Developing cost-effective, practical, accurate non-weather-based wind power production forecasting AI models

Khondoker Farihah Ahmed, Mauricio Hernandez

Received September 28, 2022
Accepted July 13, 2023
Electronic access September 30, 2023

Efficient methods of wind power production forecasting are crucial for the integration of renewable energy into the electrical grid. At present, the application of effective machine learning algorithms for this forecasting relies on three key criteria: accuracy, cost-effectiveness, and practical data collection. Although many state-of-the-art forecasting models in the field are highly accurate, most use weather data inputs such as wind speed and direction. However, cost-effectiveness and practical data collection are not yet widely considered, despite both being crucial obstacles to the actual deployment of such forecasting models. There are numerous issues with this weather data which make it costly and impractical. First, deploying and maintaining a comprehensive network of weather sensors can be challenging and resource-intensive, requiring significant investments, ongoing maintenance, and expertise to manage interpret weather data. Second, professional meteorological agencies and weather data companies have resources expertise to collect data at a higher quality and resolution than a wind power company on its own—and this third-party data tends to be expensive. Third, processing and interpreting raw data to generate useful power forecasts is a complex task requiring significant expertise in data science and meteorology—presenting a higher cost for data processing. This study hypothesizes that an alternate data source–past power production data–can be a more cost-effective and practical data source to predict future power production. 10 minutes of past power production data are used to predict the next 10 minutes of future power production. The open-source 'Wind Turbine Scada Dataset' is used from Kaggle, and are compared with state-of-the-art weather-based models and Kaggle's top-voted submission using the coefficient of determination R^2. All of the models of this study outperformed the state-of-the-art Kaggle submission, with the highest R^2 score of 0.97 with a simple neural network. Furthermore, the models also beat numerous state-of-the-art weather-based models. Thus, by checking all three criteria (high accuracy, cost-effectiveness, practicality of data collection), we demonstrate a novel, highly efficient alternative for wind power production forecasting which can be deployed in wind farms quite simply–accelerating the integration of renewable energy sources into the grid.

Introduction

Being able to forecast power generation for any energy system is crucial to its smooth operation, and for the integration of renewable forms of energy such as wind energy into the electrical grid, efficient forecasting methods are of paramount importance[1]. Wind power forecasting is indeed necessary for several reasons. First, wind power is intermittent and varies greatly due to changes in wind speed and direction. This variability can introduce instability into the power grid if not effectively managed. Accurate forecasting of wind power can help system operators schedule power resources more effectively and maintain grid stability[2]. Second, accurate forecasting can help wind farm operators optimize the operation of turbines and reduce maintenance costs. Knowing when wind power output is expected to be high can aid in scheduling maintenance during periods of low output, helping operational efficiency[3]. Third, for energy markets, accurate forecasts of wind power production can improve the reliability of power supply, reduce the costs associated with balancing supply and demand, and in turn, promote the integration of wind energy into the power system. Hence, these forecasts are crucial for market integration[4]. Fourth, accurate forecasting can also aid in investment planning and policy-making related to wind power[5].

In addition, as the contribution of wind energy to the power grid increases, so does the complexity of managing the grid due to inherent variability and uncertainty of wind. Uncertainties can lead to frequency instability, unexpected power flows, and voltage problems, posing significant risks to the secure operation of power systems[6]. Wind power forecasting tools can help alleviate these issues relating to grid stability in several ways. Accurate wind power forecasts can reduce the amount of operating reserve that system operators must maintain to account for wind power uncertainty. This reduction of operating reserves can lead to cost savings and increased overall efficiency[7]. With accurate wind power forecasts, system operators can also make better decisions about

dispatching generation units, reducing operational costs, and increasing overall system efficiency[8]. Third, in regions with organized electricity markets, accurate wind power forecasts can help market participants make more informed bidding decisions, improving market efficiency and facilitating market operations. Hence, not only is wind power forecasting essential for maintaining grid stability in the face of increasing wind power penetration levels, but it can also provide economic efficiency, improving reliability and stability of the grid operation[9].

As machine learning (ML) and artificial intelligence (AI) continue to be applied to this field, from a study by Saroha et. al, we identify three key evaluation criteria to judge the practical implementation of such forecasting models in current energy systems: the accuracy of the forecast, the cost-effectiveness of the forecast, and the practical collection of the type of data required. All of these factors are equally important for the practical deployment of such models[1].

