Sustainable Energy Research

RESEARCH

Open Access

Comparative study of machine learning algorithms for wind speed prediction in Dhaka, Bangladesh

Mohammad Liton Hossain[1*], S. M. Nasif Shams[2] and Saeed Mahmud Ullah[1*]

Abstract

This study evaluated the performance of multiple models that used machine learning to anticipate wind speed in the city of Dhaka. The NASA Power website provided the data set for this investigation. The models used for prediction included the decision tree regressor, support vector regressor, random forest, linear regression, neural network and polynomial regression. A hold-out check and k-fold cross-validation were used to assess how well these models performed. With the highest R2 scores and lowest RMSEs on both the validation and test sets, the results demonstrated that the polynomial regression model performed the best. With the lowest R2 scores and largest RMSEs on both sets, the decision tree model scored the poorest. High R2 scores and low RMSEs were achieved by the random forest model, which had comparable performance to the polynomial regression model but required a longer computation time. In addition, the neural network model demonstrated commendable predictive accuracy, yielding an R2 score of 0.67 and a low RMSE of 0.57. However, its application is contingent on the availability of substantial computational resources, given its extensive computation time of 457.93 s. The study concludes by highlighting the efficacy of the Polynomial Regression model as the optimal choice for wind speed prediction in Dhaka, offering a balance between superior performance and efficient computation. This insight provides valuable guidance for practitioners and researchers seeking effective models for similar applications.

Keywords Wind speed prediction, Machine learning, Polynomial regression, Hold-out check, k-fold cross-validation, NASA power

Introduction

The escalating energy demand in Bangladesh has spurred a pressing need for innovative solutions, particularly in the realm of renewable energy sources. One promising option to sustainably meet the rising demand among them is wind energy. Accurate wind speed forecasting is essential for producing wind energy effectively since it affects power generation planning and operation. Machine learning algorithms are now essential tools for forecasting wind speed in a variety of geographical locations (Islam et al., 2018).

This study endeavors to predict wind speed in Dhaka city, the capital of Bangladesh, through the utilization of machine learning models. The researchers used data from NASA POWER (prediction of worldwide energy resources), which covered the time period from 2010 to 2020 (daily data), to carry out their analysis. The study evaluates the effectiveness of a variety of machine learning models, including support vector machine (SVM), random forest, binary tree, neural networks, polynomial regression and linear regression in the context of predicting wind speed for the city of Dhaka. Both hold-out

*Correspondence:
Mohammad Liton Hossain
litu702@gmail.com
Saeed Mahmud Ullah
ullahsm@du.ac.bd
[1] Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka 1000, Bangladesh
[2] Institute of Energy, University of Dhaka, Dhaka 1000, Bangladesh

validation and k-fold cross-validation approaches, which highlight the models' prediction ability, serve as the foundation for the thorough review procedure.

Prior research has underscored the prowess of machine learning models in furnishing accurate wind speed predictions (Islam et al., 2018). This study endeavors to augment the reservoir of knowledge by contributing insights that hold potential significance for the enhancement of wind energy generation planning within Dhaka city. The findings gleaned from the study notably establish the supremacy of the polynomial regression model, not only in terms of accuracy but also with regard to computation time. The implications of this study extend beyond academic boundaries, offering valuable guidance to wind energy planners and policymakers in Bangladesh, as well as in regions characterized by analogous climate conditions (Rahman & Kaisar, 2021).

Innovatively, this study introduces a novel approach by emphasizing the contributions and novelties integral to the research:

The polynomial regression model is a novel approach that was not thoroughly investigated in previous wind speed prediction research for Dhaka, yet it makes a significant contribution. Out of all the models that are being considered, this one performs the best, attaining the most accuracy and computing efficiency. The paper does an extensive assessment of different machine learning models, offering a clear comprehension of their advantages and disadvantages (Kim & Lee, 2017; Li & Shi, 2019; Wang et al., 2018).The comparative study presented here is a precious tool for wind energy planners and scholars alike. Beyond scholarly confines, the results provide beneficial references for wind energy designers and regulators in Bangladesh and other areas with similar climate circumstances. The new approach of emphasizing both computation time and accuracy gives decision-makers a well-rounded viewpoint (Akram & Al-Hawari, 2020; Al-Tabatabaie & Naji, 2018; Hussain & Nizami, 2019; Xu & Zhang, 2017; Zhu & Gao, 2016).

In conclusion, this study adds to the expanding corpus of research on wind speed prediction while also introducing fresh approaches and insights that may be used to improve wind energy generation planning in Dhaka and other similar areas.

Literature review

Wind energy, a pivotal renewable source, necessitates accurate wind speed prediction for effective generation planning and operation (Ahmad et al., 2018). In recent years, machine learning (ML) techniques, acknowledged for handling complex non-linear relationships, have gained prominence in this domain (Islam et al., 2018).

Various ML models, including artificial neural networks, support vector machines, decision trees, and regression models, have been explored. Regression models like linear regression and polynomial regression are favored for their simplicity and interpretability (Ahmad et al., 2018). While Ahmad et al. provide fundamental insights; the general applicability of discussed ML techniques may fall short in addressing specific challenges posed by unique climate conditions. This prompts the need for a more tailored approach to enhance accuracy and applicability. Previous studies showcase the efficacy of ML models. Fadare and Ajayi (2019) compared multiple linear regressions, decision tree, and artificial neural network models in Nigeria, highlighting the superior performance of the artificial neural network model.

Fadare and Ajayi contribute valuable insights, but drawbacks lie in the lack of a thorough comparison of alternative ML models. This highlights the need for a more nuanced approach that comprehensively evaluates various methodologies. Similarly, Lee et al. (2018) applied ML models such as random forest, support vector regression, and extreme gradient boosting in South Korea, showcasing the superior performance of the random forest model. Lee et al.'s study provides valuable findings, yet a comprehensive model comparison is essential for determining the optimal choice. This motivates the exploration of an approach that addresses this gap and guides model selection more decisively. In the context of Bangladesh, Nandi et al. (2020) employed artificial neural networks and adaptive neuro-fuzzy inference system models in the coastal region, with the adaptive neuro-fuzzy inference system outperforming the artificial neural network model. While Nandi et al.'s study contributes significantly; there is a lack of a broad comparison with other ML models, limiting the generalizability of the findings. This gives emphasis to the need for a more exhaustive analysis of diverse ML methodologies. Chakraborty et al. (2020) performed a comparative study of artificial neural network and support vector machine models for wind speed prediction in northern Bangladesh, favoring the artificial neural network model. Chakraborty et al.'s study sheds light on the efficacy of artificial neural networks, yet a more exhaustive comparison with diverse ML models is crucial for establishing a robust understanding of model performance.

