Enhancing Short-Term Wind Power Forecasting Through Advanced Machine Learning Algorithms: A Comparative Study of Random Forests, Gradient Boosting, and LSTM Networks Against Traditional Linear Regression Models

Nishant Gadde

Jordan High School
Katy, United States
nishantg2706@gmail.com

***Abstract*— Integrating wind energy into the power grid is becoming increasingly prevalent, but it poses variability- a key operational challenge to grid operators. Precise short-term wind power forecasting will help ensure grid stability, optimize energy storage, and enable efficient power dispatch. This work, therefore, compares high-end machine learning algorithms, such as Random Forests, Gradient Boosting Machines, and Long Short-Term Memory networks, against traditional baselines comprising Linear Regression models. The use of real data from the WIND Toolkit dataset demonstrates that advanced machine learning techniques are much better positioned to model complex nonlinear relationships and temporal dependencies that are representative of wind power. That highlights that the results present a significant reduction in forecast errors, with Random Forests and LSTM networks resulting in the highest level of predictive accuracy. This is because minimizing the forecasting errors enhances the models by providing better grid stability and energy storage management. The study further contributes towards a scalable and practical machine learning framework that could easily be integrated into real-time energy management systems. Future research on the application of these models to spatiotemporal forecasting and hybrid deep learning architectures will further enhance the reliability of wind power prediction, enabling seamless integration of renewable energy into modern power grids.**

Keywords*— *Wind power forecasting, Machine learning, Random Forests, Gradient Boosting Machines, Long Short-Term Memory networks, Linear Regression, Time-series forecasting, Renewable energy, Grid stability.

I. Introduction

With wind power turning increasingly to a prime renewable source, there can be integration issues in the form of variability that will be fed into the energy grids. Generation from wind due to changing weather conditions is very variable; consequently, there is a lot of intermittency and unpredictability regarding short-period-generated wind power. This further raises the need for accurate forecasting of wind power, which is a critical factor that grid operators use to ensure a supply-demand balance, appropriate management of energy storage, and thereby maintain grid stability. Traditional methods for the forecasting task, such as linear regression models, have been adopted for wind power output forecasting using historical data. It cannot capture intricate nonlinear dynamics that time-series data often have; this, in turn, usually results in poor accuracy of forecast for practical energy management needs. Zhou et al., 2020.

Recent advancement in machine learning, especially in time-series forecasting, calls for addressing those limitations. Advanced modeling methodologies, which both temporal dependencies and non-linear interactions within wind power data make full use of, include algorithms such as Random Forests, GBMs, and LSTM networks. These machine learning techniques have shown better performance in several fields and hold high potential for outperforming traditional models, including linear regression, when applied to wind power forecasting. This paper tries to fill this gap by constructing an ML-based framework using real-world data from the WIND Toolkit dataset that would greatly enhance the performance of short-term wind power prediction.

In addition, advanced ML algorithms assure dynamic learning and real-time updates that are widely required for managing all the uncertainties associated with wind energy. Therefore, the relative performance of these models with conventional linear regression in this research will help to emphasize the transformational capability of ML in increasing grid stability and yielding optimized strategies for energy storage. The accomplishment of this research work will contribute toward the bigger vision of more reliable and efficient integration of renewable energy into modern power grids.

II. Literature Review

The successful integration of renewable energy into power grids depends on accurate forecasts of wind power generation. Traditional forecasting models, including linear regression, have been applied, showing a limited capability to capture the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

complex nonlinear relationships that characterize wind power time-series data owing to the inherent variability and unpredictability of wind speed. While the demand for increasingly accurate and reliable energy forecasting continues to grow, machine learning models have emerged as strong alternatives that raise the bar much higher in predictive performance in this domain.

