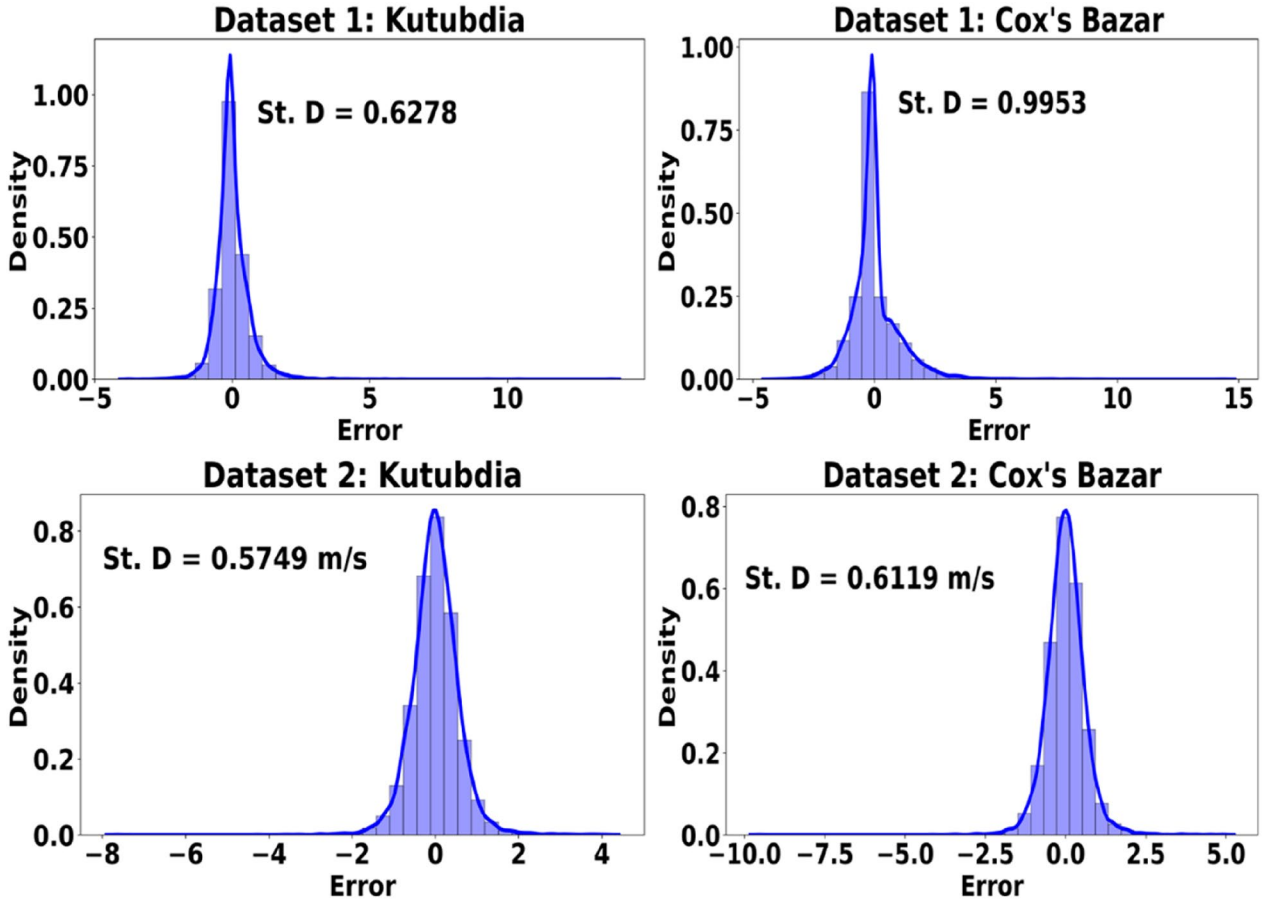


Fig. 8 Histograms and Gaussian kernel density functions of wind speed prediction for Dataset 1 and Dataset 2

distinguishes itself as the most fitting choice, demonstrating the highest capacity factor for both locations (37.17% and 46.99% for Kutubdia and Cox's Bazar, respectively). It is important to highlight that the turbines with a capacity factor equal to or exceeding 20% are considered viable for the respective sites (Islam et al., 2013).

Figure 11 shows the wind power curve or wind turbine power performance curve of the highest CF turbine (Goldwind GW 171/3850), which illustrates the relationship between observed wind speed and the electrical power output of a wind turbine for 120 m hub height. The curve shows how the turbine's power output increases with higher wind speeds until reaching the rated power (Assareh et al., 2016). The power curves exhibit identical characteristics for both stations. The wind turbine begins to generate power at the cut-in wind speed, the minimum speed required for power generation. At the rated wind speed, the turbine achieves its maximum designed power output. Beyond the cut-out wind speed, the turbine shuts down to prevent damage. This is the maximum wind speed the turbine can withstand.

In low-wind sites, ensuring a continuous power supply requires the integration of a hybrid system. This system combines a wind turbine with an alternative power source, such as solar panels or a small-scale generator, to supplement energy production during periods of low or no wind. If the wind speeds are inadequate, the hybrid system consistently shifts to an alternative power source so that it can allow the turbine to uphold operating. A reliable and uninterrupted power supply can be secured by this approach, which is particularly effective for low- and unstable wind sites. A hybrid system upgrades the overall performance and sustainability of the energy generation system in such conditions by tactically adjusting the wind and secondary energy sources.