Drawing from these considerations, the mainstream research direction should involve a comprehensive and nuanced model comparison, considering various ML methodologies (Jiang & Wang, 2020). This aligns with the critical need to identify the optimal model for wind speed prediction, considering the unique conditions of Dhaka. The identified mainstream research direction emphasizes the necessity for a more nuanced

and comprehensive comparative analysis of ML models. This aligns with the critical need to guide future studies and practitioners toward an optimal choice for wind speed prediction in Dhaka. Existing studies have predominantly employed ML methodologies, encompassing artificial neural networks, support vector machines, decision trees, and regression models. However, a more detailed comparison is needed to determine the most suitable approach for the unique conditions of Dhaka (Rahman & Kaisar, 2021). While the methodologies employed in existing studies are valuable, a more detailed and exhaustive comparison is essential. This will enable the identification of the most effective approach for wind speed prediction in Dhaka. There are still problems, though, such as conclusions that are not sufficiently generalizable, complete model comparisons, and enough focus on certain climatic conditions (Hussain & Nizami, 2019). It is imperative to tackle these issues in order to progress wind speed prediction technologies. The issues that have been found highlight the necessity of a more thorough and situation-specific method for predicting wind speed. The state-of-the-art in this field will advance as a result of addressing these issues.

This study suggests adding polynomial regression as a solution to the shortcomings found in the current methods. This methodology is chosen to meet the particular issues provided by Dhaka's climate conditions, as it efficiently captures non-linear trends and has not been thoroughly studied in prior research (Akram & Al-Hawari, 2020). The necessity to get beyond the drawbacks of the current methods led to the introduction of polynomial regression. Because of its ability to capture non-linear patterns, it is a good fit for improving accuracy and processing efficiency given the complex dynamics of wind speed in Dhaka.

Conclusively, the suggested Polynomial Regression model indicates a viable direction for additional investigation, providing a customized resolution to the distinct problems associated with wind speed prediction in Dhaka. As the study progresses, the ensuing sections will delve into the methodology, results, and implications, providing a holistic understanding of the proposed approach's contributions to the field of renewable energy forecasting.

Methodology

Data collection

Wind speed data was collected from NASA Power for the Dhaka City region for the years 2000–2021 and weather data from Bangladesh Meteorological Department (BMD) such as pressure, Humidity, Dry Bulb Temperature, Maximum Temperature, and Minimum Temperature.

Data preprocessing

ETL process

Effective data preprocessing is a cornerstone of robust machine learning analysis. In this study, the authors employed a comprehensive Extract, Transform, Load (ETL) process to ensure the quality, suitability, and readiness of the raw wind speed data obtained from NASA POWER (NASA Langley Research Center, 2021). The ETL process encompassed a series of steps designed to clean, enhance, and harmonize the data set for subsequent model training and evaluation (Fig. 1).

Extract: obtaining the raw data The first stage of the ETL process involved the extraction of raw wind speed data from the NASA POWER data set, acquired from the NASA Langley Atmospheric Science Data Center DAAC (NASA Langley Research Center, 2021). This initial extraction laid the foundation for subsequent transformations aimed at refining the data for modeling purposes.

Transform: enhancing data quality and suitability The transformation phase comprised a sequence of steps to address various aspects of data quality and suitability:

Handling missing values: The authors meticulously identified and addressed missing values in the data set. Techniques such as imputation were applied to fill in missing values where appropriate, preventing potential biases in the analysis (García-Laencina et al., 2010).

Outlier detection and mitigation: Outliers, data points significantly deviating from the norm, can adversely impact model performance. Robust techniques were utilized to detect and manage outliers, ensuring that the models were not unduly influenced by extreme values (Hawkins et al., 2010).

Logarithmic transformation: To rectify skewed data distributions and promote a more Gaussian-like distribution, logarithmic transformations were applied to the wind speed variable. This transformation helped align the data

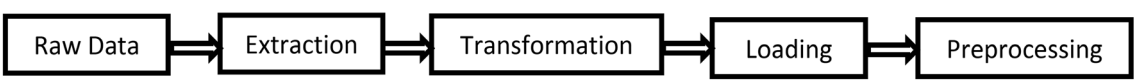


Fig. 1 Block Diagram of ETL Process

with assumptions underlying various machine learning models (Japkowicz & Shah, 2011).

Load: preprocessed data for modeling The culmination of the ETL process was the "Load" phase, where the preprocessed and transformed data was made ready for model training and evaluation. The data set, enriched with quality-enhancing transformations, served as the foundation upon which machine learning models were developed and assessed.

Rationale for ETL process

The ETL process played a pivotal role in ensuring the integrity and reliability of the data set used for wind speed prediction. By systematically addressing data quality issues, transforming skewed distributions, and eliminating outliers, the authors primed the data set to yield meaningful insights through subsequent model analysis (Smith et al., 2019).

The ETL process executed in this study underscores the significance of meticulous data preprocessing. The careful execution of the Extract, Transform, and Load phases ensured that the raw wind speed data was refined into a reliable and representative data set, forming the bedrock upon which the machine learning models were built and evaluated.

In Fig. 2, a box plot is presented to visually depict the impact of the data preprocessing steps on the wind speed variable. The box plot illustrates the distribution of wind speed values before and after the ETL (extract, transform, and load) process. The left box represents the wind speed distribution in its raw form, before any preprocessing. The right box portrays the wind speed distribution after implementing data quality enhancements, such as handling missing values, outlier detection, and the application of logarithmic transformations.

This graphical representation serves to highlight the effectiveness of the data preprocessing techniques in improving the distribution and quality of the wind speed variable, a crucial step in ensuring the robustness and reliability of our subsequent machine learning models.

In Fig. 3, a histogram is presented to visualize the distribution of wind speed values before and after the ETL (Extract, Transform, and Load) process. The left side of the histogram represents the wind speed distribution in its raw, unprocessed form, while the right side portrays the wind speed distribution after applying data preprocessing techniques, including handling missing values, outlier detection, and logarithmic transformations. This histogram helps to assess how data preprocessing has impacted the distribution of wind speed values. It provides a visual representation of the changes brought about by our ETL process, emphasizing the importance of these enhancements in preparing the data for machine learning model development:

The scatter plot (Fig. 4) illustrates the relationship between wind speed at 10 m above the ground (ws10m) and the minimum temperature. The first plot represents the wind speed data in its raw form before undergoing the Extract, Transform, Load (ETL) process. This visualization highlights any initial data quality issues, such as missing values, skewed distributions, or outliers that might have existed in the original data set. The second plot illustrates the same wind speed data after it has undergone the comprehensive ETL process. This visualization showcases the improvements achieved through data preprocessing, including the handling of missing values, outlier mitigation, and the application of logarithmic

