Proceedings of the Canadian Society for Mechanical Engineering International Congress
32nd Annual Conference of the Computational Fluid Dynamics Society of Canada
Canadian Society of Rheology Symposium
CSME-CFDSC-CSR 2025
May 25–28, 2025, Montréal, Québec, Canada

Data-Driven Feature Selection for Prediction of Wind Turbine Vibrations

Mohsen Masoomi Ardakani[1], Jianming Yang[2*]
[1,2]Mechanical and Mechatronics Engineering, Memorial University, St john's, Canada
*jyang@mun.ca

Abstract—The growing demand for renewable energy necessitates enhancing wind turbine performance and reliability. Wind turbines' operational efficiency and lifespan depend heavily on two critical vibration parameters: tower and drivetrain vibration. Accurate data-driven models for predicting wind turbine vibrations require optimal feature selection as a fundamental step. A supervisory control and data acquisition (SCADA) system dataset containing 301 features and about 30k samples provide the foundation for this study to systematically evaluate and analyze critical factors that impact tower and drivetrain vibrations. Advanced feature selection methods that combine correlation analysis with mutual information and feature importance from Random Forest and XGBoost models determine the impact of different parameters on turbine vibrations. The number of features in the dataset underwent preprocessing to maintain high-quality input by removing unnecessary and redundant data. Mutual information revealed hidden non-linear relationship patterns between features, and feature importance methods confirmed the crucial role of these parameters. The analysis shows that wind speed is the main contributing factor to tower acceleration measurements, and rotor speed is an essential variable for drivetrain vibrations. Research findings create knowledge that enables the development of predictive maintenance models and improved wind turbine reliability through feature selection methods**.**

Keywords-component; wind turbine vibrations; feature selection; wind energy systems; vibration prediction; correlation; feature importance

I. INTRODUCTION

The explosive growth of wind energy during recent years spurs researchers to find solutions for existing challenges. The scientific literature has mainly explored wind energy conversion [1], wind power prediction [2], wind speed forecasting [3], and turbine monitoring [4]. Still, it lacks a thorough investigation of wind turbine vibrations, significantly affecting turbine performance and operational life. Wind turbine acceleration prediction is a primary research target because it supports crucial tasks in wind energy system maintenance alongside efficiency optimization and stability regulation. Growing renewable energy demand requires enhanced turbine optimization and improved reliability systems to meet industry performance standards. Effective acceleration prediction remains essential for executing predictive and condition-based maintenance (CBM) programs since accurate performance forecasting supports maintenance scheduling while minimizing unexpected shutdowns and related expenses. Adopting predictive maintenance through turbine performance monitoring can substantially decrease operational expenses since problems get resolved before equipment failure happens, according to [5]. Actual predictions of acceleration enable enhanced turbine efficiency by allowing better control of operational loads during changing wind conditions [5]. Concretely, Wang and Qu demonstrate in [6] how predictive maintenance systems that base their analysis on generator current data help improve turbine operational efficiency alongside lowering maintenance costs. Accurate forecasting helps operators maximize resource distribution while enhancing maintenance planning, generating increased energy output, and minimizing operational downtime [6].

Initial wind turbine vibration research conducted studies using basic principles to develop models and perform simulations. Leithead and Connor investigated variable-speed wind turbine dynamics and control model design [7] while [8] studied tower vibration effects under voltage sags through simulation platforms TRUBSIM FAST and SIMULINK. Murtagh et al. investigated the potential of passive control systems to reduce vibrations [9]. These parametric models need several assumptions to function, reducing their ability to show real turbine behavior correctly. Wind power infrastructure growth demands advanced and accurate turbine maintenance tactics, which require better modeling techniques to ensure performance quality.