Currently, state-of-the-art AI models for wind power forecasting are mostly reliant on input data such as wind speed, wind direction, and sometimes other weather/meteorological data[10]. However, there are numerous issues with obtaining these kinds of weather data from external companies. For example, a study by Ordiano et al. (2016) focusing on solar energy highlights the following issues with obtaining weather data on a daily basis to run these models: first, purchasing this data from weather data companies creates high, additional costs for the maintenance of such models, decreasing profitability; second, constant communication with weather data services are needed to run such models, however, in the case of communication failures, these weather-based AI forecasting models come to a halt and are unable to continue operating[11]. Although the study by Ordiano et al. (2016) focuses on solar power output prediction rather than wind power output prediction, there are several reasons why it could still be relevant to the broader context of grid stability and grid operations. First, there are many shared challenges between the domains: both wind and solar power output predictions face similar challenges such as high variability and dependence on weather conditions. Hence, solutions and methodologies used in solar power output prediction could potentially be applicable or provide insights for wind power output prediction[12]. Second, grid stability is also key for both domains: both wind and solar are integral parts of renewable energy contributions to the power grid. Studies in solar power prediction can offer useful insights into handling intermittency and variability, which are crucial for maintaining grid stability[13]. Third, both domains are part of integrated, hybrid systems: in many regions, wind and solar are used in tandem. Insights from solar power output prediction could therefore be relevant when considering the overall balance of power in systems that use a mix of renewable energy sources[14]. Fourth, there is methodological relevance across both domains: machine learning techniques used for solar power output prediction may be applicable to wind forecasting as well. In this context, it is beneficial to reference techniques or results from solar power studies[15]. Fifth, the combination of different renewable sources forecasts may enhance the performance of power system operation models by diversifying the input data and thus reducing the overall forecast error[16]. Sixth, although the aforementioned study focuses on solar power production forecasting, the same type of data (weather data) is also used for wind power production forecasting. Thus, these same issues ultimately apply to wind power production forecasting as well[11].

Although wind farms can use weather sensors such as anemometers (a measuring tool which measures wind speed, direction, and pressure), there are still many critical issues with obtaining and using weather data from sensors. First, there are data collection challenges: deploying and maintaining a comprehensive network of weather sensors can be challenging and resource-intensive. It requires significant investments in hardware and infrastructure, ongoing maintenance to ensure the sensors continue to function correctly, and the expertise to manage and interpret the data they generate[17]. Second, there are issues with the quality and resolution of data: professional meteorological agencies and weather companies typically have the resources and expertise to collect weather data at a higher quality and resolution than a wind power company could achieve on its own. While sensors can capture local data, they may miss broader meteorological phenomena that could impact wind patterns[18]. Third, expertise and resources are another crucial challenge: processing and interpreting raw weather data to generate forecasts is a complex task that requires significant expertise in meteorology and data science. Many wind power companies lack these capabilities in-house and could find it more cost-effective to use third-party weather data that has already been processed and contextualized: however, this itself–buying third-party weather data–presents an additional cost burden[19].

The systematic review study by Mosavi et al. (2019) reviews numerous state-of-the-art models and studies in the wind energy field[20] , related to wind speed forecasting[21–24] wind power forecasting[25,26], and scheduling strategies for fluctuations in wind power.[27,28] The coefficient of determination R^2 range of nearly all state-of-the-art weather-based models is approximately between 0.875 to almost 1; hence we define this as the threshold R^2 accuracy for being comparable to a 'state-of-the-art' weather-based model[29].

While current state-of-the-art wind power production forecasting models in the field may fit only one of the three required criteria (high accuracy), the cost effectiveness and practical collection of the input data are still essential issues which are yet to be resolved in the field[1]. However, an alternative source of input data, which has the potential to fit all three

of these criteria, is the past power production data of a wind farm. The past power production data is a wind farm's own, readily-available, free data–solving the cost effectiveness and practical collection of data issue[11]. In regards to the accuracy of this type of input data in models, it is important to note that the most recent past power production data of a wind farm will likely be reflective of the most recent weather trends as well (recent past power production and most recent weather trends have a tendency to be highly correlated data)[5]. This paper aims to show that expensive, difficult-to-obtain pieces of weather data such as forecasted future wind speed and future wind direction do not need to be obtained to predict future wind power production, if two simple inputs–time and past wind power production–can be used to predict future wind power production. Furthermore, it is important to note that even though forecasted wind speed and direction may sometimes be provided for free by weather services, the difficulty is that such forecasts are usually for larger areas (such as an entire city), so relying on this free data about wind speed and direction would not work as it would likely not give an accurate measure of the forecasted wind speed and direction in the wind farm's exact, specific location. This is because weather data is often collected from observation stations located away from city centers, which may not accurately represent the weather conditions of a specific location within the city; in addition, small-scale spatial variations in weather conditions may not be resolved in weather models, leading to inaccuracies in predicting weather for a specific location[30].