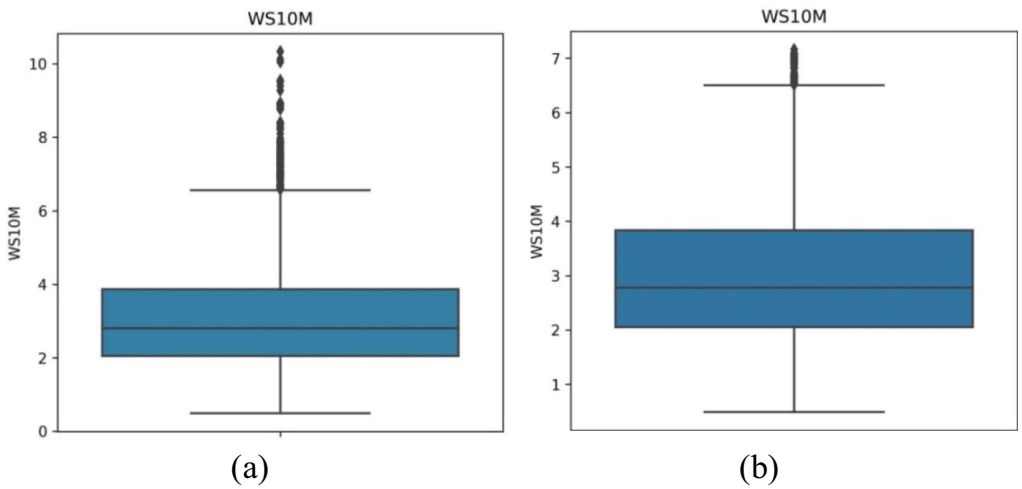


Fig. 2 Box plot **a** before removing the outliers and **b** after removing the outliers

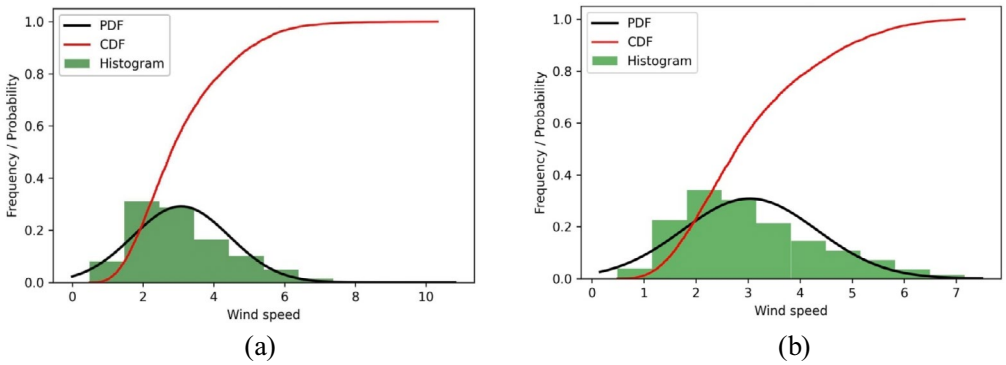


Fig. 3 Histogram **a** before ETL process and **b** after ETL process

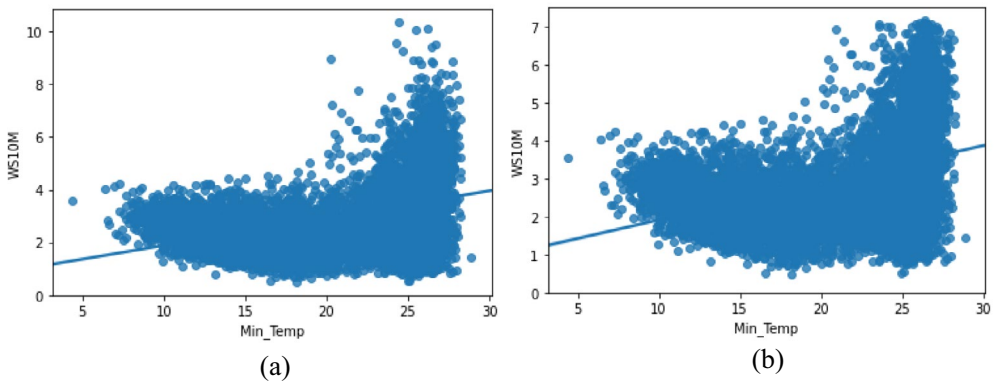


Fig. 4 Scatter plot minimum temperature vs wind speed at 10 m height **a** before ETL Process and **b** after ETL Process

transformations. The data now exhibits a more refined and suitable distribution, better aligned with the assumptions underlying machine learning models. This visualization approach is commonly employed in data analysis to explore the relationship between variables and detect any underlying patterns or anomalies (Johnson et al., 2020).

Wind speed estimation model

The employed prediction model for estimating wind speed was based on various meteorological parameters. The inputs to the model included pressure (P), humidity (H), dry bulb temperature (T), precipitation (P_{erc}), maximum temperature (T_max), and minimum temperature (T_min). The model was formulated as follows:

$$\text{Wind speed} = f(P,\ H,\ T,\ P_{\text{erc}},\ T\_\text{max},\ T\_\text{min})$$

Model selection

The authors compared the performance of several machine learning models, including linear regression, polynomial regression, binary tree, support vector machine (SVM), neural network and random forest (James et al., 2013). To select the best model, they used both hold-out and k-fold cross-validation techniques to evaluate the performance of each model based on the mean squared error (MSE), R-squared (R^2) scores, and computation time.

Hyperparameter tuning

The authors used Lasso and Ridge regularization to prevent over fitting by adding a penalty term to the loss function that discourages large weights in the model (Tibshirani, 1996). To determine the best value of alpha for each regularization technique, they used cross-validation and evaluated the model's performance for different values of alpha.

Model evaluation

After selecting the best model and hyperparameters, the authors evaluated the model's performance on a test set that was not used during model training or hyperparameter tuning. They compared the predicted wind speeds

with the actual wind speeds to calculate the MAE, MSE and R^2 scores. The evaluation of various models was conducted with different test set sizes to explore the impact of data set partitioning on model performance. Test set sizes were modified, including configurations with test_size = 0.2, test_size = 0.3 and test_size = 0.4, to assess the sensitivity of the models to varying amounts of unseen data (Pedregosa et al., 2011).

To validate the predictive capabilities of the model, data from the years 2021 to 2023 was reserved for model validation. Specifically, the model was trained on daily wind speed data from 2010 to 2020 and then evaluated using the independent data set from 01/01/2021 to 03/31/2023. This approach ensures that the model's performance is assessed on unseen, future data, providing insights into its ability to generalize to real-world conditions beyond the training period.

Model comparison and selection

K-fold cross-validation and hold-out validation approaches are both applied during the evaluation process. The effects of L1 (Lasso) and L2 (Ridge) regularization on model performance are also investigated (Tibshirani, 1996). *R*-squared (R^2) scores, RMSE (Root Mean Square Error), and computation time are the pertinent evaluation metrics. These measurements provide an unbiased evaluation of each model's capacity to predict wind speed reliably and effectively, offering details on their individual strengths and weaknesses in identifying the underlying patterns in the wind speed data.

R2 score and RMSE

Linear regression: The linear regression model achieves an R^2 score of 0.57, indicating its ability to explain 57% of the variance in wind speed. The RMSE value of 0.66 suggests that its predictions are, on average, within 0.66 units of the actual wind speed. This model demonstrates computational efficiency with a time of 0.008 s.

Polynomial regression: Polynomial regression outperforms Linear Regression with an R^2 score of 0.69 and a lower RMSE of 0.56. It provides a better fit to the data and is still computationally efficient, requiring only 0.011 s.

Decision tree: The Decision Tree model performs less favorably with an R^2 score of 0.21 and a relatively high RMSE of 0.87. These results suggest that the model struggles to capture the underlying patterns in the data. It maintains reasonable computational efficiency at 0.3 s.