Developing predictive models for wind turbine performance depends heavily on effective feature selection within data-driven approaches. Identifying essential dataset features through selection greatly benefits model performance and accuracy levels. Feature selection affects multiple essential domains. The feature selection process optimizes computational performance by simplifying input data dimensions and accelerating model training processes and prediction calculations. According to Feng [10], decreasing feature dimensions help eliminate processing bottlenecks in massive Supervisory Control and Data Acquisition (SCADA) system datasets. Removing unnecessary features enhances prediction performance because these components often produce unwanted noise, which can produce overfitting effects. [11] data classification accuracy and model generalization strength increased after removing noisy features. A significant benefit of feature selection methods includes making models easier to interpret. Wind turbines demonstrate the importance of determining key features to understand system operations better and achieve optimization targets. Several studies [11] highlight ensemble techniques and the proper selection of input variables as critical components for precision wind power prediction. The selection of advanced features through specialized techniques solves performance issues that commonly surface in wind energy applications because of large data sets. Large data volumes become manageable through methods that enhance prediction accuracy, according to research findings presented in [12].

This research describes the data and then prepares data for applying the selecting features algorithm. In the next step, three algorithms are used to find more critical features as input of the model the model output is the acceleration of the wind tower, and finally, the influence of inputs on production has been studied, and the parameters of wind turbine with high impact on the wind turbine vibration has been chosen.

II. Dataset Description and Preprocessing

A. Data Description

This research analyzes a dataset from supervisory control and data acquisition (SCADA) systems installed at a large wind farm. The data consists of 27k records sampled at a frequency of 1 Hz over one month. The selected time frame was one month since it achieved an optimal trade-off between operational variability detection and computational requirements. The analysis duration of one week might produce questionable outcomes because it fails to display adequate variations between wind patterns and turbine operational behavior. Expanding data collection from one month to two or three months would result in larger datasets that require higher computational resources and processing time while probably delivering no new insights. Analysis of turbine behavior patterns for predictive maintenance purposes can be performed effectively during a one-month observation period. The SCADA system monitors 301 parameters, and this study considers a subset of those parameters that are deemed most relevant to the research objectives. An example of the dataset used has been shown in Table 1. The values from the dataset have many time-stamped parameters, such as torque, wind speed, wind deviation, drive train acceleration, and tower acceleration.

TABLE I. Sample dataset for wind turbine

Observation Number	**Features**				
	Wind speed (m/s)	***Generator (RPM)***	***Wind direction (°)***	**……**	***Tower Acceleration Y (mm/s^2)***
1	7.38	1569	300.70	….	18.42
2	6.69	1363	295.85	….	20.04
3	6.50	1386	296.53	….	18.51
4	6.48	1357	291.88		16.83
5	6.51	1393	293.98		17.90
…				….	
27610	6.12	1280	6.360	….	14.76

The drive train and tower acceleration are particularly important as they represent vibrations in the wind turbine drive train and tower, respectively [4]. However, since sensor malfunctions, transmission issues, and subsystem failures caused some errors in the dataset, the dataset was processed to improve the accuracy of the data-driven model [13].

B. Data Preprocessing.

First, prepare the data for feature selection. Since algorithms and machine learning techniques are commonly used to extract features, data preprocessing was necessary to verify the dataset is high quality and high precision. This process removed columns or parameters with more than 10% missing values (NaN). Moreover, those parameters whose values remained constant in more than 80% of the dataset have been excluded since their contribution to the feature selection or model training is minimal. Also, the missing data from the selected output have been removed as we used the machine learning algorithm, and if there is a NaN, the algorithm could not be applied. After applying these, the dataset was reduced to 241 columns and 26063 rows. Applying these preprocessing steps improved the dataset's quality, but no dimensionality reduction techniques were used to reduce its size. This approach intentionally retained all the meaningful features for further analysis and model development.