However, the wind power production data of the past 10 minutes can likely be used to predict the power production relatively accurately. This is because it is highly unlikely that weather conditions would change drastically within a span of these 10 minutes; while it can occur, the likelihood that a given wind generator will experience less than a 10% change in wind output in one hour is 80%. However, over a span of 5 hours, the odds increase to 40% that the wind output will change by 10% or more: indicating that longer-term (over a few hours) variability is more common that minute-by-minute variability[31]. Also, again, the power production data is highly correlated with the weather trends and weather data itself, which is why there is a high likelihood of obtaining an accurate forecast of future power production using just 10 minutes of past power production data. Moreover, many wind farms would likely already have the data of power production for the past 10 minutes, so this data would be readily available and free, thus eliminating the need to go to weather data services. This would check the other two difficult criteria of cost-effectiveness and practical collection of data. However, an argument can be made that wind speeds and directions can drastically change within the span of 10 minutes, causing the future prediction from this data to be inaccurate. It is still worth noting that such volatile changes would only be a portion of the predictions made in an entire day; we don't expect high levels of volatility within a 10 minute interval for the majority of the day, rendering such a prediction algorithm to be still potentially powerful and highly accurate in predicting future power predictions for the rest of the day. However, we recognize that this may still be a drawback of this approach, and the limitation of this approach's performance in high-volatility weather locations of the world is later discussed in the Results and Discussion section.

The goal of this study is to examine ways to use alternative sources of data to check all three aforementioned criteria1. Developing AI non-weather-based models which are just as accurate as state-of-the-art weather-based ones by using past power production data (an alternative source of data) would be a giant step towards making wind power forecasting systems highly efficient, and ultimately it would check all three criteria1. We hypothesized that if we use the free and readily-available past power production data of a wind farm to predict future power production, then the accuracy of this ML model will be comparable to current state-of-the-art weather-based models.

Materials and Methods

Data

The data that will be used in this study is the open-source 'Wind Turbine Scada Dataset' from Kaggle. Scada systems measure data including wind speed, wind direction, and power generation in 10 minute intervals. This particular dataset is from a turbine in Turkey[32].

Evaluation Metrics

We will compare our models in two ways. First, we will compare the results of our non-weather-based models (which only use past power production data) with weather-based models (which use weather data). As both these models will be from the same dataset, evaluation metrics with the same units can be used to judge which approach has lower errors. We will use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to compare weather-based models with non-weather-based models from the same dataset, which will have the same units. Both RMSE and MAE are widely-used metrics for assessing the accuracy of predictive models. RMSE gives a higher weight to larger errors, while MAE provides a direct average measure of the absolute differences between predicted and observed values, thereby treating all errors equally irrespective of their magnitude.

Second, we will compare our non-weather-based models with state-of-the-art weather-based models from Mosavi et. al (2019), using the coefficient of determination R2 as it has no

units (thus cross-study comparisons can be made). In addition, R2 is used as it is one of the most commonly used metrics in the domain of wind energy forecasting that does not have units—enabling easy cross-study comparisons. In this review study, the R2 range of nearly all state-of-the-art weather-based models is approximately between 0.875 to almost 1; hence we define this as the threshold R2 accuracy for being comparable to a 'state-of-the-art' weather-based model[29].

The following equation describes the R^2 metric:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum\left((\hat{y}_i - \bar{y})^2\right)}{\sum(y_i - \bar{y})^2} \quad (1)$$

where SSR is the sum of squared regression, SST is the sum of squares total (or total sum of squares), y^i is the predicted value or a point on the regression line, y is the mean of all values, and y^i represents the actual values or points.

The following equation describes the MAE metric:

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

where n is the total number of data points, y_i is the true value, and $\hat{y}_i$ is the predicted value.

And finally, the following equation describes the RMSE metric:

$$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (3)$$

where the variables are the same as explained above.

Data Preparation

We used the Wind Turbine Scada Dataset downloaded from Kaggle. It has 5 attributes:

1. Date/Time in 10-minute intervals,
2. LV active power (kW) or the amount of power generated by the turbine in those 10 minutes,
3. Wind speed (m/s),
4. The theoretical power curve (KWh) or the theoretical power values that the turbine generates with that wind speed, which is given by the turbine manufacturer, and
5. Wind direction (°)[5]

For our non-weather-based models, we changed the format of the data. We only kept the LV active power (kW) column and eliminated the rest, as power production data is both in our input and output, and the goal of the study is to only use this type of simple data and not any weather or other complex forms of data.

We then restructured the data in the following way. First, we made two copies of the dataset. We named the first one the input dataset, and the second one the output dataset. Second, we deleted the last row of the input dataset, and the first row of the output dataset. This is because if we align their indexes now side by side, the input dataset has essentially 10-minute intervals of current active power production, and it predicts in the output dataset the next 10 minutes of active power production. So, the past 10 minutes of power production predicts the next 10 minutes, without the assistance of any meteorological or time inputs such as time of day, wind speed, wind direction, etc. This eliminates the need to purchase forecasted wind speed and forecasted wind direction data from weather data companies, thus making it a cost-effective solution and also a practical one for data collection, since past power production data is readily available to all wind farms themselves.

On the other hand, we also tried weather-based models, to compare with our non-weather-based model approach and see if there was any drop in accuracy metrics despite not using weather data. For this, we made two datasets for the input and output. The input dataset had the wind speed (m/s) and wind direction (°) data for each 10-minute interval, and the output dataset had the corresponding LV active power (kW). Hence, the weather data (wind speed and direction) were used to predict the power output for each time interval.