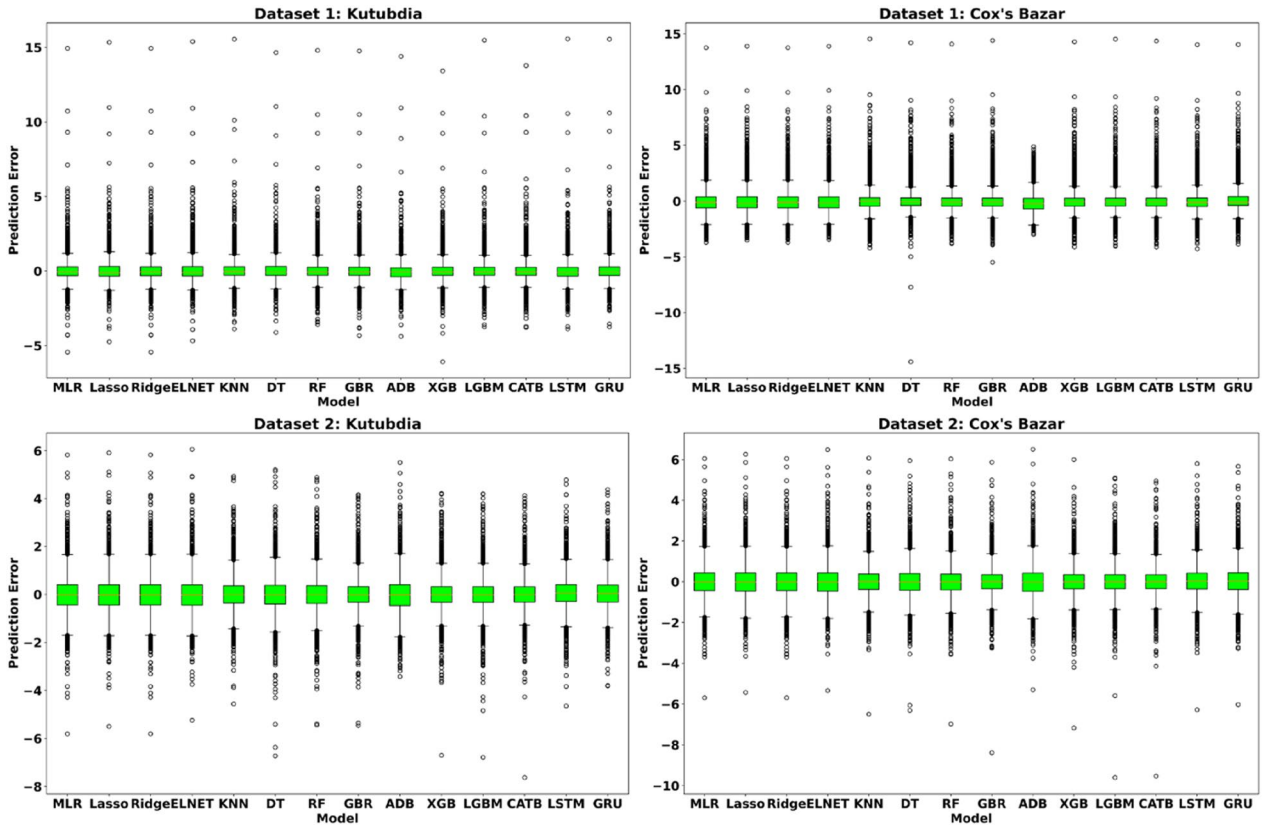


Fig. 9 Boxplots of the prediction error for Dataset 1 and Dataset 2

Table 13 Quartile percent of the prediction error for Dataset 1 (minimum Std. deviation and IQR are bolded)

Model	Quartile percentile for Dataset 1: Kutubdia					Quartile percentile for Dataset 1: Cox's Bazar				
	Std	25%	50%	IQR	75%	Std	25%	50%	IQR	75%
MLR	0.6682	− 0.3211	− 0.0657	0.6059	0.2848	1.0808	− 0.6213	− 0.1224	0.9999	0.3786
Lasso	0.6794	− 0.3396	− 0.1186	0.6483	0.3088	1.0883	− 0.6093	− 0.1852	0.9948	0.3855
Ridge	0.6682	− 0.3211	− 0.0657	0.6059	0.2848	1.0808	− 0.6219	− 0.1228	1.0009	0.3790
Elastic Net	0.6745	− 0.3303	− 0.0960	0.6256	0.2953	1.0858	− 0.6117	− 0.1659	0.9865	0.3748
KNN	0.6683	− 0.2858	− 0.0286	0.5716	0.2858	1.0619	− 0.4573	0.0000	0.7718	0.3145
DT	0.6668	− 0.2919	− 0.0797	0.6025	0.3107	1.0531	− 0.3992	− 0.0798	**0.6845**	0.2853
RF	0.6445	− 0.2855	− 0.0602	0.5458	0.2603	1.0090	− 0.4448	− 0.0886	0.7188	0.2740
GBR	0.6381	− 0.2884	− 0.0584	0.5488	0.2605	1.0012	− 0.4446	− 0.0872	0.7179	0.2732
AdaBoost	0.6975	− 0.3864	− 0.0675	0.6025	0.2161	**0.9716**	− 0.7077	− 0.2351	0.9650	0.2573
XGBoost	0.6348	− 0.2915	− 0.0573	0.5569	0.2654	1.0001	− 0.4415	− 0.0765	0.7054	0.2640
LightGBM	0.6380	− 0.2859	− 0.0617	0.5409	0.2549	0.9972	− 0.4294	− 0.0798	0.6943	0.2649
CatBoost	**0.6278**	− 0.2902	− 0.0549	**0.5408**	0.2506	0.9953	− 0.4294	− 0.0803	0.6962	0.2668
LSTM	0.6485	− 0.3372	− 0.0902	0.5819	0.2447	1.0083	− 0.4855	− 0.1176	0.7614	0.2759
GRU	0.6520	− 0.3057	− 0.0773	0.5862	0.2804	1.0177	− 0.3890	− 0.1052	0.7974	0.4083