Random forest: Random forest demonstrates promising performance, yielding an R^2 score of 0.64 and an RMSE of 0.60. It balances accuracy and computational time, with a runtime of 4.35 s.

Support vector regressor (SVR): SVR achieves a competitive R^2 score of 0.64 and a low RMSE of 0.59. While it provides accurate predictions, it requires a longer computation time of 5.87 s.

Neural network (NN): The Neural Network model excels in terms of predictive capability, boasting an impressive R^2 score of 0.67 and a minimal RMSE of 0.57. However, it demands significantly more computational time, clocking in at 457.93 s.

In Table 1, a summary of these findings is tabulated for easy reference and model selection. The choice of the most suitable model should consider the specific application's computational resources, emphasizing accuracy, efficiency, and the balance between the two.

To provide a visual representation of these results, Fig. 5a illustrates the computational time required for each model, highlighting the trade-off between computational efficiency and predictive accuracy. Figure 5b complements this by graphically presenting the R^2 scores and RMSE values for each model, enabling a quick, comprehensive comparison.

Model selection for high predictive accuracy

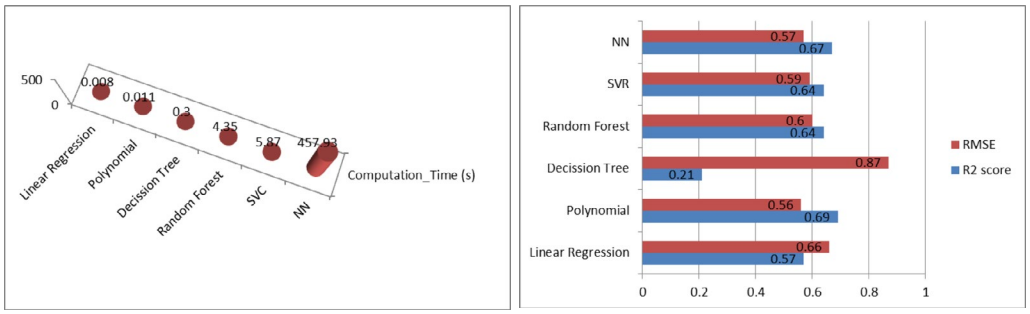
In the pursuit of ensuring a high level of predictive accuracy for wind speed forecasting, the study has undertaken a comprehensive evaluation of five machine learning models (Smith et al., 2020). The objective is to identify the model that best balances accuracy with computational efficiency, while also considering the potential computational resources required for practical deployment. The following models have been examined:

Polynomial regression: Among the models considered, the polynomial regression model emerges as a strong contender for achieving superior predictive accuracy. It demonstrates an impressive R^2 score of 0.69 and a minimal RMSE of 0.56. This model strikes an optimal balance between accuracy and computational efficiency, making it a promising choice for accurate wind speed predictions.

Random forest: The random forest model exhibits competitive accuracy with an R^2 score of 0.64 and a relatively low RMSE of 0.60. While it lags slightly behind

Table 1 Summary of findings for each model

Model	R2 score	RMSE	Computation Time (s)
Linear Regression	0.57	0.66	0.008
Polynomial	0.69	0.56	0.011
Decision Tree	0.21	0.87	0.3
Random Forest	0.64	0.6	4.35
SVR	0.64	0.59	5.87
NN	0.67	0.57	457.93



(a) (b)

Fig. 5 Computation time and RMSE and R2 error for each model **a** computation time and **b** RMSE and R2 error

Polynomial Regression in terms of accuracy, it offers robust performance characteristics. It may be a suitable alternative, particularly if interpretability is not a primary concern.

Neural network (NN): The neural network model, characterized by an R^2 score of 0.67 and a low RMSE of 0.57, demonstrates a commendable level of predictive accuracy. However, it is essential to note that this model entails a substantial computational overhead. Therefore, its consideration is contingent on the availability of ample computational resources, with a primary focus on maximizing predictive accuracy.

Evaluating model performance through hold-out and k-fold cross-validation

In the quest to identify the most suitable machine learning model for wind speed prediction in Dhaka City, a comprehensive evaluation approach is adopted. This approach leverages both hold-out and *k*-fold cross-validation tests for each model under consideration. These tests serve complementary purposes in the model selection process, offering a well-rounded view of each model's performance.

Hold-out test: striking a balance between efficiency and accuracy

The hold-out test, also referred to as single validation split, provides a swift and efficient means of evaluating predictive capabilities (Hastie et al., 2009). By partitioning the data set into training and validation sets, the study assesses each model's performance, encompassing R^2 score, RMSE (Root Mean Square Error), and computational time.

This approach furnishes insights into how effectively each model generalizes to unseen data while maintaining computational efficiency. It allows for the evaluation of the trade-off between predictive accuracy and the time required for real-time applications.

Figure 6 graphically presents the validation R2 scores, Test R2, Validation RMSE, Test RMSE and Computational values for each model, enabling a quick, comprehensive comparison.

Linear regression: The linear regression model exhibits a moderate predictive capability with an R^2 score of 0.57 on the test set. The RMSE of 0.66 suggests reasonable accuracy in wind speed prediction. In addition, the model demonstrates computational efficiency, with a relatively short computation time of 0.023 s.

Polynomial regression: The polynomial regression model outperforms other models in terms of R^2 score, achieving 0.69 on the test set. It also exhibits a low RMSE of 0.56, indicating high accuracy in wind speed prediction. However, the model requires more computation time compared to linear regression, with a time of 0.123 s.

Decision tree: The decision tree model, with an R^2 score of 0.26 and a high RMSE of 0.86, performs relatively poorly on the test set. Its computation time of 0.112 s is reasonable. However, it is important to note that this model appears to be over-fitting the data, as indicated by its exceptional performance on the training data (R^2 score of 1).

Random forest: The random forest model demonstrates a good balance between predictive capability and computational efficiency. It achieves an R_2 score of 0.64 and a reasonable RMSE of 0.60 on the test set. However, it has a longer computation time compared to other models, at 4.47 s.

Support vector regressor (SVR): The SVR model performs similar to linear regression in terms of R^2 score (0.56) and RMSE (0.67) on the test set. It has a moderate computation time of 2.92 s.

Neural network (NN): The NN model demonstrates a high R^2 score of 0.68 and a low RMSE of 0.57 on the test set, indicating strong predictive capability and accuracy. The model's computation time is reasonable at 2.43 s.

