Proceedings of the 6th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/2026.30660

High-Precision Prediction of the Remaining Service Life of Wind Power Equipment Based on Machine Learning Algorithms

Wenbin Cheng

Key Laboratory of Precision Drive and Equipment, Jiangxi University of Water Resources and Electric Power, Nanchang, China
2441469380@qq.com

Abstract. The global energy transition process is continuously accelerating. As a core pillar of clean, low-carbon and renewable energy, the installed capacity and operation scale of wind power equipment based on machine learning algorithms are steadily increasing. The prediction of the remaining service life of wind turbine gearboxes is crucial for the safe operation and maintenance of equipment. However, existing machine learning algorithms generally have problems such as feature redundancy, which limits the generalization performance of the model, and insufficient accuracy in capturing complex degradation trends. To enhance the prediction effect, this paper proposes the FVIM-XGBoost regression algorithm. Firstly, the features are optimized through correlation analysis, and then comparative experiments are conducted with GBDT, ExtraTrees, random forest, CatBoost, decision tree and BP neural network. The results show that the proposed model has the best comprehensive prediction performance. Its coefficient of determination R^2 reaches 0.947, which is significantly higher than that of other comparison models, and the fitting effect is the best. The mean square error of 1855.454, the root mean square error of 43.075, and the mean absolute error of 29.958 are all lower than those of the other models, and the prediction accuracy is better. Only the mean absolute percentage error of 62.113 is slightly higher than that of ExtraTrees, GBDT and BP neural networks. This algorithm provides a more reliable prediction scheme for the remaining service life of wind power gearboxes, which is of great practical significance for ensuring the stable operation of wind power equipment, reducing operation and maintenance costs, and promoting global energy transformation and low-carbon development. High-precision prediction of remaining service life

Keywords: Machine learning, wind turbine gearboxes, prediction of remaining useful life.

1. Introduction

Against the backdrop of the accelerated global energy transition, wind power, as a core component of clean and low-carbon renewable energy, has seen its installed capacity and operational scale continue to expand [1]. As a core transmission component of wind turbine generators, the wind turbine gearbox is responsible for converting the low rotational speed of the impeller into the high rotational speed of the generator. It is constantly exposed to complex outdoor working conditions with variable loads and rotational speeds, and is prone to faults due to wear, fatigue, and lubrication failure [2]. Gearbox failure not only leads to the shutdown of the unit, resulting in huge loss of power generation, but also brings high maintenance costs and safety risks. The traditional regular maintenance mode is difficult to accurately match the actual degradation state of components, and it is easy to have over-maintenance or under-maintenance situations [3]. Therefore, accurately predicting the remaining service life of wind power gearboxes has become a key demand for improving the operation and maintenance efficiency of wind power equipment, reducing operating costs, and ensuring the safe and stable operation of units. It is also a current research hotspot and cutting-edge direction in the field of intelligent operation and maintenance [4].

Traditional methods for predicting the remaining useful life mostly rely on physical model construction or manual empirical judgment, which are difficult to deal with the complex degradation mechanism of multi-factor coupling during the operation of wind turbine gearboxes, and have limited mining capabilities for multi-source monitoring data [5]. Machine learning algorithms, with their core advantage of being data-driven, provide an effective path to solve this problem. This type of algorithm does not require the presetting of complex physical models and can automatically mine potential degradation patterns from multi-dimensional monitoring data such as operating speed, vibration, acceleration, temperature and humidity, accurately capturing the nonlinear correlations among variables. Through the learning and training of historical operation data and fault records, machine learning algorithms can achieve quantitative prediction of the remaining service life, and have good real-time performance and adaptability. They can flexibly respond to the impact of complex working conditions such as wind speed fluctuations and load changes, significantly improving the prediction accuracy and reliability, and providing a scientific basis for the predictive maintenance of wind turbine gearboxes. Promote the transformation of the operation and maintenance mode from passive response to active prediction [6].

Although existing machine learning algorithms have made certain progress in predicting the remaining service life of wind turbine gearboxes, there are still problems such as insufficient generalization ability of the model due to feature redundancy and inaccurate capture of complex degradation trends. To further enhance the prediction performance, this paper proposes the FVIM-XGBoost regression algorithm. This algorithm is based on the gradient boosting decision tree framework and integrates the feature importance assessment mechanism. By precisely screening the key features that have a significant impact on the remaining service life, it effectively reduces the interference of redundant information and improves the training efficiency and generalization ability of the model.

2. Data sources

This dataset is open-source data and can be used for the task of predicting the remaining service life of wind turbine gearboxes. It contains 871 valid records, covering eight variables: operating speed, vibration acceleration, gearbox oil temperature, environmental temperature and humidity, cumulative operating time, load factor, and lubricating oil contamination degree. The predicted variable is the remaining service life hours. This dataset is suitable for the training and validation of machine learning regression prediction models. Select some data for display, as shown in Table 1.

DOI: 10.54254/2755-2721/2026.30660

Table 1. The statistical results of the dataset

rotational_speed_rpm	oil_temperature_c	vibration_accel_g	ambient_humidity_pct	ambient_temperature_c	operating_hours	load_factor_pct	oil_contamination_nas	remaining_useful_life_hours
1770.4	60.8	0.025	44	18.7	4087	93.1	9	341.6
1639.2	51.3	0.199	31.1	9.2	9759	49.1	10	34.2
1559.2	36	0.227	46.6	-2.9	9003	75.9	6	160.2
1293.6	41.2	0.028	41.2	-3.7	3588	79.9	7	537.2
1293.6	63.6	0.132	75.6	11.9	1834	65.2	10	424.2

Correlation analysis was conducted on each variable, and a correlation heat map was drawn, as shown in Figure 1. The remaining service life hours have the strongest correlation with the cumulative operating time, with a correlation coefficient of -0.77. This indicates that as the cumulative operating time increases, the remaining service life will significantly decrease, and there is a clear negative correlation between the two. The correlation coefficient between vibration acceleration and the remaining service life is -0.42, showing a strong negative correlation, indicating that the greater the vibration acceleration, the shorter the remaining service life. The correlation coefficients between the oil temperature of the gearbox and the contamination degree of the lubricating oil and the remaining service life are -0.18 and -0.28 respectively, showing a certain negative correlation.

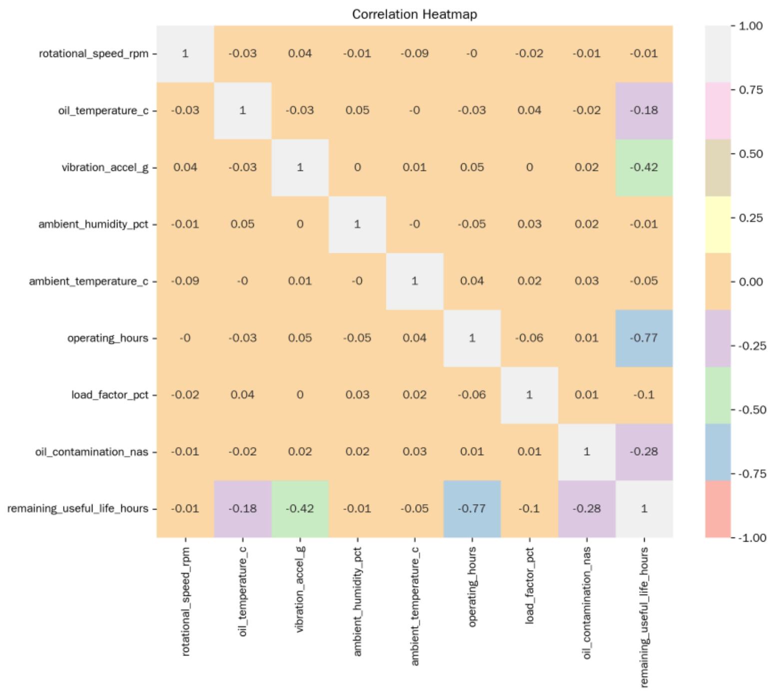

Figure 1. Correlation analysis

DOI: 10.54254/2755-2721/2026.30660

3. Method

3.1. FVIM

The core of the FVIM algorithm is to achieve the screening of key features and the elimination of redundant information by quantifying the degree of influence of features on the predicted target. The principle is based on the correlation strength between features and target variables as well as the interaction relationship among features. The importance score of each feature is calculated through statistical analysis or model derivation. The algorithm first constructs the mapping relationship between features and target variables, and uses methods such as variance contribution, information gain or partial correlation analysis to evaluate the explanatory power of individual features for the prediction results. At the same time, consider the multicollinearity among features to avoid redundant features interfering with the model performance. For mechanical engineering prediction tasks, FVIM can accurately identify core features that have a significant impact on the remaining service life of equipment, such as vibration acceleration and cumulative running time, and eliminate irrelevant variables with weak correlation. This simplifies the model input dimension, reduces computational complexity, provides high-quality feature subsets for subsequent modeling, and improves the training efficiency and generalization ability of the model.

3.2. XGBoost

The XGBoost (Extreme Gradient Boosting) algorithm is an ensemble learning algorithm based on gradient boosting decision trees, with the additive model and gradient descent as the core design ideas [7]. The principle is to train multiple weak learners iteratively, weighted and fused the prediction results of all weak learners to form a strong learner. In each iteration round, the algorithm optimizes the objective function based on the prediction residuals of the previous round's model using the gradient descent method to generate a new decision tree. Simultaneously introduce L1 and L2 regularization terms to limit the complexity and weight scale of the tree, effectively preventing overfitting. In addition, XGBoost supports the automatic processing of missing values, dynamically selects the direction of missing values through gain calculation when splitting nodes, and has parallel computing capabilities, which can significantly improve the training speed [8]. This algorithm has an extremely strong fitting ability for nonlinear relationships and high-dimensional data, and can accurately capture the complex correlations among multiple variables in the operation data of mechanical equipment. It demonstrates excellent accuracy and stability in regression prediction tasks. The network structure of XGBoost is shown in Figure 2.

DOI: 10.54254/2755-2721/2026.30660

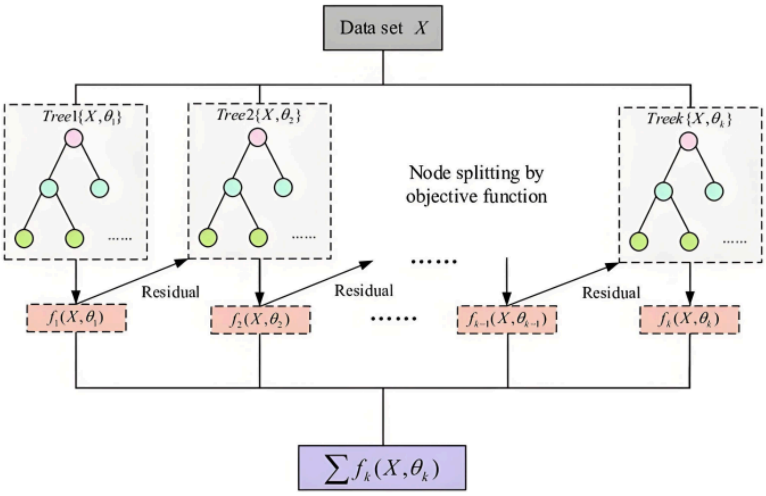

Figure 2. The network structure of XGBoost

3.3. FVIM-XGBoost

The FVIM-XGBoost algorithm is a hybrid algorithm that organically integrates the FVIM feature screening mechanism with the XGBoost ensemble learning framework. Its core is a two-step strategy of screening first and then modeling [9]. The principle is as follows: Firstly, the original high-dimensional features are preprocessed through the FVIM algorithm. Based on the quantification score of the influence intensity of the features on the remaining service life, key features such as vibration acceleration and cumulative running time are retained, and low-contribution redundant variables such as environmental humidity are eliminated to construct a streamlined and effective feature set. Subsequently, the screened features are input into the XGBoost model. By leveraging its gradient boosting mechanism and regularization characteristics, the nonlinear degradation relationship between the features and the remaining service life is precisely fitted. This fusion algorithm not only takes advantage of FVIM's strengths in reducing noise interference and simplifying model structure, but also leverages XGBoost's powerful nonlinear fitting and error correction capabilities. It effectively addresses the issues of overfitting and insufficient generalization ability of a single algorithm in high-dimensional mechanical data, achieving dual optimization of feature quality and modeling accuracy. It is particularly suitable for predicting the remaining life of complex mechanical systems with multiple operating conditions and variables, such as wind turbine gearboxes [10].

4. Result

In terms of parameter Settings, the number of search agents is set to 5, the maximum number of iterations is set to 15, and the objective function adopts linear regression. During data processing, the training set accounts for 0.7 of the total data. The data is normalized to between -1 and 1 through mapminmax. The model training adopts 5-fold cross-validation to evaluate the parameter performance.

For the comparative models, this paper conducts experiments using Decision tree, Random Forest, GBDT, Extratrees, CatBoost and BP neural network, and outputs the index evaluation tables of each model, as shown in Table 2. Output the comparison charts of each indicator, as shown in Figure 3.

DOI: 10.54254/2755-2721/2026.30660

Table 2. The results of the comparative experiment

Model	MSE	RMSE	MAE	MAPE	R^2
Decision tree	5731.235	75.705	51.553	146.983	0.832
Random Forest	3035.696	55.097	39.574	72.847	0.91
GBDT	2027.343	45.026	30.368	55.548	0.937
Extratrees	2439.693	49.393	37.609	53.903	0.92
CatBoost	3434.773	58.607	43.2	49.2	0.908
BP neural network	5483.594	74.051	58.959	51.478	0.816
Our model	1855.454	43.075	29.958	62.113	0.947

The comprehensive prediction performance of Our model is the best, and its coefficient of determination R2 reaches 0.947. The fitting effect is the best among all the models, which is higher than 0.937 of GBDT, 0.92 of Extratrees, 0.91 of random forest, 0.908 of CatBoost, 0.832 of decision tree and 0.816 of BP neural network. In terms of error indicators, the mean square error of Our model is 1855.454, the root mean square error is 43.075, and the mean absolute error is 29.958, all of which are lower than those of all other models, demonstrating more accurate predictive ability. However, the mean absolute percentage error is 62.113. It is slightly higher than Extratrees' 53.903, GBDT's 55.548 and BP neural network's 51.478. In this metric, it is slightly inferior to these three models, but its overall performance is still significantly better than that of other models such as decision trees, random forests and CatBoost.

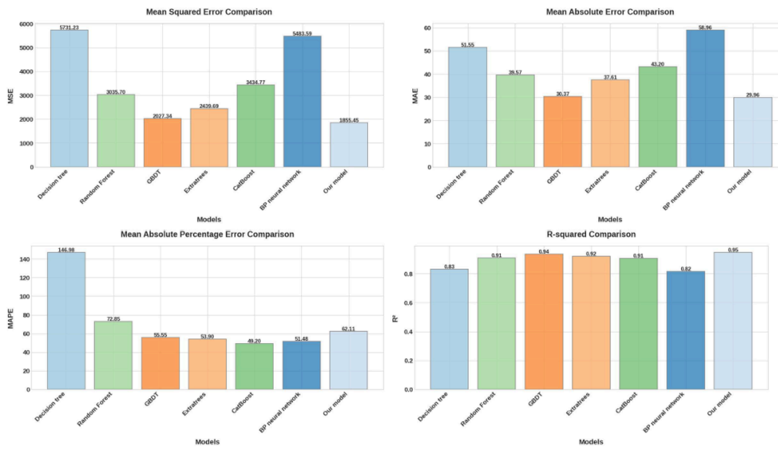

Figure 3. The bar chart comparison of each indicator

Output the line graph of the predicted value - actual value of the Our model test set, as shown in Figure 4.

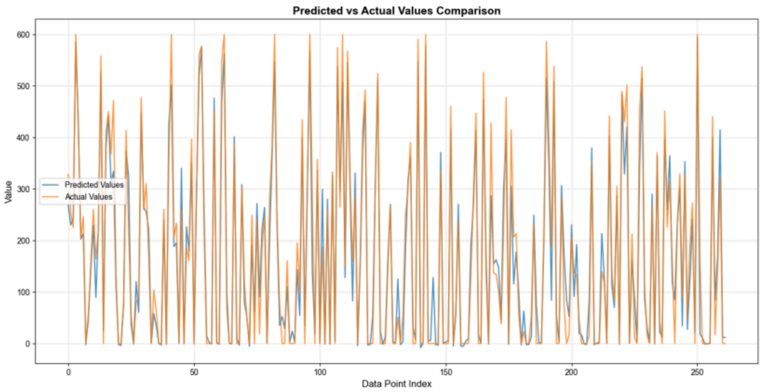

Figure 4. The line graph of the predicted values - actual values of the test set of Transformer-BiGRU

5. Conclusion

Against the backdrop of the continuous deepening of the global energy transition, wind power, as a core force of clean, low-carbon and renewable energy, has seen its installed capacity and operational scale continuously expand. The prediction of the remaining service life of wind power gearboxes is crucial for ensuring the stable operation of wind power equipment. Although existing machine learning algorithms have achieved certain results, they still face problems such as insufficient generalization ability of the model due to feature redundancy and inadequate accuracy in capturing complex degradation trends. To this end, this paper proposes the FVIM-XGBoost regression algorithm. Firstly, the data basis is optimized through correlation analysis, and then comparative experiments are conducted with GBDT, Extratrees, random forest, CatBoost, decision tree and BP neural network. The results show that the proposed algorithm has the best comprehensive prediction performance. The coefficient of determination R^2 reaches 0.947, which is higher than that of all other comparison models, and the fitting effect is the best. The mean square error of 1855.454, the root mean square error of 43.075, and the mean absolute error of 29.958 are all lower than those of the other models, and the prediction accuracy is better. Only the mean absolute percentage error of 62.113 is slightly higher than that of Extratrees, GBDT and BP neural networks. The overall performance is still significantly superior to models such as decision trees and random forests. This algorithm effectively breaks through the limitations of existing methods, providing a more reliable prediction scheme for the remaining service life of wind power gearboxes. It has significant practical significance for improving the operation and maintenance efficiency of wind power equipment, reducing operating costs, and promoting the high-quality development of the wind power industry.

References

[1] Cheng, Fangzhou, Liyan Qu, and Wei Qiao. "Fault prognosis and remaining useful life prediction of wind turbine gearboxes using current signal analysis." IEEE Transactions on Sustainable Energy 9.1 (2017): 157-167.
[2] Carroll, James, et al. "Wind turbine gearbox failure and remaining useful life prediction using machine learning techniques." Wind Energy 22.3 (2019): 360-375.
[3] Cheng, Fangzhou, et al. "Enhanced particle filtering for bearing remaining useful life prediction of wind turbine drivetrain gearboxes." IEEE Transactions on Industrial Electronics 66.6 (2018): 4738-4748.
[4] Akash, R., et al. "Data driven approach for estimation of remaining useful life of gearbox in wind energy conversion systems." Life Cycle Reliability and Safety Engineering (2025): 1-13.

DOI: 10.54254/2755-2721/2026.30660

[5] Vietze, Daniel, Michael Hein, and Karsten Stahl. "Method for a cloud based remaining-service-life-prediction for vehicle-gearboxes based on big-data-analysis and machine learning." Forschung im Ingenieurwesen 84.4 (2020): 305-314.
[6] Lázaro, Roberto, Nurseda Y. Yürüşen, and Julio J. Melero. "Determining remaining lifetime of wind turbine gearbox using a health status indicator signal." Journal of Physics: Conference Series. Vol. 1618. No. 2. IOP Publishing, 2020.
[7] Li, Linli, and Qifei Jian. "Remaining useful life prediction of wind turbine main-bearing based on LSTM optimized network." IEEE Sensors Journal 24.13 (2024): 21143-21156.
[8] Godwin, Jamie L., and Peter Matthews. "Prognosis of wind turbine gearbox failures by utilising robust multivariate statistical techniques." 2013 IEEE conference on Prognostics and Health Management (PHM). IEEE, 2013.
[9] Yu, Xiaoxia, et al. "Non-temporal neural networks for predicting degradation trends of key wind-turbine gearbox components." Renewable Energy 243 (2025): 122438.
[10] Salameh, Jack P., et al. "Gearbox condition monitoring in wind turbines: A review." Mechanical Systems and Signal Processing 111 (2018): 251-264.