III. Feature Selection of Wind Turbine Vibration

This section applies advanced feature selection techniques, including Correlation Analysis, Mutual Information, and Feature Importance using Random Forests and XGBoost. These methods are employed to systematically identify and prioritize the most relevant features for model development, ensuring optimal performance and minimizing redundancy. For the output, as we want to predict tower acceleration, the tower acceleration in the y direction is selected as the output, as shown in Fig. 1

Figure 1. Wind Turbine

A. *Correlation Analysis*

The analysis used a correlation matrix to study linear relationships between all features and the outcome variable. The analysis revealed highly correlated features because these features create model problems with redundancy and multicollinearity. The model selection process focused on features that demonstrated significant ties to y-direction tower acceleration while maintaining low connection points with other variables to achieve diverse and non-redundant predictive inputs. The output variable exhibits correlations with features that span from 0.917 through -0.57, as shown in Tables 2 and 3. The model correlations align with Wind speed and Power factors. The predictive input variables are wind speed, Wind speed, Standard deviation, power, Energy Export, and Motor current, among the highest correlated features.

The first step for model-building feature selection required identifying features that achieved either more significant than 0.7 or less than -0.5 correlation with the target variable, as this limitation helps us to reduce some irrelevant parameters. The feature reduction process depends on the importance of rankings and their correlation with other features. Correlation matrix analysis verified the presence of strongly interrelated features with other variables. Our selection of input variables included just one representative from each set of highly correlated features to prevent information duplication. Thus, selected features bring distinct and valuable information to the model structure.

TABLE II. HIGHLY CORRELATED FEATURES WITH TOWER ACCELERATION

Correlation to Tower acceleration	**High Correlated Features**				
	Wind speed (m/s)	***Wind speed, Standard deviation (m/s)***	***Power (kW)***	***Energy Export (kWh)***	***Motor current (A)***
Correlation Value	0.917	0.876	0.843	0.826	0.797

TABLE III. LOW CORRELATED FEATURES WITH TOWER ACCELERATION

Correlation to Tower acceleration	**Low Correlated Features**				
	Power factor	***Nacelle temperature (°C)***	***Generator temperature (°C)***	***Blade angel (°)***	***CPU temperature (°C)***
Correlation Value	-0.57	-0.450	-0.22	-0.18	-0.171

The VIF measurement technique evaluates multivariate relationships in datasets by showing how much features impact each other through their correlations. The VIF measurement of "Wind Speed" stood at 561, an unacceptably elevated value. The model developers removed these features one by one through successive iterations to lower the VIF scores for related components, including Wind Speed levels

Post-technique implementation, the database size decreased from 300 to 100 features. Determining linear associations is the only capability of correlation analysis when studying features with output variables. The connection between wind turbine parameters and tower acceleration is inherently nonlinear [13]. The analysis employed mutual information because it detects nonlinear feature dependencies

B. *Mutual Information*

Statistical measure Mutual Information helps evaluate variable dependencies through determined uncertainty reductions in one variable when another is known. The ability of Mutual Information to detect nonlinear and linear relationships makes it suitable for analyzing complex data from wind turbine dynamics. Mutual Information served to determine which features most effectively predicted tower acceleration. Higher MI scores determined which features provided the best understanding and prediction accuracy for turbine vibration analysis. This approach analyzed all features before displaying the 20 most impactful parameters in Fig. 2. The study reveals that wind speed substantially impacts acceleration, but other measurements, such as torque, generator speed, and rotor speed, also significantly affect the vibration.

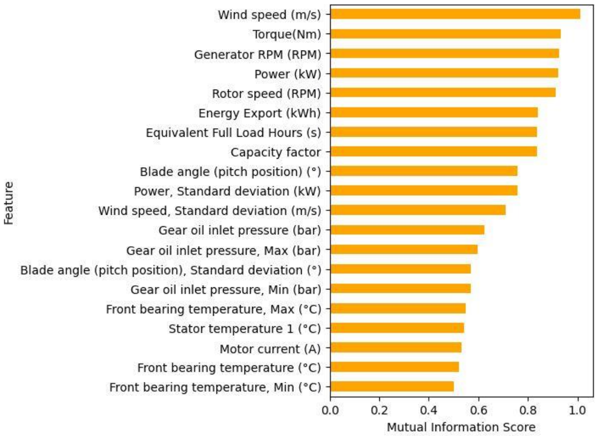

Figure 2. Top 20 features by mutual information

The analysis demonstrates challenges in feature differentiation. Fig. 2 shows that most parameters have nearly equivalent effects on acceleration, suggesting the need for more sophisticated methods to accurately identify and prioritize the most significant features driving turbine vibration.