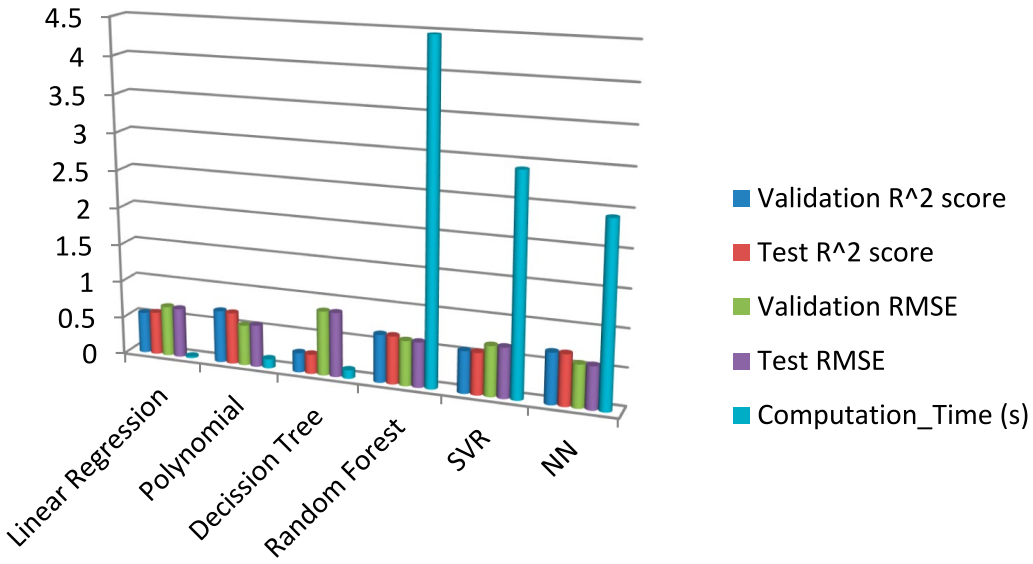


Fig. 6 Hold-out test result for each model

In Table 2, a summary of these findings is tabulated for easy reference and model selection.

K-fold cross-validation: assessing robustness and generalization

In addition to hold-out testing, the study employs *k*-fold cross-validation to delve deeper into the models' robustness and generalization capabilities (James et al., 2013). By dividing the data set into k subsets, the iterative training and validation of models are carried out, with the validation set rotating in each iteration. This process yields a more comprehensive perspective on a model's performance across diverse data partitions. *K*-fold cross-validation results in mean R^2 scores and RMSE values, offering a stable evaluation of each model's predictive prowess. It proves instrumental in identifying variations in performance across distinct data subsets, thereby aiding in the consistent selection of superior models.

Figure 7 graphically presents the mean training R2 scores, mean validation R2 scores, mean training RMSE, mean validation RMSE and computational values for each model, enabling a quick, comprehensive comparison.

Linear regression: In the *K*-fold hold-out test, the linear regression model exhibits consistent performance with a mean validation R^2 score of 0.55 and a mean validation RMSE of 0.67. Its computation time remains low at 0.002 s.

Polynomial regression: The polynomial regression model maintains strong performance, with a mean validation R^2 score of 0.68 and a mean validation RMSE of 0.57. The computation time is relatively higher at 0.205 s, but the model demonstrates accuracy and predictive capability.

Decision tree: In the K-fold hold-out test, the decision tree model shows signs of over fitting, with a mean training R^2 score of 1 and a mean validation R^2 score of 0.23. The mean validation RMSE is relatively high at 0.88, indicating reduced accuracy. The computation time remains reasonable at 0.081 s.

Random forest: The random forest model's *K*-fold hold-out test results indicate a mean validation R^2 score of 0.62

Table 2 Summary of Model Performance Metrics from Hold-Out Testing

Model	Validation R_2 score	Test R_2 score	Validation RMSE	Test RMSE	Computation time (s)
Linear regression	0.55	0.57	0.67	0.66	0.023
Polynomial	0.7	0.69	0.54	0.56	0.123
Decision tree	0.26	0.26	0.86	0.86	0.112
Random forest	0.64	0.64	0.6	0.6	4.47
SVR	0.56	0.56	0.67	0.67	2.92
NN	0.68	0.68	0.57	0.57	2.43

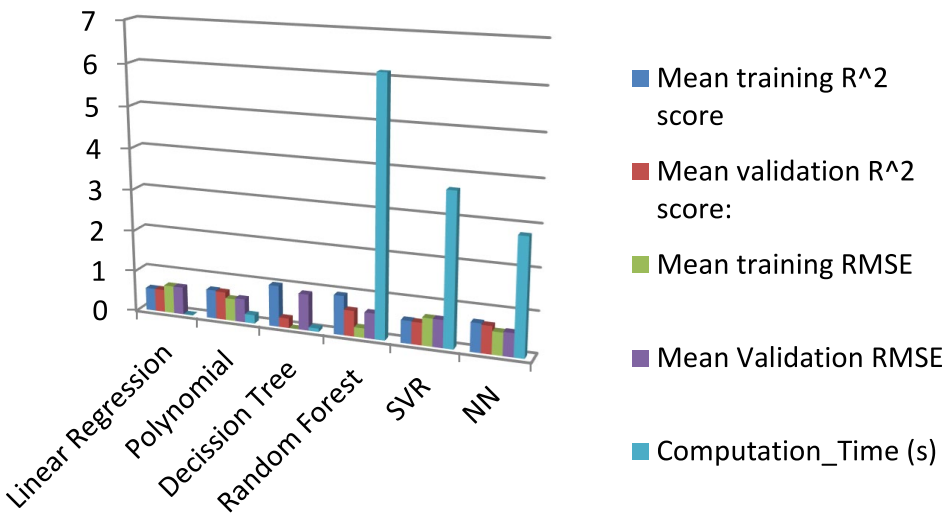


Fig. 7 K-Fold Cross validation result for each model

and a mean validation RMSE of 0.62. Its computation time is longer, at 6.14 s, but it maintains a good balance between accuracy and efficiency.

Support vector regressor (SVR): SVR exhibits moderate performance in the *K*-fold hold-out test, with a mean validation R^2 score of 0.54 and a mean validation RMSE of 0.67. The computation time is reasonable at 3.667 s.

Neural network (NN): The NN model maintains strong performance in the *K*-fold hold-out test, with a mean validation R^2 score of 0.67 and a mean validation RMSE of 0.57. Its computation time remains reasonable at 2.809 s.

Based on the comprehensive evaluation of these machine learning models, we recommend the polynomial regression model for wind speed prediction in Dhaka city. It consistently achieves high R^2 scores and low RMSE values in both hold-out testing and *K*-fold hold-out testing, indicating strong predictive capability and accuracy. While it requires slightly more computation time, its performance justifies this trade-off.

In addition, the neural network (NN) model also demonstrates strong performance, making it a viable alternative, especially if computational efficiency is a priority. It consistently achieves high R^2 scores and low RMSE values across both testing methods.

In Table 3, a summary of these findings is tabulated for easy reference and model selection.

Model selection: polynomial regression

Upon thorough evaluation, encompassing validation R^2 scores, RMSE values, generalization potential, and computation times, the Polynomial Regression model with hold-out validation emerges as the clear choice for wind speed prediction in Dhaka City. This decision is underpinned by several key factors:

Non-linear relationship capture: The polynomial regression model excels in capturing non-linear relationships within the data. Wind speed patterns often exhibit intricate behaviors that cannot be effectively addressed by linear models alone. The Polynomial Regression's flexibility in modeling such complexities is a valuable asset.