Table 14 Quartile percent of the prediction error for Dataset 2 (minimum Std. deviation and IQR are bolded)

Model	Quartile percentile for Dataset 2: Kutubdia					Quartile percentile for Dataset 2: Cox's Bazar				
	Std	25%	50%	IQR	75%	Std	25%	50%	IQR	75%
MLR	0.7008	− 0.4395	− 0.0139	0.8436	0.4041	0.7372	− 0.4412	− 0.0076	0.8657	0.4245
Lasso	0.7076	− 0.4460	− 0.0233	0.8483	0.4023	0.7475	− 0.4609	− 0.0149	0.8856	0.4247
Ridge	0.7008	− 0.4397	− 0.0138	0.8439	0.4042	0.7372	− 0.4413	− 0.0076	0.8658	0.4245
Elastic Net	0.7062	− 0.4522	− 0.0226	0.8513	0.3992	0.7471	− 0.4636	− 0.0143	0.8886	0.4250
KNN	0.6226	− 0.3589	− 0.0039	0.7144	0.3556	0.6634	− 0.3722	0.0039	0.7433	0.3711
DT	0.6740	− 0.4009	− 0.0105	0.7798	0.3789	0.7052	− 0.4224	− 0.0109	0.8216	0.3992
RF	0.6433	− 0.3806	− 0.0050	0.7451	0.3645	0.6672	− 0.4013	− 0.0085	0.7703	0.3690
GBR	0.5871	− 0.3304	− 0.0112	0.6538	0.3234	0.6191	− 0.3473	0.0003	0.6915	0.3442
AdaBoost	0.7181	− 0.4642	− 0.0178	0.8704	0.4061	0.7544	− 0.4810	− 0.0311	0.8953	0.4143
XGBoost	0.5856	− 0.3321	− 0.0104	0.6529	0.3208	0.6199	− 0.3510	− 0.0034	0.6966	0.3456
LightGBM	0.5871	− 0.3327	− 0.0109	0.6521	0.3194	0.6192	− 0.3464	− 0.0026	0.6899	0.3436
CatBoost	**0.5749**	− 0.3198	− 0.0052	**0.6369**	0.3171	**0.6119**	− 0.3374	− 0.0007	**0.6730**	0.3356
LSTM	0.6088	− 0.2954	0.0561	0.7048	0.4095	0.6580	− 0.3558	0.0254	0.7682	0.4124
GRU	0.6078	− 0.3200	0.0394	0.7097	0.3897	0.6757	− 0.3779	0.0311	0.8108	0.4329

Various strategies are involved in the reliable operation of a wind power plant to ensure the effectual and firm performance of the wind turbines as well as the overall plant. Accurate prediction of wind speeds lays out informative perceptions that are devoted to the optimization and stability of the operation of plants. Some key techniques regarding the reliable operation of wind power plants are mentioned here (Commission, 2022):

- Variations in wind conditions can be anticipated by the operators using wind speed predictions. The plant can optimize energy production and maximize the efficiency of power generation by balancing the pitch and yaw of the turbines based on predicted wind speeds.
- The employment of advanced control systems to manage loads on the turbines can be enabled using prediction of wind speeds in advance. The operating parameters of the turbines can be adjusted by the control algorithms to guarantee optimal performance and minimize wear and tear, benefiting to the long-term stability of the equipment.
- Accurate prediction of wind speeds can be used to manage the integration of wind power into the electrical grid. Uncertain swings can be anticipated by grid operators in energy production. Thus, proactive measures, such as adjusting energy reserves or activating alternative sources, can be undertaken to maintain grid stability.
- Operators can antedate the time period of increased stress on turbine components using predictions of wind conditions. This allows planning maintenance activities during periods of lower wind speeds, reducing downtime and confirming the reliability of the plant.
- Wind speed forecast can help distribute effective resources, including human resources and spare parts. Operators can maintain inspections, repairs, and maintenance tasks relying on predicted wind conditions. Thus, they can optimize the allocation of resources to enhance the system reliability.
- Grid operators, energy market participants, and plant owners who rely on a stable and predictable energy output for planning and operational decision-making can be anticipated by wind speed predictions.
- Precise wind speed predictions can be used by utilities and grid operators for long-term planning and grid development. Predicting future wind conditions helps in determining the feasible locations for new wind projects and planning the extension of the existing grid infrastructure to assist increase renewable energy capacity.