Machine Learning Algorithms

Before being fed into the ML models, the data was split into a training and test set with an 80:20 ratio, respectively. The data was not shuffled, as the time order of the data is crucial for this study.

Five simple, traditional models with default configurations and seven simple neural networks are tested. The five simple, traditional models are:

1. Linear Regression (LR)
2. K-Nearest Neighbors (KNN)
3. Decision Tree Regressor (DT)
4. Multi-Layer Perceptron Regressor (MLP)
5. Random Forest (RF)

For the seven neural networks tested, the architecture of each one is described in Table 2. The ReLU (Rectified Linear Unit) activation function is used for all the neural networks so that the models can learn non-linear relationships in a fast and efficient way. All the models were compiled with a mean squared error (MSE) loss function.

Table 1 - Neural Networks (NN) Model Architecture: 7 neural networks (NNs) were designed, each with 3 dense layers. All layers had one input (past power production of 10

Table 1 Description of Neural Network Models

Model	Layers	Number of units each layer	Number of epochs	Activation function	Loss function
NN1	3 dense layers	1 (input) - 500 - 8 - 1 (output)	10	ReLU for all layers	MSE
NN2	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	10	ReLU for all layers	MSE
NN3	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	15	ReLU for all layers	MSE
NN4	2 dense layers	1 (input) - 100 - 1 (output)	10	ReLU for all layers	MSE
NN5	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	20	ReLU for all layers	MSE
NN6	3 dense layers	1 (input) - 100 - 1000 - 1 (output)	30	ReLU for all layers	MSE
NN7	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	30	ReLU for all layers	MSE

minutes) and one output (power production of the next 10 minutes). The number of units between the input and output layer varied. Models were tested between the range of 10 to 30 epochs. The activation function used was ReLU for all layers, and the loss function used to compile all models was mean squared error (MSE).

Results

Using the 'Wind Turbine Scada Dataset' from Kaggle, we only used the past power production data from this dataset, and compare our results to a weather-based model using the same dataset, and other weather-based state-of-the-art models in the field. Comparing our weather-based models to the non-weather-based models, we evaluate using RMSE and MAE. We also evaluate whether or not our models have crossed the 'state-of-the-art model threshold' for R^2 defined in the Introduction section, from 0.875 to almost 1.

Table 2 shows the results of the evaluation metrics, MAE, RMSE, and R^2, for our weather-based and non-weather-based models. Note that R^2 is from a scale of 0 to 1, where 1 indicates that the model perfectly fits the data, and 0 indicates that the model is not better than simply using the mean to predict the value. Our goal is to try to get as close to 1 as possible, without having potential overfitting of the model. The R^2 values were reported by comparing the predictions of the models on the test set (not the training set) to the actual values. In addition, RMSE and MAE are in the units of the dataset, and our goal is for these two error metrics to be as low as possible. We will evaluate whether RMSE and MAE are similar or lower in the non-weather-based approach compared to the weather-based approach.

Table 2 – Evaluation Results for All Models: A total of 12 models are tested for each approach (weather-based and non-weather-based), with 5 traditional approaches (LR, KNN, DT, MLP, and RF) and 7 designed NNs (NN1 to NN7). All values are rounded to four significant figures.

For the non-weather-based approach, the best models were between NN6, NN5, and NN4, with NN6 achieving the best scores in two metrics; on the other hand, the worst model was RF consistently.

For the weather-based approach, the best model was consistently NN7, and the worst model was consistently NN6 (in fact, NN6 even achieved a negative R2 score, indicating that it predicted worse than the mean of the power output values we tried to predict; hence NN6 was likely not a well-suited model for the weather-based approach).

To compare the weather-based approach with the non-weather-based approach: the average RMSE and MAE for the weather-based approach were higher than the average RMSE and MAE for the non-weather-based approach. In addition, the average R2 was much higher for the non-weather-based approach than the weather-based approach. Furthermore, all of the non-weather-based models' R^2 score fits within the threshold R^2 score we identified from Mosavi et. al between 0.875 to nearly 1: as all our models performed above 0.95, and the average R^2 score was 0.9691 for the weather-based approach.

Hence, the results indicate that the non-weather-based models (using only past power production to predict future power production) outperform the weather-based models (using wind speed and wind direction to predict power production). Furthermore, every single model was in the threshold for a state-of-the-art model, and all of them were on the higher end of the boundary (above 0.95 for R2 score), indicating comparable results to state-of-the-art models, and that many of our models are actually even better than some of the state-of-the-art weather-based wind models reviewed by Mosavi et al. (2019)[19].

Figure: 1 shows an example of one of our models (NN1) and its prediction with our method. It shows the predicted vs. actual power generation for the first 24 hours of data in the dataset. The graph shows that the model was able to clearly grasp the pattern of the data without overfitting.