C. Feature Importance With Random Forests And XGBoost

The Random Forest algorithm constructs data-driven trees across subsets while assessing feature importance by measuring how tree impurity reduces whenever each feature splits the data. It receives a higher importance score by measuring a feature's ability to partition data for improved prediction accuracy. XGBoost method builds trees one after another in a sequential process while each new tree works to address previous errors. The feature importance evaluation in "gain" metrics shows how using a specific variable leads to better splitting performance. The algorithms determine vital prediction components through their built-in ranking, establishing features based on their error reduction capacity. This essential characteristic evaluation makes them excellent for uncovering fundamental decision drivers of model performance. A visual representation of the random forest algorithm result is displayed in Fig. 3. The most important features are Motor current, power, wind speed, and blade angle.

The XGBoost algorithm illustrated in Fig. 4 also supports the importance of the key parameters enhanced through the random forest technique with significant predictive importance. Accurate current and power measurements are obtained but are based on essential variables such as wind velocity and actions performed by the generator. Although the use of MI and feature importance is helpful in determining features relevant to the description of turbine vibration, these techniques present some drawbacks. For example, MI rankings may indicate current and power as significant due to correlation coefficients, while these variables are the second-order effects infested by wind speed and control settings. MI may not capture some of the most important aspects and characteristics of the turbine dynamics since it assesses statistical dependencies and not physical processes [13].

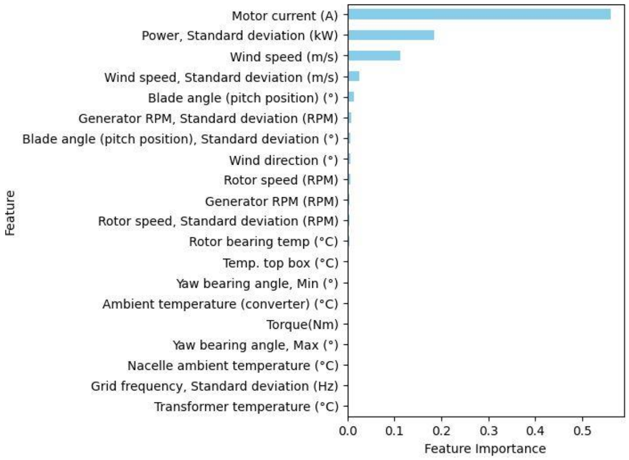

Figure 3. Top 20 features importance by Random Forest

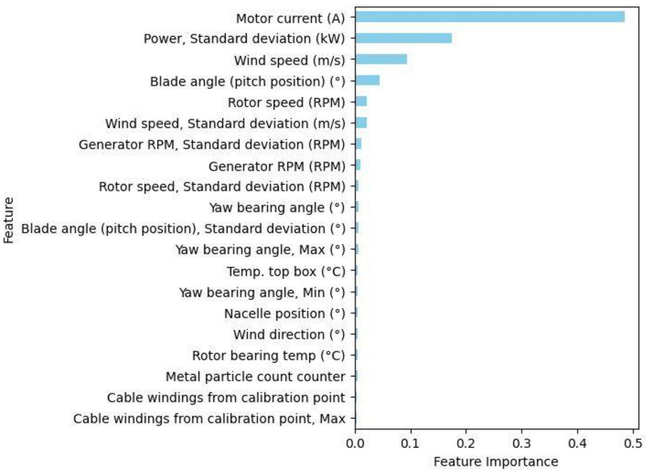

Figure 4. Top 20 features importance by XGBoost

Four significant variables significantly affect the mechanical behavior of wind turbines: wind velocity, deviation angle, rotor velocity, and blade angle [13]. These variables control speed and angles, resulting in aerodynamic and structural forces that create vibrations and accelerations. While current and power are related to electrical performance, these parameters directly affect mechanical and aerodynamic forces that determine vibration [4]. They can achieve this since the effects of the parameters are minimized, and each parameter presents information that is different from the rest. The wind speed and the blade's angle control aerodynamic fluctuations, while the speed of the rotor gives information on the mechanical system. Engineering studies focus on these parameters since they define key forces acting on turbines, aligning with engineering research principles [13].