Conclusion and recommendations

The unpredictability of wind turbine production due to variations in wind speeds poses a challenge for wind power plants. To address these, accurate wind speed forecasting emerges as a pivotal strategy for operational

Table 15 Performance comparison of the suggested models with the models from prior studies

Ref	Year	Region/country	Data resolution	Methods	Best performer	Performance
Shawon et al. (2021)	2021	NA	Hourly	ARMA, ARIMA, SVR, and ANN	Polynomial SVR	RMSE = 0.552 MAPE = 5%
Mohsin et al. (2021)	2021	NA	3– h interval	BNN, and Lasso	BNN	MAPE = 19.01% NMAE = 0.003
Hanoon et al. (2022)	2022	14 regions in Malaysia	Daily	GPR, SVR, and BTs	GPR	RMSE = 0.18144 MSE = 0.03292 NSE = 0.26957 MAE = 0.13498 R^2 = 0.38115
S. Kumar P (2019)	2019	Waterloo, Canada	15-min interval	BPN, BPN with MIFS, RBF, RBF with MIFS, NARX, and NARX with MIFS	NARX with MIFS	RMSE = 0.5814 MAE = 0.4381
Elsaraiti & Merabet (2021)	2021	Halifax, Canada	Hourly	ARIMA, and LSTM	LSTM	RMSE = 3.124 MAE = 2.457
Liu & Chen (2019)	2022	East Jerusalem, Palestine	3-h interval	MLR, ridge, lasso, RF, SVR, and LSTM	RF	MAE = 0.894 MSE = 1.345 MAD = 0.715 R^2 = 0.435
Xie et al. (2021)	2021	Yanqing, and Zhaitan, Beijing, China	Hourly	ARMA, single-variable LSTM, and MV-LSTM	MV-LSTM	RMSE = 1.1460 MAE = 0.8468 MBE = 0.0276 MAPE = 0.6412
Malakout (2023)	2023	Turkey	Monthly	LightGBM, GBR, AdaBoost, Elastic net, lasso, and ensemble method (LightGBM and AdaBoost)	Ensemble method	RMSE = 0.2080 MAE = 0.1410 MAPE = 0.0292 R^2 = 0.997
Krishnaveni et al. (2021)	2021	Las Vegas, USA	Hourly	MLR, Lasso, SVR, and MPFFNN	SVR	MSE = 0.011217 MAE = 0.080115
This study	2023	Kutubdia and Cox's Bazar, Bangladesh	3-h interval	MLR, Ridge, Lasso, Elastic Net, KNN, DT, RF, GBR, AdaBoost, XGBoost, LightGBM, LSTM and GRU	CatBoost	MSE = 0.3744 MAE = 0.4415 R^2 = 0.8670

NREL: National Renewable Energy Laboratory; MLR: multiple linear regression; LR: linear regression; SVR: support vector regression; ARMA: autoregressive moving average; ARIMA: autoregressive integrated moving average; ANN: artificial neural network; GPR: Gaussian progress regression; BTs: bagged regression trees; BNN: Bayesian neural network; BPN: back propagation network; NARX: nonlinear autoregressive model process with exogenous inputs; MIFS: mutual Information feature selection; MPFFNN: multiple perceptron feed-forward neural network; RBF: radial basis function; RF: random forest; LSTM: long short-term memory; MV-LSTM: multivariate long short-term memory; GB: gradient boosting; XGBoost: extreme gradient boosting; CatBoost: category boosting; AdaBoost: adaptive boosting; KNN: K-nearest neighbors; DTR: decision tree regressor; SVM: Support vector machine; RVM: Relevance vector machine; LightGBM: Light gradient boosting machine; R^2: coefficient of determination; RMSE: root mean square error; MAE: mean absolute error; MAD: mean absolute deviation; MAPE: mean absolute percentage error; MBE: mean bias error; MRE: mean relative error; NMAE: normalized mean absolute error; NSE: Nash–Sutcliffe efficiency

stability. Site feasibility and turbine choosing also to rely on the forecasts of wind speed. This study, conducted in the coastal region of Bangladesh, evaluates fourteen ML models for short-term wind speed prediction. Among them, the CatBoost model surpasses other models, demonstrating regression coefficients exceeding 50%–60% for Dataset 1 and surpassing 85% for Dataset 2. This showcases the model's substantial potential for accurate prediction of wind speed in the realm of wind energy potential. Additionally, the research underscores the necessity of site-specific wind speed feasibility studies associated with the capricious nature of wind prior to project implementation. The Weibull model parameters indicate that the wind power density of Cox's Bazar is greater than that of Kutubdia for both observed and predicted speed data. Moreover, the Goldwind model emerges as a viable turbine option with a favorable capacity factor for both locations. While this study has shed light on the dynamics of wind speed forecasting, it is important to acknowledge its limitations.

However, BMD data only record the integer value that caused the round-off error. As a result, all ML models end up doing moderately well on average in predicting the BMD data. In contrast, NASA data display a notable