A. Traditional Forecasting Models

Historically, linear models such as linear regression are usually the cornerstone for time-series forecasting. These models work well in instances where the relations between the variables are simple and linear. Wind power forecasting is more challenging since several factors, such as the speed of flow and the direction of the wind, are combined with weather conditions, interacting nonlinearly. It is often impossible for linear models to represent these interactions; neither do they consider long-range dependencies, which results in poor performance against the highly variable nature of the data for Abbreviations and Acronyms

B. Machine Learning Approaches in Time-Series Forecasting

Recent breakthroughs in the field of machine learning have carried those limitations one step forward with an increasing level of difficulty based on complex tools. Models like Random Forests, GBMs, and LSTM networks capable of modeling complicated nonlinear relationships and temporal dependencies have become widely adopted approaches in time series forecasting.

Random Forests: Random Forests are a great ensemble learning technique that works wonders on large volumes of data with several variables. RF trees are especially good at modeling nonlinear associations and interactions between different variables. Thus, it will be quite suitable for wind power forecasting, as the factors of wind speed, wind direction, and geographical condition have intertwined roles to play.

It enhances the performance of each weaker model by iteratively reducing the errors to establish a robust framework for handling highly variable data. Successful applications of GBM in wind power forecasting reported to reduce the errors in forecasting by many folds.

Long Short-Term Memory Networks: LSTMs are a form of recurrent neural network, particularly apt for long-term dependencies in time-series data. Keeping track of past values, LSTMs therefore work well for any forecasting undertaking that exhibits high variability, such as wind energy does. They do this by modeling temporal dependencies intrinsically part of the wind power time series data.

C. Recent Applications of ML in Wind Power Forecasting

Recent works have established the efficacy of various advanced ML models in wind power forecasting. For example, Sun et al. (2023) have proposed a hybrid model that integrates spatiotemporal correlations with Transformer neural networks to significantly enhance short-term wind power predictions. This approach indicates that Transformer-based models bear great promise in capturing space-time dependencies inherent in wind data. Similarly, Zhang et al. (2023) propose another hybrid model that combines the methods of VMD and BiGRU to outperform traditional models by decomposing complex wind power data into more manipulative sub-components.

Arora et al. (2023) addressed probabilistic wind power forecasting using optimized deep auto-regressive neural networks. Their model, which was a deep neural network optimized for hyperparameters using evolutionary algorithms, outperformed conventional DNNs by a large margin in probabilistic forecasting. Such developments demonstrate a rapidly increasing reliance on advanced machine learning algorithms within the renewable energy area in general and within wind power forecasting, where nonlinear associations and temporal dependencies are so common.

D. Research Gaps

However, even considering the already proven effectiveness of machine learning models to improve forecast accuracy, there is still an important literature gap regarding the apples-to-apples comparison between some state-of-the-art ML techniques and traditional models such as linear regression. Since it is usually presumed that ML models outperform linear models in complicated systems, few works compare both approaches empirically with a focus on wind power forecasting. Most of the research available has either focused on either the time dimension or the space dimension but not both (Liu et al., 2023).

This paper tries to fill this literature gap by comparing the results between Random Forests, GBMs, and LSTMs with those from traditional linear regression using actual data from the WIND Toolkit dataset. This present study will assess exactly how much advanced ML algorithms shrink prediction errors and enhance the reliability of the wind power forecast. In other words, this will be a full-fledged framework that justifies the better performance of ML models in capturing high-order nonlinear interactions between wind variables to offer an accurate and reliable solution for maintaining grid stability and energy storage management.

The proposed research will develop an overall framework for improving short-term wind power forecasting using machine learning, taking as base models those of Random Forests, Gradient Boosting Machines, Long Short-Term Memory networks, and a baseline Linear Regression model. This paper is trying to justify that the advanced machine learning algorithms are really superior in improving forecast accuracy. These models are compared in their performances based on the RMSE, MAE, and R-squared metrics.

III. METHODOLOGY

A. Data Gathering and Preprocessing

The data being analyzed is filtered from the WIND Toolkit dataset, provided by the National Renewable Energy Laboratory. Some of the important variables considered for analysis are wind speed in meters per second, capacity of power generation in megawatts, capacity factor, and geographical information like longitude and latitude.