Table 2 Evaluation Results for All Models

Model	Non-weather-based approach			Weather-based approach		
	RMSE	**MAE**	**R2**	**RMSE**	**MAE**	**R2**
LR	227.6	135.8	0.9712	552.7	413.1	0.8304
KNN	244.1	143.2	0.9669	560.3	270.5	0.8257
DT	243.6	138.4	0.9671	625.1	272.7	0.783
MLP	227.8	134.7	0.9712	601.6	470.9	0.799
RF	285.9	166.3	0.9546	562.6	268.6	0.8243
NN1	227.6	130.0	0.9712	523.0	344.7	0.8481
NN2	228.6	130.5	0.9710	537.4	350.2	0.8396
NN3	228.5	130.1	0.9710	459.5	295.2	0.8828
NN4	228.1	129.0	0.9711	549.0	404.9	0.8326
NN5	227.5	131.6	0.9713	475.8	294.1	0.8743
NN6	227.4	132.3	0.9713	1987.0	1466.0	-1.193
NN7	227.6	130.3	0.9712	401.7	210.5	0.9104
Average Score	235.4	136.02	0.9691	653.0	421.8	0.6714
Best Model	227.4 (NN6)	129.0 (NN4)	0.9713 (NN5 and NN6)	401.7 (NN7)	210.5 (NN7)	0.9104 (NN7)
Worst Model	285.9 (RF)	166.3 (RF)	0.9546 (RF)	1987.0 (NN6)	1466.0 (NN6)	-1.193 (NN6)

Figure 1. Predicted power output vs. actual power output by Neural Network 1 (NN1) for the first 24 hours of data in the dataset.

The performance of one of the models, NN1, has been shown below. The model's predictions for power output (in blue) for the first 24 hours of data in the dataset (from January 1, 2018, 12:10 AM to January 2, 2018, 12:10 AM are shown). 10 minutes of the past power output data is used to predict the next 10 minutes of power output. They are compared with the actual power output values (in red). 144 values are used to create the graph below, as there are 144 10-minute-intervals within 24 hours. The results indicate that the model was able to capture the trends of the data without over-fitting. Figure 2 shows a zoomed-in version of figure 1, of the critical time between 15 and 20 hours. The results indicate that at times the model over-predicted, and at times the model under-predicted; however, in general, the model was able to grasp the trend of the power production data without over-fitting.

Figure 2. Predicted power output vs. actual power output by NN1 between 15 and 20 hours.

This graph is an extension of figure 1, focusing in on the time between 15 and 20 hours. The results show that while the model was able to grasp the general trend of the power output (the shapes of the predicted and real power output graphs are similar), the model over-predicted the power output at times and under-predicted the power output at times. Hence, this is also a sign that the model did not over-fit, and MSE and RMSE scores explains the gap between the two graphs.

Furthermore, Figure 3 shows a comparison of the prediction between the best weather-based model (NN7 for all metrics) and the best non-weather-based model (we used NN6 since it won the bet in two metrics).

Figure 3. Best weather-based model's predictions (NN7) vs best non-weather-based model's predictions (NN6) vs actual power output.

A random day is shown between the hours 15 and 20 (critical hours of power production). This particular sample shows that the non-weather-based prediction and the weather-based prediction were both close to the actual power production. Between hour 15 and 17, the weather-based model appears to be much closer to the actual power production graph. However, between hours 17 and 20, it shows that the weather model vastly underpredicts results, and the non-weather-based model is much closer to the actual power output. Hence, while both models are close to the acutal power output, the non-weather-based model was closer for more time samples (explaining its lower error scores).

Furthermore, additional statistics were calculated from the dataset's wind speed in order to measure how volatile weather conditions truly were in our data, and hence how robust our model is against variable, volatile wind speeds. The mean wind speed was found to be 7.56 m/s, with a range between a minimum of 0.00 m/s and a maximum of 25.2 m/s. The variance of the wind speeds was found to be 17.87 m2/s2. Relative to the mean, minimum, and maximum, this indicates a relatively high variance or 'volatility' in our wind speed data; however other regions could be more or less 'volatile' in weather conditions and wind speeds compared to this. Table