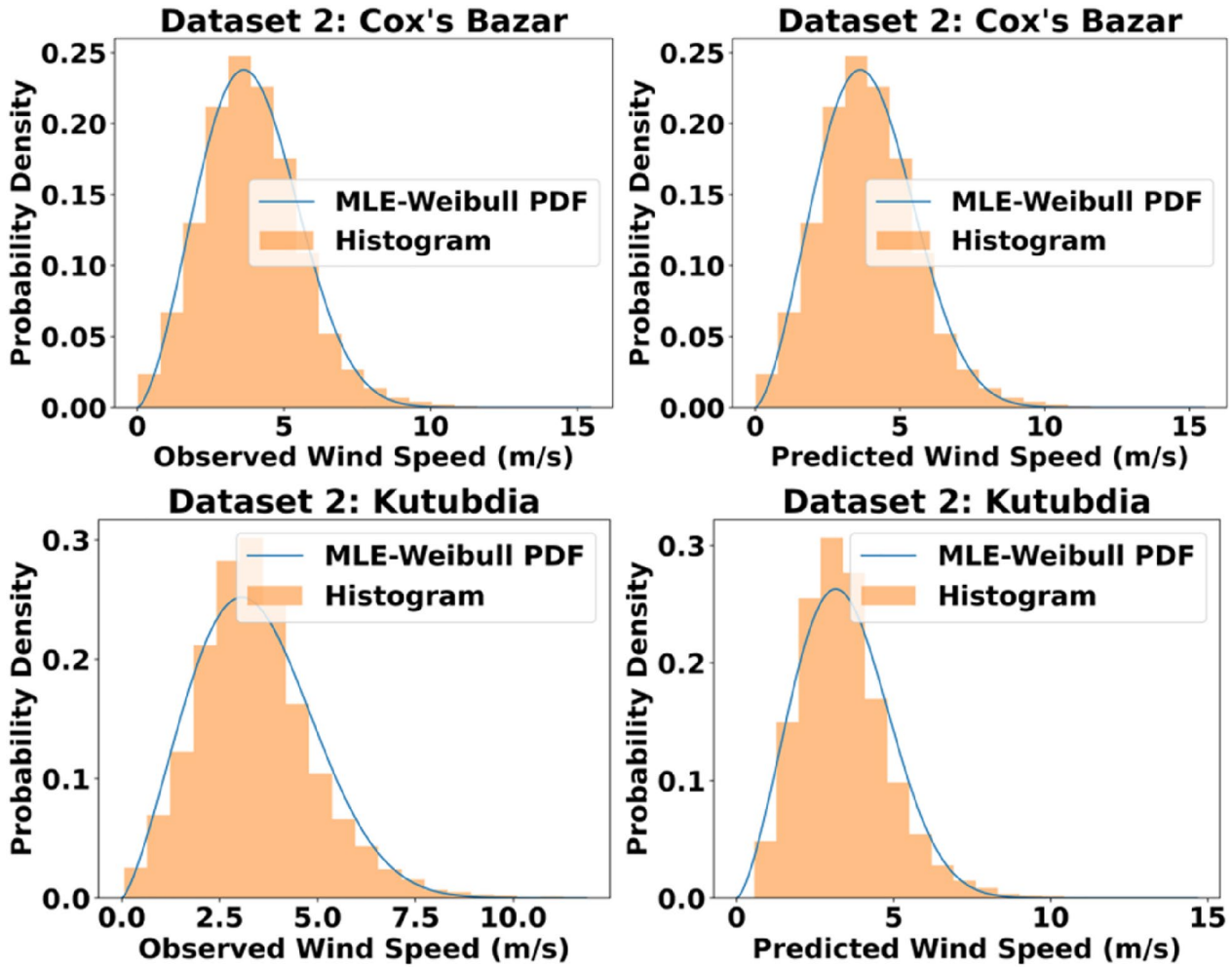


Fig. 10 Probability density function of observed and predicted wind speed data for both stations

Table 16 Weibull k and c parameters, mean wind speed, wind power density, and generation scale at 50 m

Dataset\factor		k	c	Vavg(m/s)	Wind power density (W/m2)	Wind power class	Generation scale
Dataset 2 (observed)	Kutubdia	2.3794	5.6301	4.9903	124.6224	1	Poor
	Cox's Bazar	2.4376	6.3896	5.6660	179.0635	1	Poor
Dataset 2 (predicted)	Kutubdia	2.5164	5.6347	5.003	120.2248	1	Poor
	Cox's Bazar	2.5959	6.3872	5.6729	171.7581	1	Poor

k: shape parameter; c: scale parameter; Vavg: average wind speed

Table 17 Weibull k and c parameters, mean wind speed, wind power density, and generation scale at 120 m

Dataset\factor		k	c	Vavg(m/s)	Wind power density (W/m2)	Wind power class	Generation scale
Dataset 2 (observed)	Kutubdia	2.3795	6.5937	5.8443	200.1749	2	Marginal
	Cox's Bazar	2.4376	7.4830	6.6355	287.6126	2	Marginal
Dataset 2 (predicted)	Kutubdia	2.5164	6.5990	5.8560	193.1091	1	Small
	Cox's Bazar	2.5959	7.4802	6.6437	275.8756	2	Marginal

k: shape parameter; c: scale parameter; Vavg: average wind speed

Table 18 Characteristics of some on-shore wind turbines for the chosen sites

Turbine model	Rotor diameter (m)	Hub height (m)	Rated power (kW)	Cut-in speed (m/s)	Rated wind speed (m/s)	Cut-out speed (m/s)
Gamesa G114-2.0 M	114	93/120/140	2000	2.5	12	25
Enercon E-160 EP5 E1	160	120/166	4600	2.5	12	22
Doosan WinDS3000/100	100	90/site specific	3000	3	12	25
Adwen AD 8–180	180	Site specific	8000	3	12	30
Goldwind GW 155/4500	155	95/110/140/project specific	4500	2.5	10.8	26
Goldwind GW 171/5000	171	100—185/site specific	2500	2.5	9.5	24
Goldwind GW 171/3850	171	100—185/site specific	3850	2.5	8.8	17
Senvion 2.4M114	113	95/120	2430	3	13	18
Vestas V172-7.2 EnVentus	172	112/117/150/164/166/175/site specific	7200	3	12	25
Envision EN140-3.0	140	90 m/110 m/125 m/140 m/site specific	3000	3	12	20
Nordex N149/5.X	149.1	up to 164	5000	3	12	20
Nordex N133/4800 Delta	133.2	78/83/110/site specific	4800	3	12	20