IV. Results

The results of the feature selection analyses conducted to identify the most significant parameters for predicting tower and drivetrain vibrations are summarized in Tables 4 and 5. The performance of key features, including wind speed, Rotor speed, wind speed standard deviation, and blade angle, was evaluated using four different methods: correlation analysis, mutual information (MI), feature importance by using Random Forest, and XGBoost. it is essential to note that the values presented for each method are independent and should not be directly compared across methods. Each method evaluates feature importance using a different approach, and the scales of their outputs may vary.

Across each evaluation methodology, wind speed proved to be the key driver of tower acceleration Table 4. The wind speed and tower acceleration demonstrated a robust positive relationship according to correlation analysis results of 0.916, yielding the highest mutual information score of 1.028 among all features examined in Table 4. The two ensemble learning methods, XGBoost and Random Forest revealed wind speed as the most critical predictor through their importance scores, which stood at 0.7619 and 0.7385, respectively.

TABLE IV. FEATURES RANKINGS FOR TOWER ACCELERATION

Features	Feature Importance Based on Different Methods			
	Correlation	*MI*	*XGBOOST*	*Random Forest*
Wind speed	0.916	1.028	0.7619	0.7385
Wind speed, Stdv	0.878	0.723	0.0776	0.0883
Rotor speed	0.671	0.898	0.1418	0.1480
Blade angle (pitch position)	-0.303	0.768	0.0187	0.0253

As shown in Table 5, the drivetrain acceleration analysis demonstrated that rotor speed functioned as the primary determining factor of special significance to machine learning models. The XGBoost and Random Forest models selected rotor speed because it was the principal predictor based on both importance scores, respectively 0.765 and 0.7501, and outpaced all other features. There is a moderate correlation between rotor speed and tower measurements 0.671 and drivetrain acceleration measurements 0.735 Table 4,5. The mutual information analysis detected strong non-linear associations between tower acceleration, which reached 0.898, and drivetrain acceleration, which hit 0.980.

Wind speed standard deviation affected both acceleration types moderately or intensely. Still, it significantly affected drivetrain acceleration more than tower acceleration based on XGBoost 0.190 and Random Forest 0.1801 models Table 5.

The poor associations of blade angle with the tower acceleration -0.23 and drivetrain acceleration -0.31, as noted in Table 4,5, indicate minimal linear behavior. Mutual information analysis suggested moderate non-linear contributions, with scores of 0.768 for tower acceleration and 0.66 for drivetrain acceleration. XGBoost and Random Forest modeling techniques confirmed that blade angle holds a lower importance rating among features when predicting both target variables, reflecting its minimal influence on these outcomes.

Statistical evaluation through correlation and mutual information produces different results than the scores calculated through machine learning models, indicating that the operational parameters could exhibit non-linear connections to turbine accelerations. Wind speed standard deviation is an example of a situation where traditional methods produced stronger correlations than ensemble methods.

TABLE V. FEATURES RANKINGS FOR DRIVE TRAIN ACCELERATION

Features	Feature Importance Based on Different Methods			
	Correlation	*MI*	*XGBOOST*	*Random Forest*
Wind speed	0.868	0.9437	0.0287	0.0474
Wind speed, Stdv	0.846	0.7495	0.190	0.1801
Rotor speed	0.735	0.9802	0.765	0.7501
Blade angle (pitch position)	-0.368	0.6602	0.0154	0.0224

V. CONCLUSION

Based on the feature selection analysis, this study offers important information on the interaction of the operational parameters and vibration characteristics. Analyzing the responses with the help of mainstream statistical methods and machine learning algorithms also identified separate patterns of the influence for tower and drivetrain accelerations.

Therefore, tower acceleration is determined mainly by the wind speed, as evidenced by all the evaluation metrics. This correlation strongly indicates that wind speed is a parameter that the monitoring system should closely monitor for condition monitoring of towers. On the other hand, drivetrain acceleration depends more on rotor speed maximization, especially in machine learning models, and therefore, health monitoring of the drivetrain should focus on the rotor speed variations. This indicates that wind turbines exhibit non-linear behavior, so the importance scores obtained from machine learning models differ from the statistical measures. This was even more so for the standard deviation of wind speed, which had different significance levels in various analysis methods. The low influence of blade angle in all the measures indicates that while pitch control is still vital for power enhancement, it is probably not the key factor in the vibration analysis.

To improve the prediction accuracy of the tower vibration, the lagged variables in the analysis framework, especially in the XGBoost model, can consider the structure's time dependence and dynamic responses to the wind loading condition. This approach also recognizes that the current vibration states depend on the previous operational states.

Moreover, data preprocessing by removing noise before feeding to the feature selection and model training phases will enhance the signal-to-noise ratio and improve the model's reliability. These preprocessing steps are more critical when the collected data is high-frequency vibration, which may contain some measurement noise or environmental interferences

REFERENCES

[1] H.-S. Ko, K. Y. Lee, M.-J. Kang, and H.-C. Kim, “Power quality control of an autonomous wind-diesel power system based on hybrid intelligent controller,” Neural Networks, vol. 21, no. 10, pp. 1439–1446, 2008.

[2] A. Kusiak, H.-Y. Zheng, and Z. Song, “Short-term prediction of wind farm power: A data-mining approach,” IEEE Trans. Energy Convers., vol. 24, no. 1, pp. 125–136, 2009.

[3] M. Monfared, H. Rastegar, and H. M. Kojabadi, “A new strategy for wind speed forecasting using artificial intelligent methods,” Renewable Energy, vol. 34, no. 3, pp. 845–848, 2009.

[4] A. Schröder, H.-Y. Zheng, and Z. Song, “On-line monitoring of power curves,” Renewable Energy, vol. 34, no. 5, pp. 1487–1493, 2009.

[5] L. Schröder, N. Dimitrov, D. Verelst, and J. Sorensen, “Using transfer learning to build physics-informed machine learning models for improved wind farm monitoring,” Energies, vol. 15, no. 2, p. 558, 2022.

[6] Q. Wang and L. Qu, “Prognostic condition monitoring for wind turbine drivetrains via generator current analysis,” Chinese Journal of Electrical Engineering, vol. 4, no. 3, pp. 80–89, 2018.

[7] W. Leithead and B. Connor, “Control of variable speed wind turbines: Dynamic models,” Int. J. Control, vol. 73, no. 13, pp. 1173–1188, 2000.

[8] R. Fadaeinedjad, G. Moschopoulos, and M. Moallem, “Investigation of voltage sag impact on wind turbine tower vibrations,” Wind Energy, vol. 11, no. 4, pp. 351–375, 2008.

[9] P. J. Murtagh, A. Ghosh, B. Basu, and B. M. Broderick, “Passive control of wind turbine vibrations including blade/tower interaction and rotationally sampled turbulence,” Wind Energy, vol. 11, no. 4, pp. 305–317, 2008.

[10] C. Feng, M. Cui, B. Hodge, and J. Zhang, “A data-driven multi-model methodology with deep feature selection for short-term wind forecasting,” Applied Energy, vol. 190, pp. 1245–1257, 2017.

[11] K. Pes R. and S. Jacob, “Improved random forest algorithm for software defect prediction through data mining techniques,” Int. J. Comput. Appl., vol. 117, no. 23, pp. 18–22, 2015.

[12] S. Ally, “Modular deep learning approach for wind farm power forecasting and wake loss prediction,” 2023.

[13] A. Kusiak and Z. Zhang, “Control of wind turbine power and vibration with a data-driven approach,” Renewable Energy, vol. 43, pp. 73–82, 2012.