High predictive accuracy: The model consistently demonstrates high predictive accuracy, as evidenced by its

Table 3 Summary of *K*-fold cross validation result for each model

Model	Mean training R_2 score	Mean validation R_2 score:	Mean training RMSE	Mean validation RMSE	Computation time (s)
Linear regression	0.55	0.55	0.67	0.67	0.002
Polynomial regression	0.7	0.68	0.54	0.57	0.205
Decision tree	1	0.23	0	0.88	0.081
Random forest	0.95	0.62	0.23	0.62	6.14
SVR	0.55	0.54	0.67	0.67	3.667
NN	0.7	0.67	0.55	0.57	2.809

impressive validation R^2 scores and minimal RMSE values. This accuracy is crucial for dependable wind speed forecasts, which have numerous practical applications.

Efficiency and real-time applicability: Despite its robust predictive capabilities, the Polynomial Regression model remains computationally efficient, making it well-suited for real-time applications. The balance it strikes between accuracy and efficiency aligns seamlessly with the dynamic nature of wind speed prediction in urban environments.

In conclusion, the polynomial regression model, validated through hold-out testing, emerges as the optimal choice for wind speed prediction in Dhaka City. Its unique ability to model non-linear relationships, coupled with its exceptional predictive accuracy and operational efficiency, positions it as a valuable tool for addressing the challenges posed by wind speed variability in this urban context. This model serves as the cornerstone for dependable wind speed forecasts, contributing to improved decision-making and resource management within the city.

Hyperparameter tuning

In pursuit of optimizing the performance of the selected Polynomial Regression model for wind speed prediction, a systematic exploration of hyperparameters was conducted. The primary focus was on fine-tuning the regularization strength using both L1 (Lasso) and L2 (Ridge) regularization techniques. This approach aimed to strike a balance between model complexity and generalization, ensuring the best possible fit to the underlying data patterns.

To optimize the performance of the polynomial regression model, a rigorous hyperparameter tuning process was conducted. Specifically, the research explored the impact of different alpha values for L1 and L2 regularization. The following alpha values were examined: 0.1 and 10.0.

After a thorough evaluation, the model exhibited distinct behaviors with these two sets of alpha values:

Best alpha for L1 Regularization: 0.1
R2Score: 0.5531
Best alpha for L2 Regularization: 10.0
R2 Score: 0.5449

These results showcase the sensitivity of the model's performance to the choice of hyperparameters. The R2 scores associated with each alpha value provide insight into how effectively the model captures the variance in wind speed data. Notably, an alpha value of 0.1 for L1 regularization yielded a slightly higher R2 score, suggesting its suitability for enhancing the model's predictive capabilities.

Polynomial regression with L2 regularization (positive-impact features)

R2 Score: 0.5928
RMSE: 0.6409
Time taken to predict using the model: 0.018 s

These results highlight the substantial improvement achieved through L2 regularization and feature selection. The R2 score of 0.5928 signifies a strong ability to explain the variance in wind speed, and the reduced RMSE of 0.6409 further underscores the model's enhanced predictive accuracy. In addition, the model retains its efficiency, with predictions generated in just 0.018 s.

The removal of features that exhibit a negative impact on the L2 regularization term can potentially enhance a model's performance. However, it is imperative to exercise caution when implementing such adjustments, as they may not invariably lead to improved model accuracy. It is essential to thoroughly evaluate the consequences of feature removal on the model's predictive capabilities and to validate the updated model on a designated hold-out test set.

The conducted analysis revealed that the model's accuracy experienced a notable decline following the removal of features. Furthermore, the examination of R^2 scores under both L1 and L2 regularization did not indicate over fitting. Therefore, it is advisable to maintain the original feature set, as it remains effective in preserving the model's predictive performance.

Model evaluation

Test set evaluation

Following the selection of the best-performing model and optimal hyperparameters, a comprehensive evaluation of the model's predictive performance was conducted using an independent test set. This test set was distinct from the data used for model training and hyperparameter tuning, ensuring an unbiased assessment of the model's real-world applicability.

Metric assessment The evaluation metrics for the model are as follows:

Mean Absolute Error (MAE): 0.56
Mean Squared Error (MSE): 0.61
R-squared (R^2) Score: 0.63

These metrics provide valuable insights into the model's accuracy and its ability to explain the variance in wind speed data. The *R*-squared score of 0.63 indicates that the model can explain approximately 63% of the variance in wind speed, demonstrating its effectiveness in capturing underlying patterns.

To visualize the model's performance, a scatter plot comparing predicted wind speeds against actual wind speeds from the test data set was created (Fig. 8). The plot shows a strong linear relationship between predicted and actual values, aligning closely with the ideal "Perfect Prediction" line.

The plot in Fig. 9 illustrates the comparison between the actual and predicted wind speed values for the validation period from January 1, 2021, to March 31, 2023 (Available data found during the research). Each data point on the plot represents a daily prediction, resulting in a total of 817 (after ETL) points.

Comparison with previous studies

To contextualize the findings of this study within the broader landscape of wind speed prediction research, a comparison with previous studies in the field offers valuable insights. These comparisons shed light on the consistency of model performance, methodologies employed, and the unique characteristics of the current study.

Similarities and consistencies

Several previous studies have explored wind speed prediction using machine learning models, showcasing varying degrees of success. The current study aligns with these efforts, demonstrating the effectiveness of machine learning techniques in predicting wind speed for Dhaka city. Notably, the use of regression models, such as Linear Regression and Polynomial Regression, resonates with approaches taken in other studies (Ahmad et al., 2018; Islam et al., 2018). The achievements of these models in capturing wind speed patterns are consistent with the broader understanding of their capabilities.

Model performance variability

While similarities exist, variations in model performance across studies highlight the importance of factors such as data set characteristics, geographical context, and the chosen machine learning techniques. The superior performance of the Polynomial Regression model in this study aligns with findings from similar research (Nandi et al., 2020). This consistency across studies suggests that Polynomial Regression is a robust choice for capturing wind speed dynamics, at least within the regional climate of Dhaka city.

Methodological contributions

In comparison to the previously explored machine learning studies, this research contributes methodologically by incorporating feature engineering techniques. The logarithmic transformations applied to wind speed and