Table 19 Annual energy output and capacity factor of considered turbines for 120 m hub height

Weather station	Kutubdia		Cox's Bazar	
Turbine model	Ea (kWh/yr)	CF (%)	Ea (kWh/yr)	CF (%)
Gamesa G114-2.0 M	3,412,288.31	19.48	4,708,609.98	26.88
Enercon E-160 EP5 E1	7,672,150.08	19.04	10,829,764.53	26.88
Doosan WinDS3000/100	4,543,199.44	17.29	6,605,123.45	25.13
Adwen AD 8–180	12,115,198.50	17.29	17,613,662.95	25.13
Goldwind GW 155 / 4500	9,592,666.43	24.33	13,122,283.71	33.29
Goldwind GW 171 / 5000	7,021,850.59	32.06	9,159,535.18	41.82
Goldwind GW 171 / 3850	12,535,495.38	37.17	15,848,188.27	46.99
Senvion 2.4M114	3,009,007	14.14	4,453,514.73	20.92
Vestas V172-7.2 EnVentus	10,903,678.65	17.29	15,852,296.28	25.23
Envision EN140-3.0	4,543,197.21	17.29	6,604,677.47	25.13
Nordex N149/5.X	7,571,995.36	17.29	11,007,795.78	25.13
Nordex N133/4800 Delta	7,269,115.54	17.29	10,567,483.94	25.13

Ea: annual average energy output; CF: capacity factor

improvement, achieving an accuracy increase of over 20% compared to BMD data despite covering a broader geographical range compared to a specific point location. Training the ML models with parameter tuning with 62,808 data samples consumes more time compared to training a single model. The model can be susceptible to overfitting, particularly in situations with limited sample sizes. In site and turbine selection, a limitation of the study is the exclusive reliance on the MLE-Weibull model for wind resource assessment, without considering the potential use of alternative models employing various optimization methods. Moreover, a constraint of this study lies in the challenge of presenting a detailed illustration of the correlation between reliable wind plant operation and the accuracy of wind speed predictions through data analysis.

Future investigations could extend this paradigm to long-term forecasting, further enhancing the efficacy of wind power ventures. Long-term data analysis could unveil seasonal wind patterns, providing valuable insights for project planning. Advanced ML and DL methods can be adopted, which may help mitigate the uncertainties associated with wind speed prediction. Accurate predictions of wind speed for site and turbine selection necessitate dependable ground station measurements alongside thorough site inspections. Additionally, incorporating additional environmental factors like terrain, land use, and geographical features could enhance predictive accuracy as well as wind resource assessment. Various roughness lengths can be employed in evaluating wind resources when applying the height conversion logarithmic law. Evaluating economic feasibility and conducting thorough environmental impact assessments are crucial steps for comprehensive project planning.

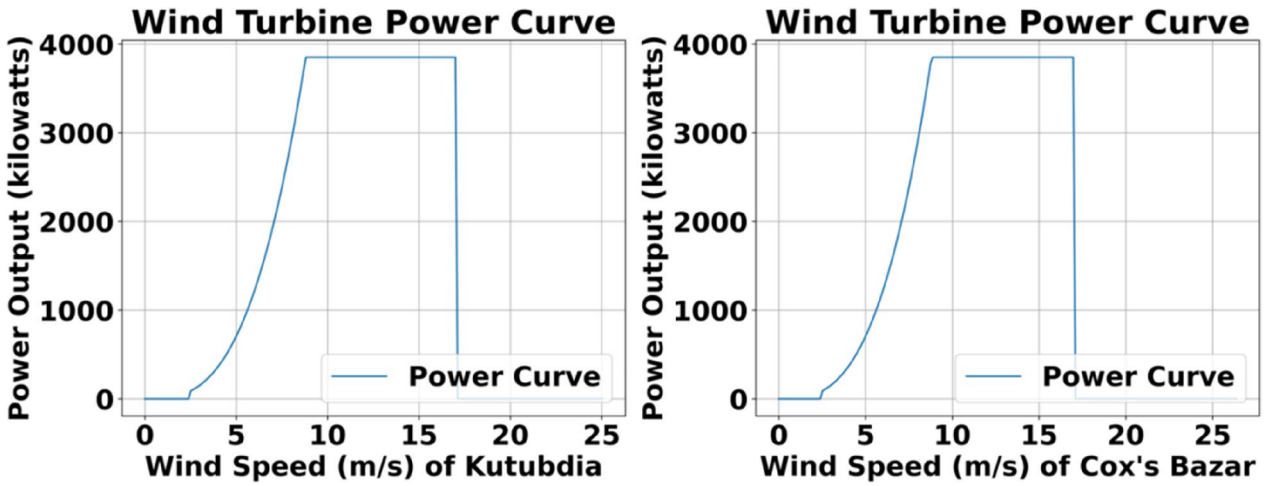


Fig. 11 Power curve of Goldwind GW 171/3850 for 120 m hub height

Author contributions
Conceptualization, T. M; investigation, T. M; methodology, T. M; software, T. M; supervision, G. H, and S. R. S; validation, T. M, G. H, and S. R. S; writing—original draft, T. M; writing—review and editing, T. M and S. R. S; project administration, T. M, G. H, and S. R. S. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Availability of data and materials
On request, we will provide the information.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 23 November 2023 Accepted: 8 January 2024
Published online: 01 February 2024