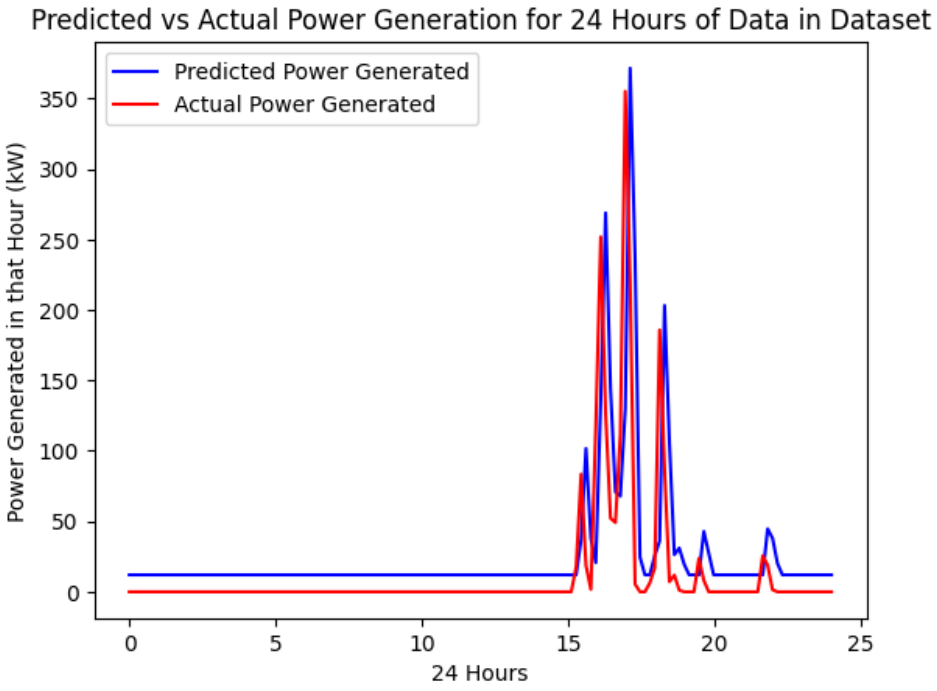

Fig. 1 Predicted power output vs. actual power output by Neural Network 1 (NN1) for the first 24 hours of data in the dataset.

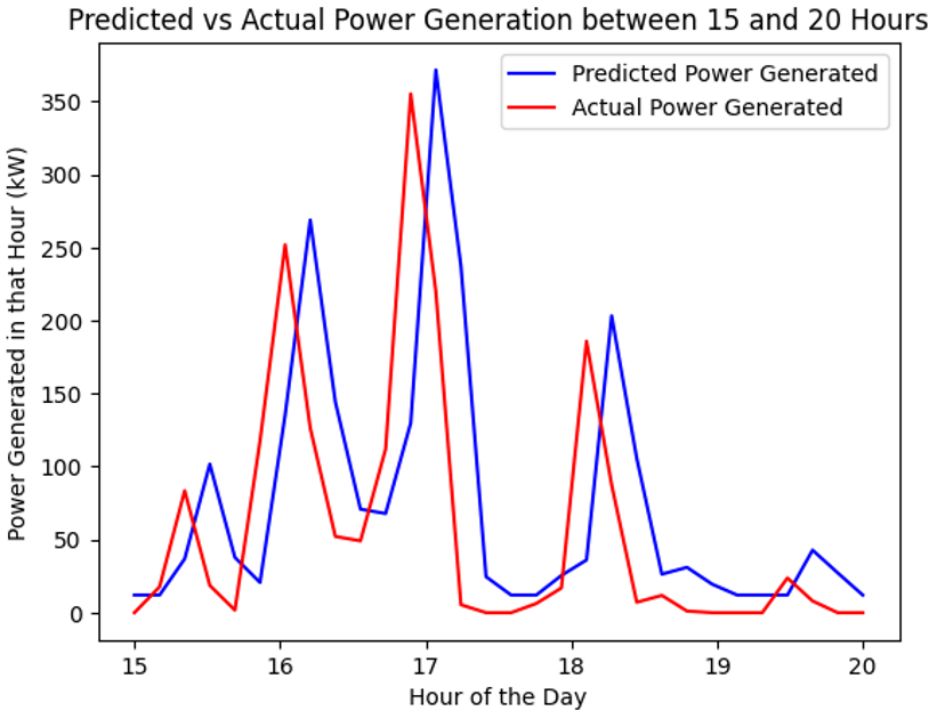

Fig. 2 Predicted power output vs. actual power output by NN1 between 15 and 20 hours.