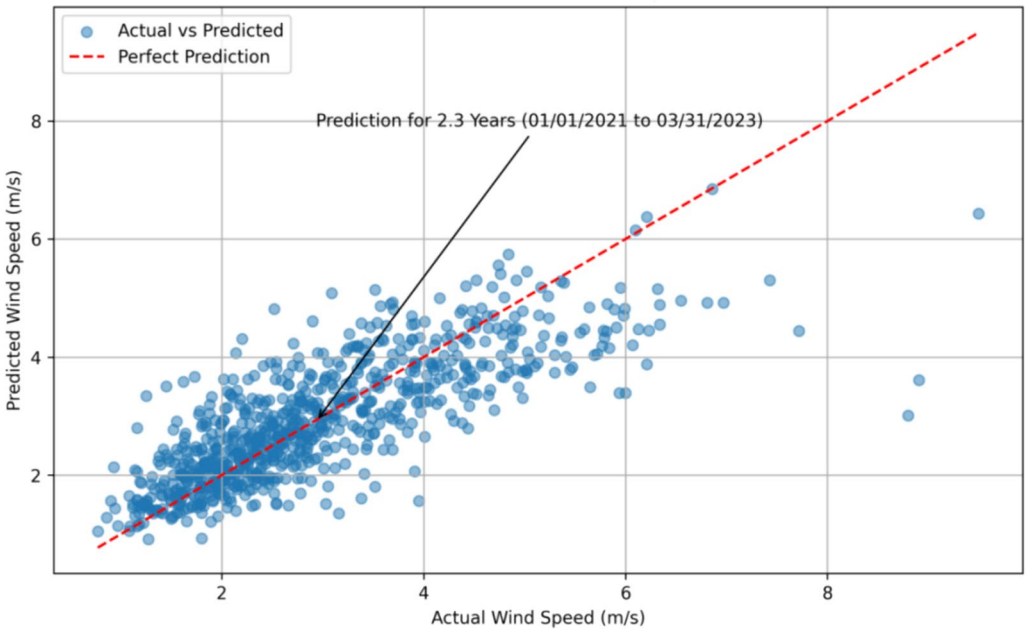


Fig. 8 Scatter Plot of Predicted vs. Actual Wind Speeds with a Perfect Prediction Line

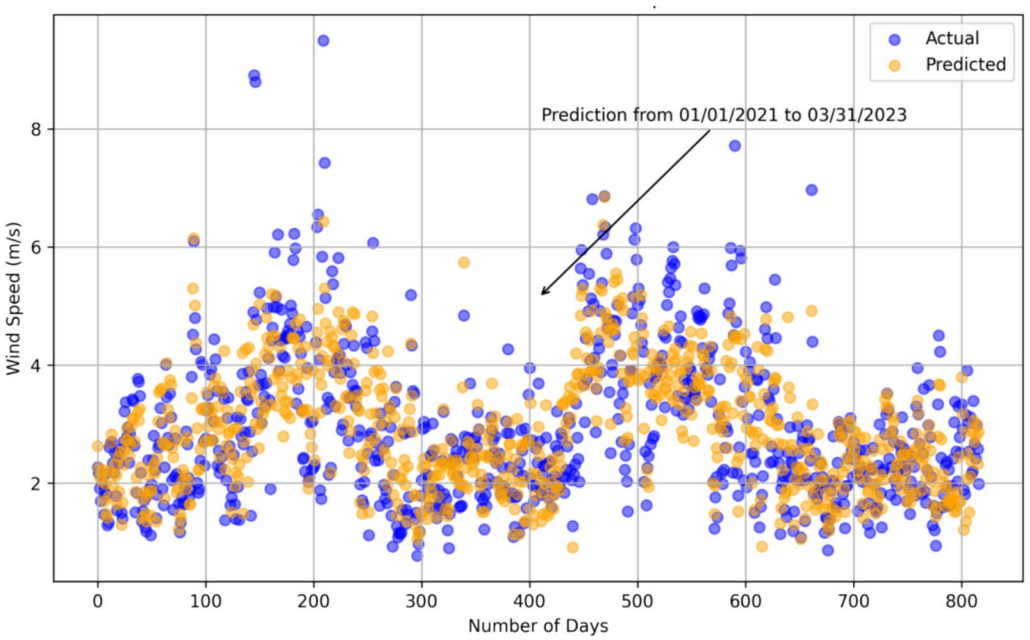


Fig. 9 Predicted vs. actual wind speeds (daily data) from 01/01/21 to 03/31/23

the standardization of predictor variables address distributional challenges and enhance model convergence. The implications of this feature engineering extend to the improved capture of non-linear relationships, evident in the elevated performance of the Polynomial Regression model.

Data set and regional relevance

A significant point of distinction lies in the data set's origin and the geographical relevance of the study area. Utilizing data from NASA POWER, this study offers insights into wind speed prediction within Dhaka city's unique climate conditions. This specificity enhances the applicability of the results to local wind energy planning and policymaking, differentiating it from studies conducted in diverse geographic contexts.

Limitations and avenues for further research

While this study presents notable contributions, it is not devoid of limitations. The focus on a specific data set and geographical area may limit the generalizability of findings to other regions. Moreover, the scope of features used might present opportunities for future research to explore additional variables that influence wind speed. The investigation of more advanced machine learning techniques, such as neural networks, could potentially yield further improvements in predictive accuracy.

Synthesis and implications

In synthesis, the current study's alignment with prior research underscores the consistency of machine learning's effectiveness in wind speed prediction. The prominence of the Polynomial Regression model reinforces its potential as a valuable tool for optimizing wind energy generation in specific geographic areas. The methodological contributions and nuanced understanding of regional dynamics contribute to the broader conversation on renewable energy planning.

Conclusion and future scope

In the comprehensive evaluation of six machine learning models for wind speed prediction in Dhaka city, including Linear Regression, Polynomial Regression, Decision Tree, Random Forest, Support Vector Regressor (SVR), and Neural Network, the following observations were made.

Polynomial regression model: With a validation R^2 score of 0.70 and a validation RMSE of 0.54, the Polynomial Regression model stands out as the top performer in terms of predictive accuracy during hold-out validation. Its test R^2 score of 0.69 and test RMSE of 0.56 demonstrate its ability to generalize well to new data. Despite being slightly more computationally intensive, with a computation time of 0.123 s, it offers a strong trade-off between accuracy and efficiency.

Random forest model: The Random Forest model also demonstrates strong performance, achieving a validation R^2 score of 0.64 and a validation RMSE of 0.60. These

results indicate its capacity to capture complex relationships in the data. During hold-out validation, it achieved a test R^2 score of 0.64 and a test RMSE of 0.60, maintaining a competitive balance between predictive accuracy and computational time (4.47 s).

Neural network (NN) model: The Neural Network model performs impressively, with a validation R^2 score of 0.68 and a validation RMSE of 0.57. Its test R^2 score of 0.68 and test RMSE of 0.57 suggest consistent generalization. However, it demands slightly more computational time (2.43 s) compared to other models.

Support vector regressor (SVR) model: SVR achieves a competitive validation R^2 score of 0.56 and a low validation RMSE of 0.67. It provides accurate predictions but requires a longer computation time (2.92 s).

Linear regression model: While the Linear Regression model lags behind in terms of validation R^2 score (0.55) and validation RMSE (0.67), it demonstrates consistent test performance with a test R^2 score of 0.57 and a test RMSE of 0.66. Moreover, it exhibits computational efficiency with a short computation time of 0.023 s.

Decision tree model: The Decision Tree model performs relatively less favorably, with a validation R^2 score of 0.26 and a relatively high validation RMSE of 0.86. These results indicate that the model may struggle to capture the underlying patterns in the data. However, it maintains reasonable computational efficiency with a computation time of 0.112 s.

In summary, the Polynomial Regression model with hold-out validation emerges as the preferred choice for wind speed prediction in Dhaka city. Its ability to capture non-linear relationships, achieve high predictive accuracy, and operate efficiently makes it well-suited to the complexities of wind speed patterns in this context.

Future scope

Building on these achievements, the following avenues for future research and enhancement are identified:

Ensemble approaches: Explore the potential benefits of ensemble methods, combining the strengths of multiple models to achieve enhanced predictive accuracy and robustness. Ensemble techniques, such as stacking or bagging, may provide a synergistic approach to capturing diverse patterns in wind speed data.