References

Anjum, L. (2014). Wind resource estimation techniques-an overview. *International Journal of Wind and Renewable Energy, 3*(2), 26–38.
Assareh, E., Poultangari, I., Tandis, E., & Nedaei, M. (2016). Optimizing the wind power generation in low wind speed areas using an advanced hybrid RBF neural network coupled with the HGA-GSA optimization method. *Journal of Mechanical Science and Technology, 30*(10), 4735–4745. https://doi.org/10.1007/s12206-016-0945-4
Babu, Md. T., Nei, H., & Kowser, Md. A. (2022). Prospects and Necessity of Wind Energy in Bangladesh for the Forthcoming Future. *J. Inst. Eng. India Ser. C, 103*(4), 913–929. https://doi.org/10.1007/s40032-022-00834-8
Baloch, M., et al. (2017). A Research on Electricity Generation from Wind Corridors of Pakistan (Two Provinces): A Technical Proposal for Remote Zones. *Sustainability, 9*(9), 1611. https://doi.org/10.3390/su9091611
L. Bauer. Wind turbines database. https://en.wind-turbine-models.com/turbines. Accessed 29 Dec 2023.
Bharani, R., & Sivaprakasam, A. (2022). A meteorological data set and wind power density from selective locations of Tamil Nadu, India: Implication for installation of wind turbines. *Total Environment Research Themes, 3–4*, 100017. https://doi.org/10.1016/j.totert.2022.100017
Chang, W.-Y. (2014). A Literature Review of Wind Forecasting Methods. *J. Power Energy Eng., 02*(04), 161–168. https://doi.org/10.4236/jpee.2014.24023
European Commission. Joint Research Centre., Clean Energy Technology Observatory, Wind energy in the European Union: status report on technology development, trends, value chains and markets : 2022. LU: Publications Office, 2022. Accessed: Dec. 29, 2023. https://data.europa.eu/doi/https://doi.org/10.2760/855840
Darwish, A. S., Shaaban, S., Marsillac, E., & Mahmood, N. M. (2019). A methodology for improving wind energy production in low wind speed regions, with a case study application in Iraq. *Computers & Industrial Engineering, 127*, 89–102. https://doi.org/10.1016/j.cie.2018.11.049
Das, N. K., Chakrabartty, J., Dey, M., Gupta, A. K. S., & Matin, M. A. (2020). Present energy scenario and future energy mix of Bangladesh. *Energy Strategy Reviews, 32*, 100576. https://doi.org/10.1016/j.esr.2020.100576
Elsaraiti, M., & Merabet, A. (2021). A Comparative Analysis of the ARIMA and LSTM Predictive Models and Their Effectiveness for Predicting Wind Speed. *Energies, 14*(20), 6782. https://doi.org/10.3390/en14206782
Energy Institute - Statistical Review of World Energy (2023) – with major processing by Our World in Data. Share of primary energy consumption that comes from wind power – Using the substitution method [dataset]. Energy Institute, "Statistical Review of World Energy" [original data]. https://ourworldindata.org/grapher/wind-share-energy. Retrieved 18 Jan 2024.
Hanoon, M. S., et al. (2022). Wind speed prediction over Malaysia using various machine learning models: Potential renewable energy source. *Eng. Appl. Comput. Fluid Mech., 16*(1), 1673–1689. https://doi.org/10.1080/19942060.2022.2103588
Hulio, Z. H. (2021). Assessment of Wind Characteristics and Wind Power Potential of Gharo, Pakistan. *Journal of Renewable Energy, 2021*, 1–17. https://doi.org/10.1155/2021/8960190
F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., Automated Machine Learning: Methods, Systems, Challenges. in The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2019. doi: https://doi.org/10.1007/978-3-030-05318-5.
Iea, "Renewable electricity – analysis," IEA, https://www.iea.org/reports/renewable-electricity (accessed Jun. 7, 2023).
Islam, M. S., Islam, A., Hasan, M. M., & Khan, A. H. (2013). Feasibility study of wind power generation in Bangladesh: A statistical study in the perspective of wind power density and plant capacity factor. *International Journal of Renewable Energy Research, 3*(3), 476–487.

Jagdale, K. R., Shelke, C. J., Achary, R., Wankhede, D. S., & Bhandare, T. V. (2022). Artificial Intelligence and its Subsets: Machine Learning and its Algorithms, Deep Learning, and their Future Trends. *JETIR, 9*, 112–117.

T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analysis of Gradient Boosting, Adaboost, Catboost Algorithms in Water Quality Classification," JuTISI, vol. 8, no. 2, Aug. 2022, doi: https://doi.org/10.28932/jutisi.v8i2.4906

Jiang, H., Wang, J., Wu, J., & Geng, W. (2017). Comparison of numerical methods and metaheuristic optimization algorithms for estimating parameters for wind energy potential assessment in low wind regions. *Renewable and Sustainable Energy Reviews, 69*, 1199–1217. https://doi.org/10.1016/j.rser.2016.11.241

Khadem, S. K., & Hussain, M. (2006). A pre-feasibility study of wind resources in Kutubdia Island, Bangladesh. *Renewable Energy, 31*(14), 2329–2341. https://doi.org/10.1016/j.renene.2006.02.011

S. Krishnaveni, J. Singh, K. Verma, A. Pachaury, G. Kashyap, and A. Bhatia, "A Machine Learning Approach for Wind Speed Forecasting," in 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India: IEEE, Mar. 2021, pp. 507–512. doi: https://doi.org/10.1109/ICACITE51222.2021.9404563.