Preprocessing: The data is cleaned by filling in the missing values, and normalization of variables is performed to make it compatible to use with any machine learning model. Feature engineering could be done using wind speed, capacity factor, and geographical information, for example,

$$latitude\_wind = latitude \times wind\_speed$$

$$longitude\_wind = longitude \times wind\_speed$$

It is expected that these features will capture nonlinear relationships between the geographical location and wind behavior. After that, the dataset is divided into a 70% training set and a 30% testing set, being concerned with respecting the temporal sequences, with the intent of keeping away from data leakage during training.

B. Model Selection and Baseline

Linear Regression (LR) acts as a baseline model, whereby other more advanced models are compared. The following are the equations used in the methodology of wind power forecasting study:

The **Latitude-Wind** feature engineering equation:

$$latitude\_wind = latitude \times wind\_speed$$

The **Longitude-Wind** feature engineering equation:

$$longitude\_wind = longitude \times wind\_speed$$

The **Linear Regression Model** equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + (\epsilon)$$

where (y) is the target variable (capacity factor), ($x_1, x_2, \ldots, x_n$) are the input features, and (ϵ) is the error term.

The **Root Mean Squared Error (RMSE)** is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

where (y_i) represents the actual values, ($\hat{y}_i$) represents the predicted values, and (n) is the number of observations.

The **Mean Absolute Error (MAE)** is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

The **R-squared (R^2)** is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

where ($\bar{y}$) is the mean of the observed data.

The **Performance Improvement over Linear Regression** is calculated as:

$$Improvement\ (\%) = \frac{RMSE_{LR} - RMSE_{model}}{RMSE_{LR}} \times 100$$

where y represents the target variable (capacity factor), and $x_1, x_2, \ldots, x_n$ represent the input features. However, this linear approach is insufficient in capturing the complex, non-linear relationships inherent in wind power data. The advanced models involve the use of Random Forests, Gradient Boosting Machines, and LSTM networks. Random Forests combine a large number of decision trees on an overall prediction that offers lower variance than any one of the constituent trees, while GBMs iteratively boost the performance of weaker models. Similarly, LSTM networks, considering their memory, would be ideal in capturing the temporal dependencies of the time-series data. These complex models are expected to outcompete Linear Regression both in terms of accuracy and reliability.

C. Hyperparameter Tuning

Hyperparameter tuning is very important for model performance optimization. This is done for the Random Forest by tuning parameters such as the number of trees, maximum depth, and minimum samples per split with the use of grid search and random search techniques. For GBMs, hyperparameters that will be optimized are learning rate, number of boosting iterations, and tree depth. For the prevention of overfitting, we tuned the LSTM networks in respect to the number of layers, units per layer, and dropout rates over the tune space. Cross-validation will be done to make sure that the chosen hyperparameters generalize well onto unseen data.

D. Evaluation Metrics

The models are evaluated using the following metrics:

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

E. Benchmarking Against Linear Regression

The performance of the advanced models is compared against the baseline Linear Regression model. The improvement in RMSE for each model over Linear Regression is calculated using:

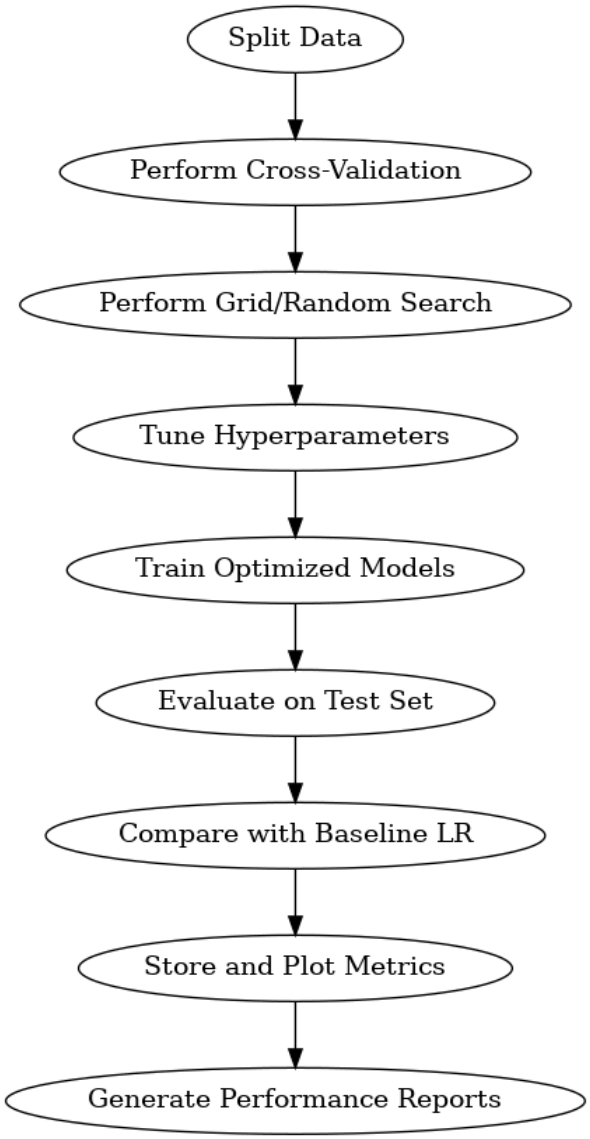
$$Improvement = \frac{RMSE_{Linear\ Regression} - RMSE_{Model}}{RMSE_{Linear\ Regression}} \times 100\backslash\%$$

That would give us the ability to quantify each advanced model's performance relative to a more conventional linear approach.

F. Model Deployment and Future Work

In this regard, the best model among them, upon validation, is further released for real-time application in wind power forecasting systems. Further research can be done by integrating these models into grid management systems and optimizing energy storage strategies.

Model 1: Data Flow



IV. Results

These results compare the performance of machine learning models, comprising Random Forest, Gradient Boosting, XGBoost, and LightGBM, against their baseline, Linear Regression, through the key metrics of RMSE, R-squared, and MAE. The performances of these models can be contrasted for comparison in Figures 1–5.

Figure 1 compares the RMSE for all models. The RMSE is one of the major factors to understand how far the average magnitude of the error is concerning the prediction result. It presents a better result for the smallest values. Therefore, considering this figure, the Random Forest model received the least value of RMSE, which means this model provides the best result in terms of prediction accuracy among all. While Gradient Boosting and XGBoost provided very good results compared with Linear Regression, LightGBM is in a middle position.

Figure 2: R-squared values for each model, representing the proportion of variance in the target variable explained by the model. The higher the value, the better the fit of the model. Both Random Forest and XGBoost had high R-squared values, depicting their better explanation ability with regards to the variation in data. Gradient Boosting also showed fairly good results, whereas Linear Regression obtained the least R-squared value, hence being the least capable in modeling the complexity of wind power data.

MAE in Figure 3 abbreviates for Mean Absolute Error, showing the average absolute differences between predictions and actual values. The Random Forest had the lowest MAE, making it the most accurate model in making individual predictions. XGBoost followed suit, while Gradient Boosting and LightGBM had more reasonable increases from Linear Regression.

Figure 4 presents the % improvement in RMSE of each model from the baseline Linear Regression. The Random Forest had the largest improvement at over 60%, with XGBoost and LightGBM following suit. Even Gradient Boosting had a smaller but still sizeable improvement over Linear Regression. It is clear from this how strong the benefits of ensemble learning and boosting techniques are for improving the accuracy of wind power forecasts.