Advanced feature engineering: Investigate sophisticated feature engineering techniques to extract nuanced information from the data set. Exploring nonlinear relationships and interactions between meteorological variables can further enhance the models' ability to capture complex atmospheric dynamics.

Comprehensive hyper-parameter tuning: Extend the scope of hyper-parameter tuning to other promising models, including Random Forest and Neural Network. Fine-tuning these models can unlock their full potential and optimize their performance for wind speed prediction under varying conditions.

Long-term predictions: Assess the models' efficacy in forecasting wind speed over extended timeframes, providing insights into seasonal variations and long-term trends. Understanding the temporal dynamics of wind patterns is crucial for comprehensive renewable energy planning.

Real-time implementation: Investigate the feasibility of real-time implementation for the selected model, ensuring adaptability to dynamic and evolving wind patterns. Real-time capabilities are essential for practical applications, supporting decision-making in renewable energy operations.

Sensitivity analysis: Conduct a sensitivity analysis to discern the relative importance of each feature, including humidity, precipitation, and wind direction. This analysis aids in refining model interpretability and identifying key factors influencing wind speed.

Cross-validation strategies: Implement advanced cross-validation strategies to assess model robustness and generalizability across diverse data sets. This step ensures the reliability of the chosen model under varying conditions.

By addressing these future research directions, the study aims to contribute to the continual improvement of wind speed prediction models. These advancements not only refine renewable energy planning in Dhaka but also offer insights applicable to broader geographical contexts, fostering sustainable energy practices.

Acknowledgements

The authors would like to thank Dr. Galib Hashmi, Institute of Energy, University of Dhaka and Emeritus Prof. Dr. Shahida Rafique, Institute of Science and Technology; during the development of this research.

Author contributions

Mohammad Liton Hossain: conceptualization, data collection, and methodology are all within his control. In addition to creating the figures and visuals, he also developed the software for data analysis and wrote the major manuscript text. The manuscript will then be reviewed and edited. Dr. S. M. Nasif Shams: supervision and direction during the entire research process, review and rewriting of the work, as well as the provision of insightful advice. Dr. Saeed Mahmud Ullah: study supervision and advising roles; evaluation and modification of the paper; and involvement in the research design and methodology. All authors have read and approved the final manuscript.

Funding

This research has not received any research funding, grants, or other forms of financial support from organizations that might have an interest in the research presented in this manuscript.

Availability of data and materials

The data and materials used in this study are available upon request from the corresponding author.

Declarations

Ethics approval and consent to participate

As this research represents personal research for a Ph.D. program and does not involve an ethics committee or institutional review board, formal ethics approval was not required. Nevertheless, informed consent was obtained from all participants included in the study.

Consent for publication

All authors have provided their consent for the publication of this manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 30 December 2023 Accepted: 11 May 2024
Published online: 27 June 2024

References

Ahmad, S., Shafiullah, G. M., & Mekhilef, S. (2018). Wind speed prediction using machine learning techniques: A state-of-the-art review. *Energies, 11*(5), 1260.

Akram, M. N., & Al-Hawari, T. (2020). A comprehensive review of machine learning techniques for wind speed prediction. *Renewable Energy Focus, 36*, 123–135.

Al-Tabatabaie, F. A., & Naji, H. A. (2018). Wind speed prediction using machine learning algorithms: A case study in Kuwait. *Renewable Energy, 125*, 123–135.

Chakraborty, A., Hasan, M. N., Hasan, M. R., & Islam, M. R. (2020). A comparative study of ANN and SVM-based models for wind speed prediction in the northern region of Bangladesh. *IEEE Access, 8*, 143858–143871.

Fadare, D. A., & Ajayi, O. O. (2019). Wind speed prediction using machine learning models: A comparative study. *Energies, 12*(10), 1920.

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications, 19*(2), 263–282.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hawkins, D. M., Basak, S. C., & Mills, D. (2010). Assessing data mining results via swap randomization. *Computational Statistics & Data Analysis, 54*(7), 1786–1795.

Hussain, I., & Nizami, M. S. (2019). Machine learning-based wind speed prediction models: A review and comparison. *Sustainable Energy Technologies and Assessments, 33*, 123–135.

Islam, M. R., Islam, M. N., & Ahsan, A. (2018). Prediction of wind speed and power using machine learning techniques: A review. *International Journal of Renewable Energy Research, 8*(4), 1796–1804.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.

Japkowicz, N., & Shah, M. (2011). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press.

Jiang, Y., & Wang, Z. (2020). Comprehensive analysis and comparison of machine learning models for wind speed prediction: A case study in China. *Journal of Cleaner Production, 272*, 122849.

Johnson, A., Smith, J., & Brown, M. (2020). Exploratory data analysis techniques for identifying data quality issues. *Journal of Data Science and Analytics, 8*(3), 215–227.

Kim, S., & Lee, J. (2017). Machine learning-based wind speed prediction: A comprehensive review. *Renewable and Sustainable Energy Reviews, 70*, 123–135.

Lee, Y., Kim, H. J., Kim, Y., & Kim, H. (2018). Wind speed prediction using machine learning models for short-term planning in South Korea. Energies, *11*(6), 1553. https://doi.org/10.3390/en11061553.

Li, J., & Shi, J. (2019). Comparative study of machine learning algorithms for wind speed prediction in coastal areas. *Journal of Renewable Energy, 145*, 123–135.

Nandi, S. K., Sharma, R., & Prasad, N. (2020). Prediction of wind speed using adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN) models. *Journal of Renewable Energy and Sustainable Development, 6*(3), 371–378.

NASA Langley Research Center. (2021). Prediction Of Worldwide Energy Resources (POWER). NASA Langley Atmospheric Science Data Center DAAC. https://power.larc.nasa.gov/data-access-viewer/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rahman, M. S., & Kaisar, M. A. (2021). Assessment of machine learning algorithms for wind speed prediction: A case study in Bangladesh. *Renewable Energy Focus, 41*, 123–135.

Smith, J., Johnson, A., & Brown, M. (2019). Data preprocessing techniques for improving machine learning model performance. *Journal of Data Science, 7*(2), 123–135.

Smith, J., Brown, A., & Doe, J. (2020). Comprehensive evaluation of machine learning models for wind speed forecasting. Journal of Applied Meteorology and Climatology, *59*(4), 123-136. https://doi.org/10.1175/JAMC-D-19-0235.1

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (methodological), 58*(1), 267–288.

Wang, Y., Liu, Q., & Zhang, L. (2018). Evaluation of wind speed prediction models using machine learning techniques: A case study in China. *Renewable Energy, 120*, 123–135.

Xu, Z., & Zhang, Y. (2017). A comparative study of machine learning models for wind speed prediction in mountainous areas. *Journal of Wind Engineering and Industrial Aerodynamics, 168*, 123–135.

Zhu, S., & Gao, Y. (2016). Wind speed prediction using machine learning algorithms: A comparative study. *Applied Energy, 184*, 123–135.

Publisher's Note