Liu, H., & Chen, C. (2019). Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy, 249*, 392–408. https://doi.org/10.1016/j.apenergy.2019.04.188

Liu, Z., Zhu, Z., Gao, J., & Xu, C. (2021). Forecast Methods for Time Series Data: A Survey. *IEEE Access, 9*, 91896–91912. https://doi.org/10.1109/ACCESS.2021.3091162

Malakouti, S. M. (2023). Improving the prediction of wind speed and power production of SCADA system with ensemble method and 10-fold cross-validation. *Case Stud. Chem. Environ. Eng., 8*, 100351. https://doi.org/10.1016/j.cscee.2023.100351

Sana Mohsin, Sofia Najwa Ramli, and Maria Imdad, "Medium-Term Wind Speed Prediction using Bayesian Neural Network (BNN)," Int. J. Syst. Innov., vol. 6, no. 5, Sep. 2021, doi: https://doi.org/10.6977/IJoSI.202109_6(5).0002.

Mollick, T., Hashmi, G., & Sabuj, S. R. (2023). A perceptible stacking ensemble model for air temperature prediction in a tropical climate zone. *Discov Environ, 1*, 15. https://doi.org/10.1007/s44274-023-00014-0.]

S. Nurunnahar, D. B. Talukdar, R. I. Rasel, and N. Sultana, "A short term wind speed forcasting using SVR and BP-ANN: A comparative analysis," in 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh: IEEE, Dec. 2017, pp. 1–6. doi: https://doi.org/10.1109/ICCITECHN.2017.8281802.

"POWER | Data Access Viewer." Accessed: Dec. 29, 2023. [Online]. Available: https://power.larc.nasa.gov/data-access-viewer/

S. Kumar P, "Improved Prediction of Wind Speed using Machine Learning," EAI Endorsed Trans. Energy Web, vol. 6, no. 23, p. 157033, Jun. 2019, doi: https://doi.org/10.4108/eai.13-7-2018.157033.

Salah, S., Alsamamra, H. R., & Shoqeir, J. H. (2022). Exploring Wind Speed for Energy Considerations in Eastern Jerusalem-Palestine Using Machine-Learning Algorithms. *Energies, 15*(7), 2602. https://doi.org/10.3390/en15072602

M. Santhosh, C. Venkaiah, and D. M. Vinod Kumar, "Current advances and approaches in wind speed and wind power forecasting for improved renewable energy integration: A review," Eng. Rep., vol. 2, no. 6, Jun. 2020, doi: https://doi.org/10.1002/eng2.12178.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci., 2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

S. M. R. H. Shawon, M. A. Saaklayen, and X. Liang, "Wind Speed Forecasting by Conventional Statistical Methods and Machine Learning Techniques," in 2021 IEEE Electrical Power and Energy Conference (EPEC), Toronto, ON, Canada: IEEE, Oct. 2021, pp. 304–309. doi: https://doi.org/10.1109/EPEC52095.2021.9621686.

Shi, J., et al. (2022). Wind Speed Forecasts of a Mesoscale Ensemble for Large-Scale Wind Farms in Northern China: Downscaling Effect of Global Model Forecasts. *Energies, 15*(3), 896. https://doi.org/10.3390/en15030896

S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL: IEEE, Dec. 2018, pp. 1394–1401. doi: https://doi.org/10.1109/ICMLA.2018.00227.

Siddique, A. H., Tasnim, S., Shahriyar, F., Hasan, M., & Rashid, K. (2021). Renewable Energy Sector in Bangladesh: The Current Scenario, Challenges and the Role of IoT in Building a Smart Distribution Grid. *Energies, 14*(16), 5083. https://doi.org/10.3390/en14165083

Talekar, B. (2020). A Detailed Review on Decision Tree and Random Forest. *Biosci. Biotech. Res. Comm, 13*(14), 245–248. https://doi.org/10.21786/bbrc/13.14/57

Tao, S., Li, B., Ren, C., & Mao, B. (2022). Grain Temperature Prediction based on Gated Recurrent Unit Deep Fusion Model. *Procedia Computer Science, 214*, 763–770. https://doi.org/10.1016/j.procs.2022.11.239

Tarek, Z., et al. (2023). Wind Power Prediction Based on Machine Learning and Deep Learning Models. *Comput. Mater. Contin., 74*(1), 715–732. https://doi.org/10.32604/cmc.2023.032533

Waqas Khan, P., Byun, Y.-C., Lee, S.-J., & Park, N. (2020). Machine learning based hybrid system for imputation and efficient energy demand forecasting. *Energies, 13*(11), 2681. https://doi.org/10.3390/en13112681

Xie, A., Yang, H., Chen, J., Sheng, L., & Zhang, Q. (2021). A Short-Term Wind Speed Forecasting Model Based on a Multi-Variable Long Short-Term Memory Network. *Atmosphere, 12*(5), 651. https://doi.org/10.3390/atmos12050651

Yousuf, M. U., Al-Bahadly, I., & Avci, E. (2019). Current Perspective on the Accuracy of Deterministic Wind Speed and Power Forecasting. *IEEE Access, 7*, 159547–159564. https://doi.org/10.1109/ACCESS.2019.2951153

Zhang, Y., et al. (2022). The prediction of spark-ignition engine performance and emissions based on the SVR algorithm. *Processes, 10*(2), 312. https://doi.org/10.3390/pr10020312

Publisher's Note