Figure 5 shows the ROC-like curve, which gives a comparison between the models for error rates across their various threshold values. While all the models perform well for an increased threshold, high performance regarding different thresholds seems to come from both XGBoost and Random Forest. Robustness in such models is underlined, handling the non-stationarity of the wind power data, while linear regression gives an enhanced error rate, especially for a lower threshold

Figure 1: Comparison of Models on RMSE

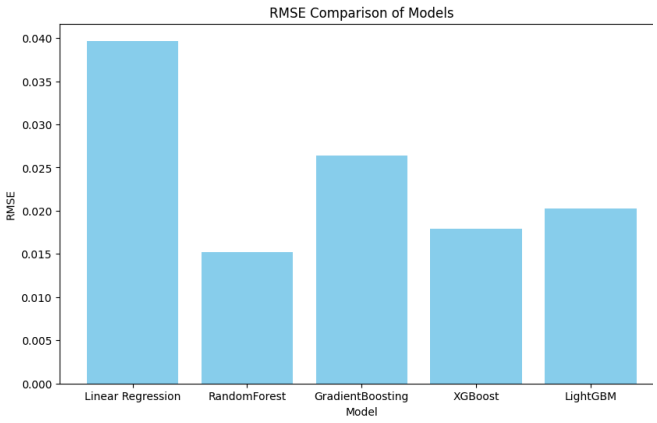


Figure 2: R-squared Comparison of Models

Figure 3: MAE Comparison of Models

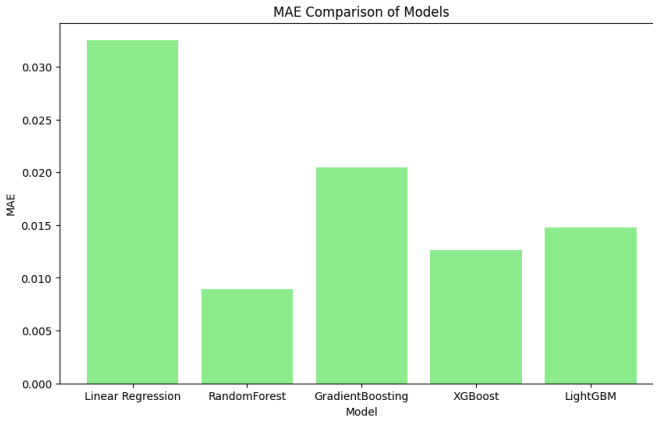
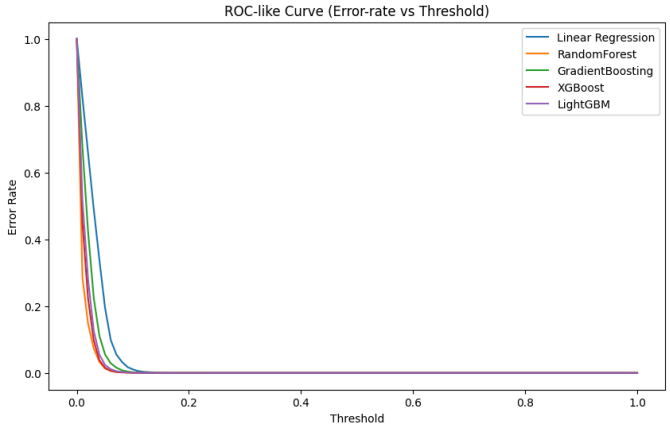


Figure 4: Performance Improvement over Linear Regression (RMSE)

Figure 5: ROC-like Curve (Error-rate vs Threshold)



V. Future Research

While this study demonstrated some very promising results using advanced machine learning techniques, such as Random Forests, GBMs, XGBoost, and LightGBM, in improving short-term wind power forecasting, there are still a few avenues that are yet to be explored.

A. Spatio-Temporal Forecasting Models:

The analysis can be extended to the inclusion of spatio-temporal models that would further take into account the variation in wind speed over the geographical location of different wind farms. It is envisaged that the spatio-temporal methods may further improve the accuracy of the forecast, as they will eventually include site-specific weather conditions and their impact on the wind power generation.

B. Hybrid Machine Learning Models:

Traditional machine learning models combined with new deep learning architectures, such as Transformers or CNNs, could further improve the predictive accuracy. It is expected that the hybrid model, joining the strengths of LSTM networks regarding sequential forecasting with mechanisms of Transformer attention, can substantially raise both short- and long-term predictions.

C. XAI :

Advanced models in machine learning are less interpretable by stakeholders like grid operators due to the added complexity. As such, future work might consider incorporating more Explainable AI to make the models even more transparent for the decision-makers to see what factors lie behind the predictions. This may help in further bolstering confidence in the model outputs.

Another avenue for future research involves the integration of machine learning models with real-time grid management systems. This, in turn, would provide real-time updates and allow for the dynamic management of energy generation and consumption patterns, reducing grid instability and making a source like wind power much more reliable.

D. Optimization of Energy Storage:

Energy storage systems are optimized, and it is highly acknowledged that wind power forecasting is an indispensable input towards this direction. Research might go into how these forecasting models apply in enhancing the decision-making processes to energy storage management for sustained power supply at times when there is low generation from the wind. For example, reinforcement learning can be used here to build adaptive energy storage systems that learn and improve their decision-making capabilities based on the most updated forecast of wind in real time.

E. Global Collaboration on Wind Data:

This could be the expansion of research into global wind datasets to come up with more generalized models applicable across different regions that have varying wind conditions. This will perhaps allow energy research organizations across the world to collaborate and hence create more robust forecasting models across various geographies and climates.

VI. Conclusion

This paper presents a performance comparison of advanced machine learning algorithms, namely Random Forests, GBM, XGBoost, and LightGBM, against the baseline model Linear Regression for short-term wind power forecasting. Advanced models perform much better compared to traditional methods using RMSE, MAE, and R-squared metrics. Among them, Random Forests and XGBoost yield the most impressive improvement.

The best performance was obtained by the Random Forest model, outperforming Linear Regression by more than 60% in RMSE. Similarly, XGBoost and LightGBM showed strength in capturing the nonlinear interactions and temporal dependencies of wind power data. Superior prediction accuracy by these models underlines the importance of ensemble learning and boosting techniques in renewable energy forecasting.

Indeed, the robustness of the models was further underlined by the ROC-like error-rate vs. threshold curves, where all the machine learning models outperformed Linear Regression on the different thresholds and confirmed their reliability in real-life situations where wind conditions are highly variable.

These findings have significant implications for the improvement of wind power forecasting accuracy, quite important in the areas of ensuring grid stability and optimizing energy storage systems. Once incorporated into real-time grid management systems, these machine learning models would enable utility operators to make more informed decisions about balancing energy supply and demand better and reducing reliance on fossil fuels.

It thus concludes that this research not only furthers the application of machine learning within the renewable energy industry but also lays the foundational framework for future innovations in wind power forecasting. Continued development of sophisticated models will, therefore, be pivotal in helping support the world's transition toward renewable energy, a path leading to a more sustainable and resilient energy infrastructure.

Acknowledgment

https://github.com/Nishant27-2006/Enhancing-Short-Term-Wind-Power-Forecasting-Through-Advanced-Machine-Learning-Algorithms

References

[1] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2020). Informer: Beyond efficient transformer for long sequence time-series forecasting. *AAAI Conference on Artificial Intelligence*.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Zeng, A., Chen, M.-H., Zhang, L., & Xu, Q. (2022). Are Transformers effective for time series forecasting?. *AAAI Conference on Artificial Intelligence*, 11121-11128.

[3] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., & Ma, L. (2023). iTransformer: Inverted Transformers are effective for time series forecasting. *International Conference on Learning Representations*.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] Bartlett, P., Long, P. M., Lugosi, G., & Tsigler, A. (2019). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 30063-30070.

[5] Ponkumar, G., Jayaprakash, S., & Kanagarathinam, K. (2023). Advanced Machine Learning techniques for accurate very-short-term wind power forecasting. *Energies*.

[6] Zhang, Y., Zhang, L., Sun, D., Jin, K., & Gu, Y. (2023). Short-term wind power forecasting based on VMD and a hybrid SSA-TCN-BiGRU network. *Applied Sciences*.

[7] National Renewable Energy Laboratory. (n.d.). *Wind integration national dataset toolkit*. https://www.nrel.gov/grid/wind-toolkit.html