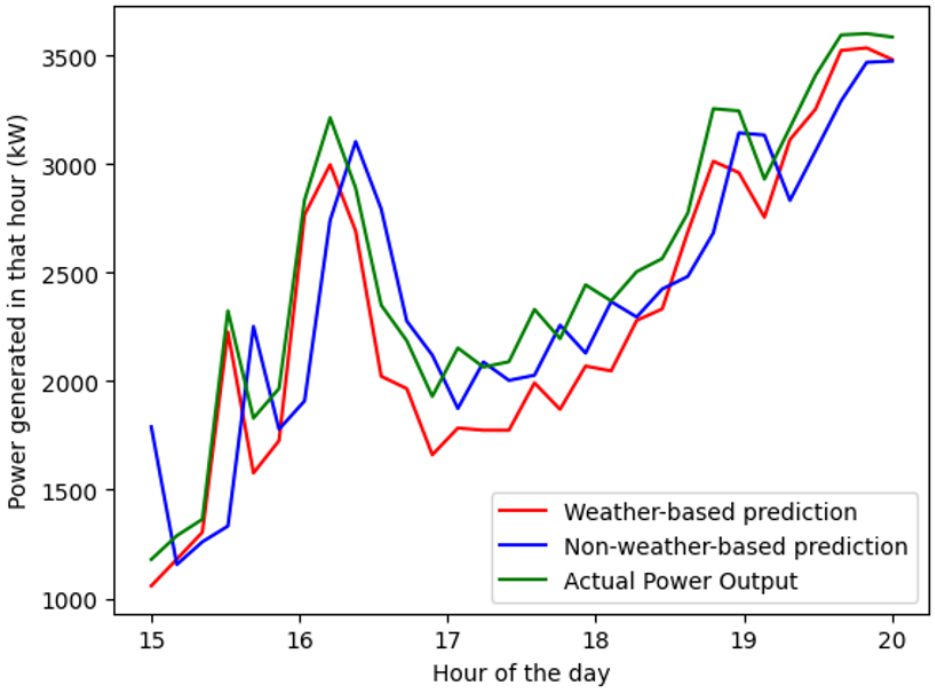

Fig. 3 Predicted power output vs. actual power output by NN1 between 15 and 20 hours.

2 shows more detailed information about the statistical details given for the dataset used in this study.

Table 3 – Statistical Details for Dataset: The dataset has values in ten minute intervals from January 1st, 2018 at 00:00 to 31st December, 2018 at 23:50. We used three types of data in both types of models: LV ActivePower (kW), wind speed (m/s), and wind direction (°). The mean, median, mode, kurtosis, standard deviation, maximum, minimum, and range are provided below for each of these three data inputs.

Discussion

Our hypothesis (that if we use the cheaper data source of a wind farm's past power production data, then the accuracy will be comparable to current state-of-the-art weather-based models) was observed to be accurate. All of our models outperformed the weather-based models from the same Kaggle dataset using wind data, and our models also reached the higher boundary of the state-of-the-art models' R2 threshold from the Mosavi et. al (2019) review study.[29]

However, there are some limitations of this study and corresponding opportunities for future studies to explore. First, the primary limitation of this approach is weather volatility. Although our wind speed data's variance compared to its minimum, maximum, and mean wind speeds indicated a relatively high variance, the next step for future studies is to perform this approach on areas known to have highly volatile wind speeds and weather conditions. Our models' high accuracies indicates that they can perform well in settings with high variance or 'volatility' in wind speeds, but testing should be done in areas with even more volatility and high variance in order to prove model robustness. However, even still, this model can be applied to regions with extremely volatile weather conditions where wind farms have energy storage solutions (ESS) or battery capacities; the ESS would be able to account for any prediction inaccuracies of the model. Furthermore, in areas with high wind speed volatility, the model's data input can be changed to account for minute-by-minute predictions; so instead of using the past 10 minutes to predict the next 10 minutes, the past minute could be used to predict the next minute in order to account for higher weather volatility. This would not require any change in the model itself, only a small change in how the data is collected to make it more granular. Second, another opportunity for future studies is to perform cross-validation to further prove model robustness. Third, it is recommended that other machine learning techniques, such as data preprocessing, feature engineering, and ensemble methods, are used by future studies to improve the results further. Additional pieces of weather data which are free to use can also be added to this model through these techniques to further improve model capabilities and robustness, especially in settings with high wind speed volatility. Fourth, it is rec-

Table 3 Statistical Details for Dataset

	Mean	Median	Mode	Kurtosis	Std. Deviation	Maximum	Minimum
LV ActivePower (kW)	1307.68	825.84	0	-1.16	1312.46	3618.73	-2.47
Wind Speed (m/s)	7.56	7.10	0	0.06	4.23	25.21	0
Wind Direction (°)	123.69	73.71	0	-0.75	93.44	360	0

ommended that other metrics aside from R2 are also tested and measured in future studies to further assess and improve model performance with other metrics. Fifth, an important research area to be investigated further is the application of this technique to other forms of renewable energy forecasting with volatile power production, especially wave and tidal power to further advance their greater integration into the electrical grid.

Overall, this study shows the vast potential of using past power production data as the sole input to predict future power output checks all three criteria: our method was demonstrated to be highly accurate (even more accurate than the traditional weather-based models), cost-effective (as this data is open-source to wind farms), and practical for data collection (as this data is readily available to wind farms). By being comparable to state-of-the-art weather-based models in accuracy especially (the most important criteria for deployment) this study's method helps to take a giant step towards making wind power forecasting systems highly efficient, as it checks all three criteria[1]. The implications of this study are that this forecasting technique can be applied realistically and easily to nearly any wind farm as it is cost-effective and practical, so this study helps to further accelerate the integration of renewable energies to the electrical grid and shift away from fossil fuel usage.

$$\frac{\hat{a}}{\frac{\sum(x_i)^{\hat{b}}}{\sum x_i^{\hat{c}}}} = x_i^{\hat{d}} \tag{4}$$

Acknowledgement

Thank you to Kayla Saucedo from Harvard University for being an awesome mentor during the Harvard Student Agencies Impact Research Fellowship!

References

1 S. Saroha, S. Aggarwal and P. Rana, *Forecasting in Mathematics - Recent Advances, New Perspectives and Applications. IntechOpen*.
2 A. Costa, A. Crespo, J. Navarro, G. Lizcano, H. Madsen and E. Feitosa, *Renewable and Sustainable Energy Reviews*, **12**, 1725–1744.
3 B.-M. Hodge, E. Ela and M. Milligan, *Wind Engineering*, **36**, 509–524.
4 D. Swider and C. Weber, *European Transactions on Electrical Power*, **17**, 151–172.
5 F. Santos-Alamillos, D. Pozo-Vázquez, J. Ruiz-Arias, V. Lara-Fanego and J. Tovar-Pescador, *Renewable Energy*, **69**, 147–156.
6 J. Smith, E. Demeo, B. Oakleaf, K. Wolf, M. Schuerger, R. Zavadil, M. Ahlstrom and W. Nakafuji, *Grid Impacts of Wind Power Variability: Recent Assessments from a Variety of Utilities in the United States*, https://www.nrel.gov/docs/fy06osti/39955.pdf.
7 L. Soder, L. Hofmann, A. Orths, H. Holttinen, Y. Wan and A. Tuohy, *IEEE Transactions on Energy Conversion*, **22**, 4–12.
8 J. Usaola, *Electric Power Systems Research*, **80**, 528–536.
9 H. Holttinen, P. Meibom, A. Orths, B. Lange, M. O'Malley, J. Tande, A. Estanqueiro, E. Gomez, L. Söder, G. Strbac, J. Smith and F. Hulle, *Wind Energy*, **14**, 179–192.
10 E. Music, A. Halilovic, A. Jusufovic and J. Kevric, The International Symposium on Computer Science - ISCS.
11 J. Ordiano, S. Waczowicz, M. Reischl, R. Mikut and V. Hagenmeyer, *Computer Science - Research and Development*, **32**, 237–246.
12 H. Ye, B. Yang, Y. Han and N. Chen, *Frontiers in Energy Research*, **10**, year.
13 J. Kleissl, *Solar Energy Forecasting Advances and Impacts on Grid Integration solar resource and forecasting laboratory*, https://www.energy.gov/sites/prod/files/2016/08/f33/1.Jan_Kleissl-PVSCPlenary.pdf, Retrieved June 11, 2023, from.
14 I. Mitra, D. Heinemann, A. Ramanan, M. Kaur, S. Sharma, S. K. Tripathy and A. Roy, *International Journal of Energy and Environmental Engineering*, **13**, 515–540.
15 Y. Ren, P. Suganthan and N. Srikanth, *Renewable and Sustainable Energy Reviews*, **50**, 82–91.
16 N. Benti, M. Chaka and A. Semie, *Sustainability*, **15**, 7087.
17 A. Foley, P. Leahy, A. Marvuglia and E. McKeogh, *Renewable Energy*, **37**, 1–8.
18 J. Fuchsberger, G. Kirchengast and T. Kabas, *Earth System Science Data*, **13**, 1307–1334.
19 W. M. Organization, *Valuing Weather and Climate: Economic Assessment of Meteorological and Hydrological Services*, https://library.wmo.int/doc_num.php?explnum_id=3314.
20 P. Chatziagorakis, C. Ziogou, C. Elmasides, G. Sirakoulis, I. Karafyllidis, I. Andreadis, N. Georgoulas, D. Giaouris, A. Papadopoulos, D. Ipsakis, S. Papadopoulou, P. Seferlis, F. Stergiopoulos and S. Voutetakis, *Neural Computing and Applications*, **27**, 1093–1118.
21 Q. He, J. Wang and H. Lu, *Applied Energy*, **226**, 756–771.
22 Z. Qu, K. Zhang, J. Wang, W. Zhang and W. Leng, *Advances in Meteorology*, 3768242.
23 A. Khosravi, L. Machado and R. Nunes, *Applied Energy*, **224**, 550–566.
24 M. Mana, M. Burlando and C. Meissner, *International Journal of Renewable Energy Research (IJRER*, **7**, 1629–1638.
25 A. Sharifian, M. Ghadi, S. Ghavidel, L. Li and J. Zhang, *Renewable Energy*, **120**, 220–230.
26 M. Anwar, M. El Moursi and W. Xiao, *IEEE Transactions on Power Systems*, **32**, 1–1.
27 J. Sarshar, S. Moosapour and M. Joorabian, *Energy*, **139**, 680–693.
28 L. Cornejo-Bueno, L. Cuadra, S. Jiménez-Fernández, J. Acevedo-Rodríguez, L. Prieto and S. Salcedo-Sanz, *Energies*, **10**, 1784.
29 A. Mosavi, M. Salimi, S. Faizollahzadeh Ardabili, T. Rabczuk, S. Shamshirband and A. Varkonyi-Koczy, *a Systematic Review. Energies*,

12, 1301.
30 C. Bianchi and A. Smith, *SoftwareX*, **10**, 100299.
31 *National Renewable Energy Laboratory*, https://web.archive.org/web/20120607000124/http://www.nrel.gov/wind/systemsintegration/system_integration_basics.html.
32 B. Erisen, *Wind turbine scada dataset. Kaggle; Kaggle*, https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset.