scientific reports

Check for updates

OPEN

A novel diagnosis methodology of gear oil for wind turbine combining Stepwise multivariate regression and clustered federated learning framework

Huihui Han[1], Ye Zhao[1], Hao Jiang[2], Muxin Chen[1], Song Zhou[2], Zihan Lin[3], Xin Wang[1], Boman Mao[4], Xinyue Yang[4] & Yuchun Li[3]✉

Data-driven approaches demonstrate significant potential in accurately diagnosing faults in wind turbines. To enhance diagnostic performance, we introduce a clustered federated learning framework (CFLF) for wind gear oil diagnosis. Initially, a stepwise multivariate regression (SMR) model is introduced and optimized after data processing, which integrates multiscale features and an AIC-diagnosis feature. Subsequently, to tackle data heterogeneity among different indicators, a series of canonical correlation representations are extracted from the SMR models, and a combined model of CFLF method and SMR is proposed to assess the performance of gear oil. Actual data analysis of wind turbine gear oil showcase the superior performance of the proposed model over the single SMR model with higher prediction accuracy of 35.73%. This study provides a new technique for evaluating gear oil in the wind energy sector.

Keywords Stepwise multivariate regression, Gear oil evaluation, Clustered federated learning framework, Wind turbines

Wind energy is a very desirable green choice among many new energy sources because it is a renewable and clean energy source with recyclability and high efficiency[1–3]. The main gear lubrication systems play a key role in the normal operation of wind turbines. However, the large-scale wind power poses a critical impact on the main gear lubrication systems, makes gear oil a key monitoring target for wind turbine system. Addressing this challenge requires a reliable analysis method for comprehensive analysis of gear oil and evaluation model, facilitating optimal maintenance time selection for wind turbines[4]. As a result, main gear oil analysis and modeling has become a focal point, particularly with the application of deep learning methods. A keyword search for 'gear oil' and 'wind turbines', there are 23,558 documents from 2014 to 01 to 2024-11[5], and the annual average number of documents issued is 2,142. As shown in Figs. 1 and 2023 reached a peak of 3,069 annual publications, and 2019 has the fastest growth rate of 12.64%, suggesting rapid development and continued growth in this field. 23,558 papers were retrieved, and the top 30 journals in terms of the number of publications are shown in Fig. 1, in which the journal with the most publications is Energies (795 articles); Journal of Physics: Conference Series ranks second, with 769 articles; Renewable Energy ranks third, with 511 articles.

Stepwise multivariate regression (SMR) analysis is one of the linear regression methods, which combines multivariate and stepwise modeling techniques. Limitations of linear models lie in poor fit for nonlinear relationships, sensitivity to outliers and insufficient handling of multicollinearity problems[6–9]. Deep learning, as a branch of machine learning, is widely applied in industry due to its remarkable feature learning capabilities and capacity for handling high-dimensional data[10–13]. To address data inhomogeneity characteristic and overcome data sparsity, Federated Learning (FL) has been introduced as a collaborative distributed machine learning approach in fault diagnosis[14–18]. A hybrid deep learning model and a convolutional neural network (CNN)

[1]East China Electric Power Test and Research Institute, China Datang Corporation Science and Technology Research Institute Co., Ltd, Hefei 230022, China. [2]Datang Guoxin Binhai Offshore Wind Power Co., Ltd., Yancheng 224000, China. [3]School of Chemistry and Pharmaceutical Engineering, Changsha University of Science and Technology, Changsha 410114, China. [4]School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410114, China. ✉email: liych@csust.edu.cn

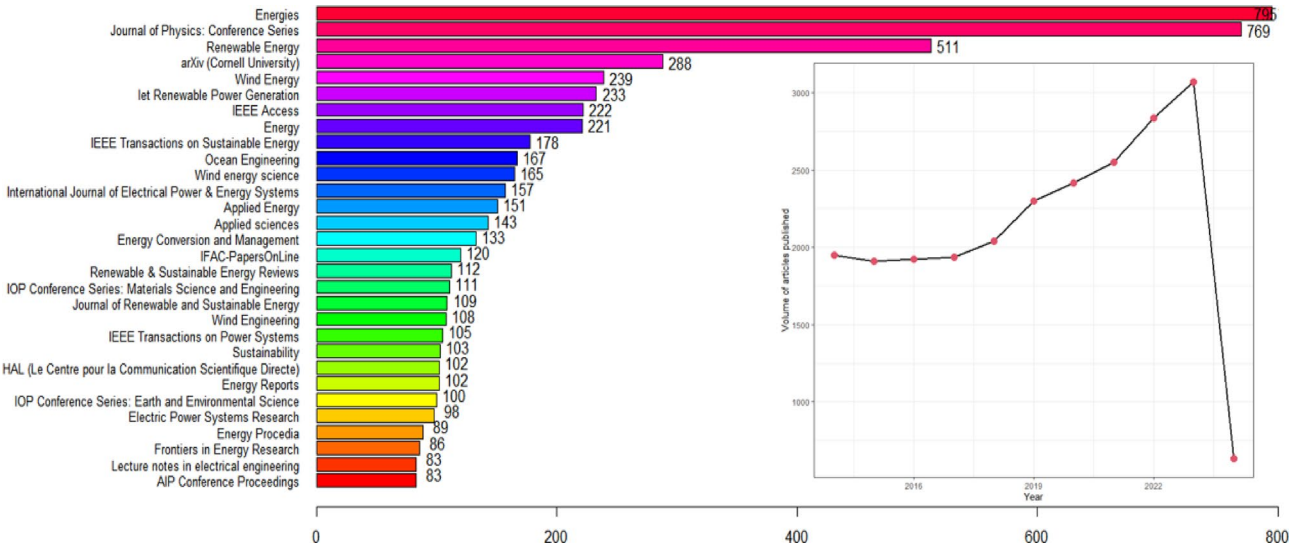


Fig. 1. Literature Information Chart with keyword of gear oil/wind turbines.

Type	Indicators	Memo
General data	Date, Wind.No., Oil_brand	2022.12 ~ 2024.6 datasets of gear oil originating from a wind farm
Chemical data	Viscosity (40℃), acid value, moisture	
Pollution data	PQ	
Metallic leaching elements	Ag, Al, Ba, Cd, Cr, Cu, Fe, K, Li, Mg, Mn, Mo, Na, Ni, Pb, Sb, Si, Sn, Ti, V	
Additive precipitating elements	P, Zn, B, Ca	

Table 1. Fundamental information of gear oil datasets.

can also be utilized for spatial feature extraction to capture the temporal patterns[19], and a hybrid approach combining deep learning and signal processing for bearing fault diagnosis was also explored under imbalanced samples and multiple operating conditions[20]. A robust hybrid model integrating Wavelet Coherence Analysis (WCA) with deep learning architectures VGG16 and ResNet50 was successfully implemented for accurate fault detection and classification in centrifugal pumps[21]. A multi-input CNN that simultaneously processes acoustic emission and vibration signals was employed for developing a model capable of detecting faults in a milling machine[22].

To tackle the challenges in wind turbine fault diagnosis using federated learning, this paper compares a SMR model and clustered federated learning framework (CFLF) for gear oil diagnosis of wind turbines. On this basis, a combined model of SMR and CFLF was proposed, which leverages the multi-scale residual to extract spatial features for gear oil diagnosis, experimental datasets are subjected to these modeling approaches for comparing the diagnostic performance of wind gear oil.

Data description and process

Data description

Table 1 presents datasets of gear oil for a wind farm in south China. In addition to the general categories of data there are four main categories of data, i.e. chemical data, pollution data, metallic leaching elements and additive precipitating elements. The viscosity (40℃), acid value, moisture and PQ value are four critical variables for gear wear diagnosis, metallic leaching elements and additive precipitating elements are two types of supporting diagnostic indicators. Additionally, the dependent variables, Results, is designated as variable 'Y', which reflects the quality of the gear oil.

Data preprocessing

There are notable variations in gear oil variables during wind gear oil degradation. After handling outliers and aggregating all gear oil data of wind turbines, Fig. 2 demonstrates considerable heterogeneity in the distribution of these variables across different wind turbines. Data heterogeneity is commonly observed in comprehensive fault analysis tasks involving multiple turbines. Failing to account for this heterogeneity when aggregating indicator models can lead to decreased diagnostic performance and slower convergence of the global model. In order to remove differences in numerical values and the influence of units among different variables, we need to apply a normalization function to scale each variable in the input dataset. The normalized variable is transformed according to the following function expression:

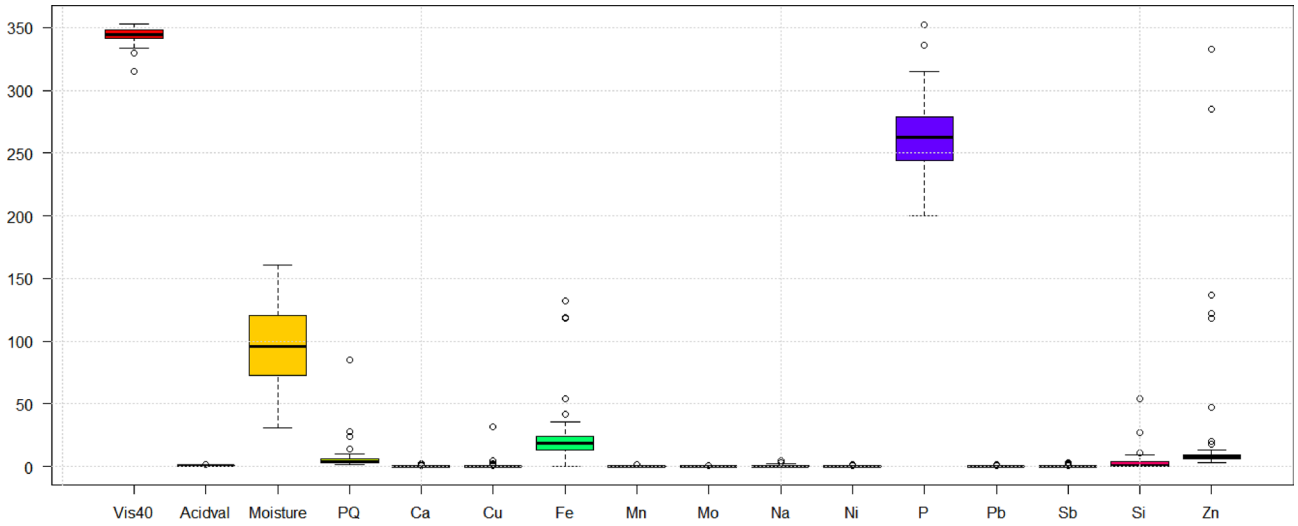


Fig. 2. Distribution scheme of gear oil variables across different wind turbines.

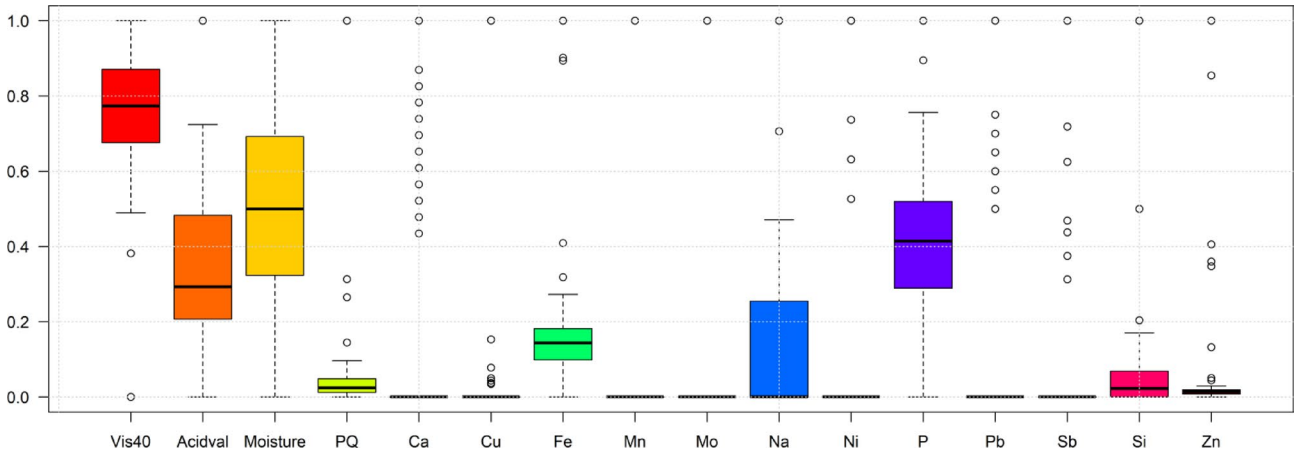


Fig. 3. Distribution scheme of gear oil variables after [0,1] normalization process.

$$X_{norm} = \frac{X - \min(\mathrm{X})}{\max(X) - \min(\mathrm{X})} \tag{1}$$

where *min(X)* and *max(X)* denote the minimum and maximum values of vector X. The normalization process maps the data uniformly onto the [0,1] interval for all variables, as shown in Fig. 3.

Methodology

SMR modeling

The R software (version 4.4.1) was used to build the SMR model[23]. Since the dependent variables (Y) were continuous in the study, a SMR model was used for analysis. In this multivariate model, a stepwise approach was used for higher accuracy[24,25]. Chi-square analysis was conducted to evaluate the relationship between the variables and the dependent variable in the estimated model. Finally, the model coefficients were reported, and the estimated model was expressed as a function[26–29]. The reference value was taken as $P \leqq 0.05$ to test the statistical significance. Table 2 indicates 11 stages of the SMR modeling process through continuous optimization of AIC value. In the first step of the modeling process, the dependent variable Y is linearly regressed on all the independent variables, and the AIC value of the model is obtained as −292.37. Based on the feedback from the model analysis, it can be seen that optimizing the independent variable Ni further reduces the AIC value to −294.37. This process of continuous optimization is repeated until the 11th step, where the AIC value reaches −309.76 and no longer decreases. This indicates that the model obtained in the 11th step has the best goodness-of-fit. Additionally, the Chi-square value of model is 23.44502 with a p-value far less than 0.0001. The optimal model function is shown in the following function expression:

Modeling step	Parameters and statistical significance
1 st step	Start: AIC=−292.37 Y ~ Vis40 + Acidval + Moisture + PQ + Cu + Fe + Mn + Mo + Na + Ni + Pb + Sb + Si + Ca + P + Zn
2nd step	Step 2: AIC=−294.37 Y ~ Vis40 + Acidval + Moisture + PQ + Cu + Fe + Mn + Mo + Na + Pb + Sb + Si + Ca + P + Zn
3rd step	Step 3: AIC=−296.37 Y ~ Vis40 + Acidval + Moisture + PQ + Fe + Mn + Mo + Na + Pb + Sb + Si + Ca + P + Zn
…	…
11th step	Step 11: AIC=−309.76 Y ~ Vis40 + PQ + Mn + Si + Ca + Zn Chisquare = 23.44502, Df = 1, p = 1.2853e−06 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 1.326375 0.144801 9.160 6.60e−14 *** Vis40 −0.275937 0.154620 −1.785 0.0783 PQ −0.838147 0.185524 −4.518 2.25e−05 *** Mn 0.604340 0.347726 1.738 0.0863 Si −0.796624 0.179905 −4.428 3.14e−05 *** Ca 0.097442 0.091887 1.060 0.2923 Zn −1.590522 0.605119 −2.628 0.0104 *

Table 2. SMR modeling process of gear oil datasets.

$$Y = 1.326 - 0.276\,\mathrm{Vis40} - 0.838\,\mathrm{PQ} + 0.604\,\mathrm{Mn} - 0.797\,\mathrm{Si} + 0.097\mathrm{Ca} - 1.591\,\mathrm{Zn} \quad (2)$$

where the PQ and Si indicators show the highest levels of significance in the model, with P-values of 2.25e−05 and 3.14e−05 respectively.

The diagnosis scheme of the above SMR model includes four parts i.e. residuals ~ fitted values plot, standardized residuals ~ quantiles plot, root standardized residuals ~ fitted values plot and standardized residuals ~ leverage plot. The residual ~ fitted plot shows that the dependent variable is linearly related to the independent variable, the residual values are not systematically related to the predicted (fitted) values; the Q-Q residuals plot shows that the assumption of normality is basically satisfied and the points on the plot fall on a straight line at a 45^0 angle; in the scale-location plot, the points around the horizontal line should be randomly distributed, and this plot seems to satisfy that assumption, which indicates that the model is a good fit. As shown in Fig. 4, the residuals ~ leverage plot provides information about the individual observations, and from the graph it is possible to identify outliers, high leverage points and strong influence points. The most typical anomalies are records 8,12 and 69.

CFLF modeling

Gear oil performance diagnosis is a nonlinear, multivariate comprehensive process. Therefore, it is necessary to introduce nonlinear clustering-based federated learning algorithms for comprehensive analysis and evaluation. Following the development of local multiscale residual networks for each wind turbine, we proposed a representational canonical correlation clustering method to group these local indicators into distinct clusters[30–36]. Initially, NbClust package in R language was used to determine the optimal number of clusters. This package defines dozens of evaluation metrics and evaluates cluster numbers from 2 to 9. As the number of clusters increases, the size of each cluster becomes smaller and more similar, causing the Dindex values to steadily decrease. When the slope of this decline flatterns, it indicates that further increasing the number of clusters does not improve clustering effectiveness. This inflection point, or "elbow point" is considered the optimal number of clusters. In this study, the Dindex values dropped sharply from 1 to 7 clusters and then more gradually afterward, suggesting that 7 is the optimal number, as shown in Fig. 5a, b and c demonstrates that the number of supporting indicators is maximized when the number of clusters is 7. The results of the above clustering can be visualized through the fviz_cluster function in the FactoMineR package, as shown in Fig. 6. Dim1 and Dim2 in the graph represent the percentage in the gear oil data set can be explained by Principal Component 1 and Principal Component 2 respectively. Each dot represents a row in the data frame, and the distance between the dots reflects the similarity. Figure 6 shows the relative concentration of data points in clusters with cluster numbers 4, 5 and 6. The optimized cluster analysis shows that record 8 is outlier, indicating that the gear oil is in a warning state and requires attention.

By employing spectral clustering, the graph is partitioned to effectively divide the indicators into separate clusters. Subsequently, the clustered federated learning model carries out local training tasks, which performs cluster-internal model aggregation. After determining the number of clusters for optimal clustering, it is necessary to create a data frame of the clustering results. On this basis, neural network modeling is performed for each cluster of data groups to obtain the model combination as shown in Fig. 7. In R language, using the neuralnet package for neural network regression prediction is a relatively direct process. This package provides a simple method for building and training neural network models, suitable for regression problems. The neuralnet() function in the neuralnet package was used to create a neural network model, and the number of neurons for hidden layer is set to 5 and a threshold of 0.01. The seven network diagrams in Fig. 7 correspond to the neural network model diagrams of the seven clustered data groups, and the correspondences are illustrated on red color in the middle-right side of the figure. From the comparison of model deviation data in the lower right corner of

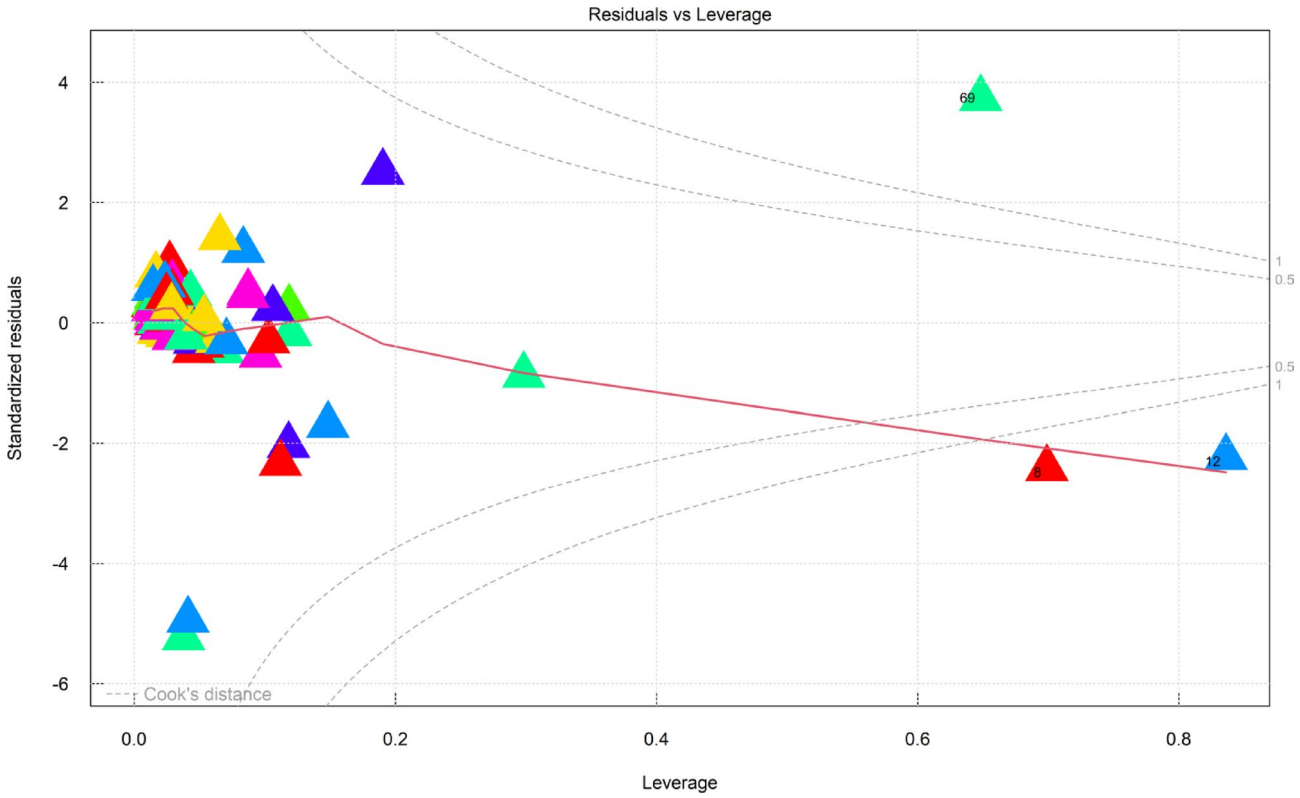


Fig. 4. Residual ~ leverage diagnosis scheme of the SMR model.

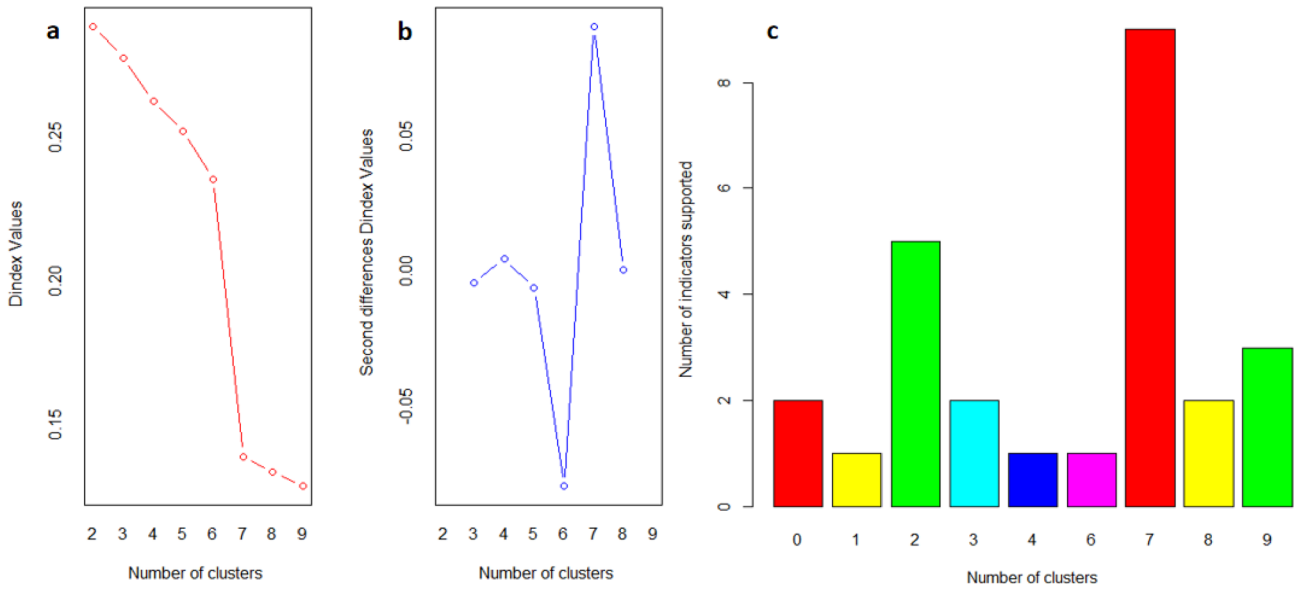


Fig. 5. Determination of best cluster number by NbClust package, (**a**) Dindex value with clusters; (**b**) Second differences Dindex values with clusters; (**c**) Number of indicators supported with clusters.

the figure, it can be seen that the deviation ranges from 0.000001 to 0.006, indicating that the accuracy of this CFLF modeling method is high.

Evaluation for the combined model of SMR & CFLF

The SMR model has advantages in predicting or explaining the quantitative effects between variables, but it also has shortcomings such as such as overfitting and weak non-linear modeling capabilities; CFLF belongs to unsupervised learning, which has the advantage of not requiring preset labels or target variables, and relying

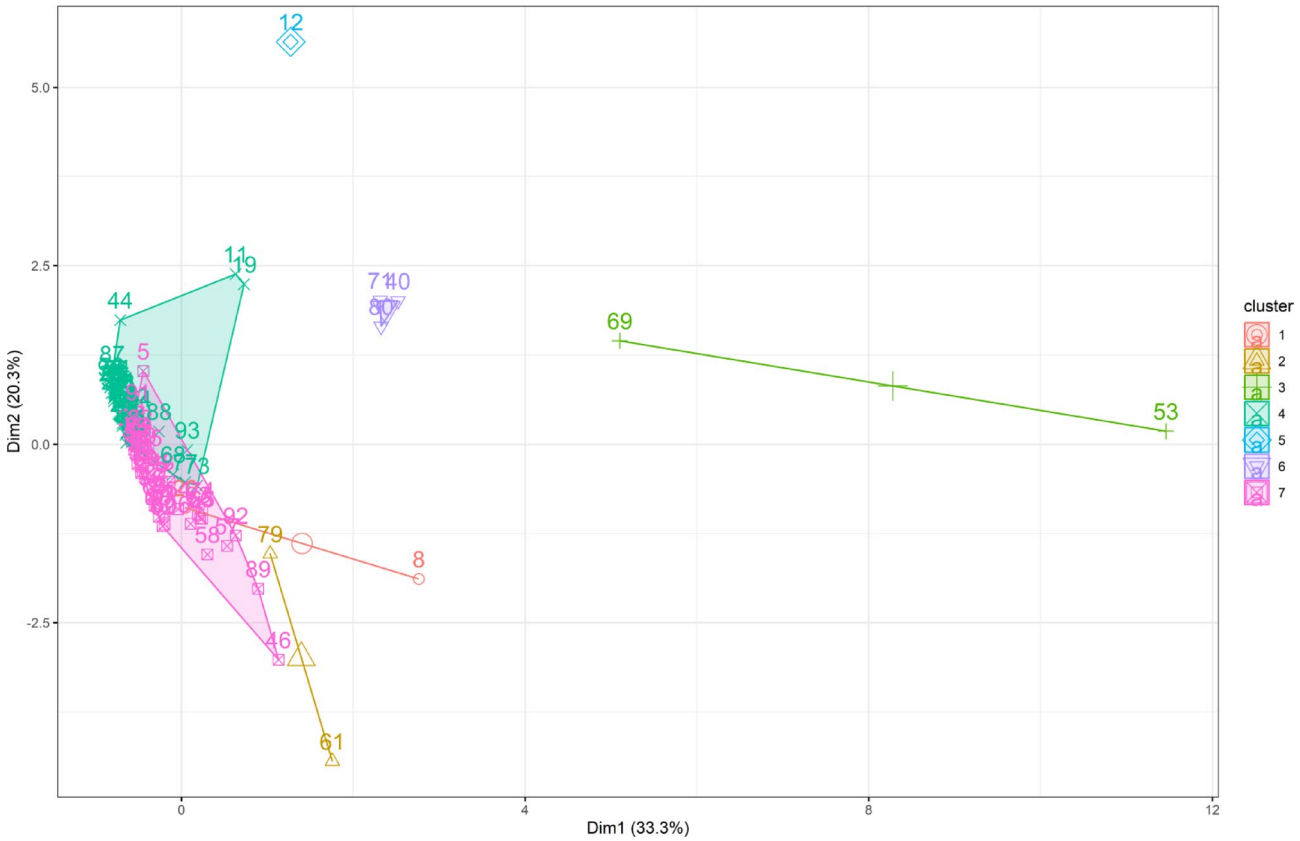


Fig. 6. Visualization of clustering results by fviz_cluster() function.

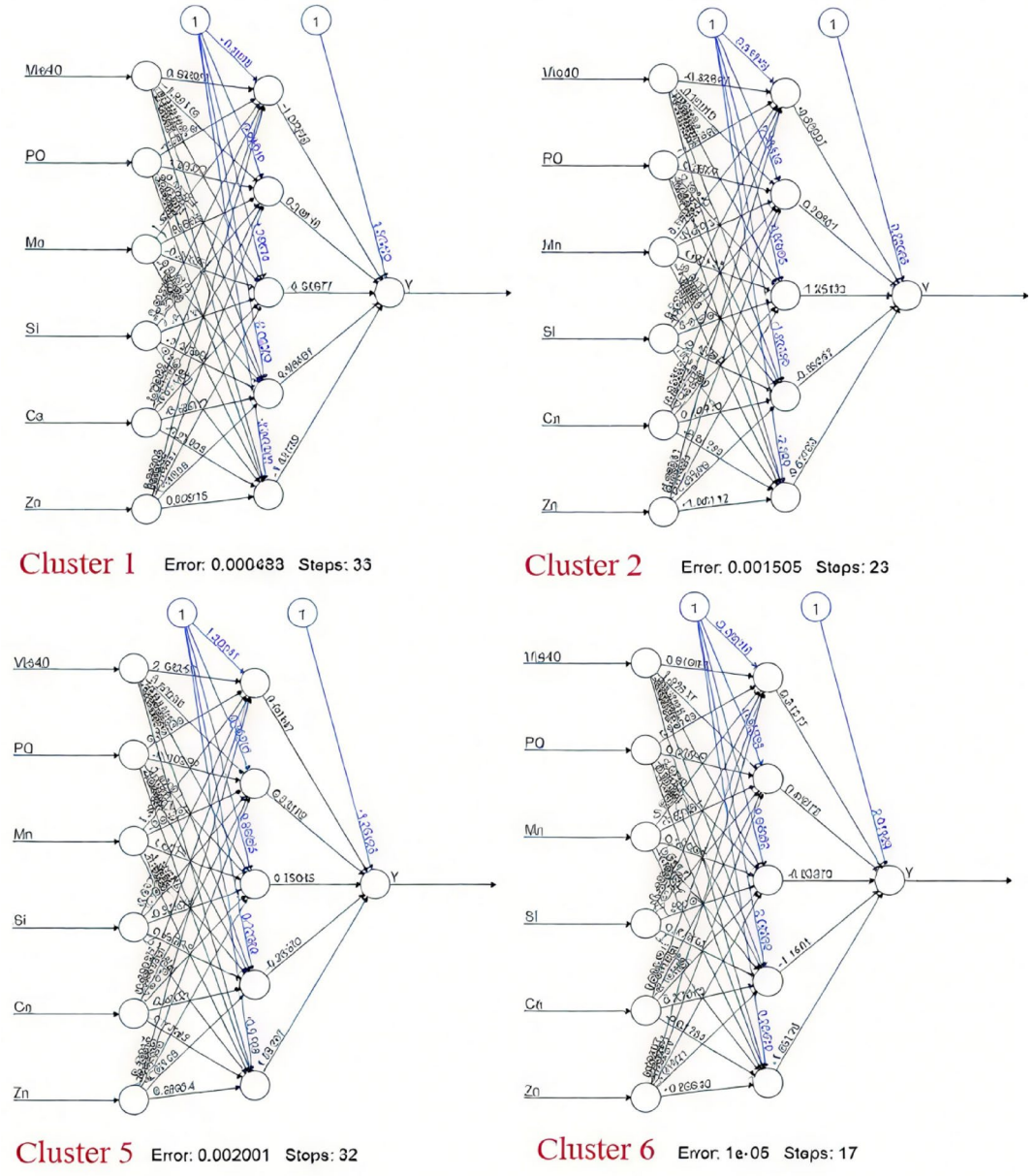
solely on similarity measures of feature variables (such as Euclidean distance). A disadvantage of CFLF is that it only outputs class labels or cluster assignments of samples. The combining model of SMR and CFLF can be proposed as SMR-CFLF composite model, which is based on the highly canonical correlation series obtained from SMR, afterwards CFLF modeling is subsequently carried out on this basis, which can take full advantage of both models to play a role in wind turbine gear oil metrics analysis. Gear oil performance dataset can be extracted through SMR modeling process, and canonical series include Vis40, PQ, Mn, Si, Ca and Zn.

By extracting 15% of the data volume for the prediction of the above SMR model, the obtained Y-values were analyzed for deviation from the actual Y-values, and the root mean squared error (RMSE) was 0.0708, demonstrating the relatively high accuracy of this model. The same validation data set was used for the prediction of the CFLF model and the root mean square error RMSE value obtained was 0.0729. However, based on the efficient quantitative analysis between variables and the precision of the Euclidean distance for feature variables, the combining SMR-CFLF model achieved an RMSE value of 0.0455 on the same validation data set predictive analysis, which fully demonstrated the advantages of combining SMR and CFLF. Figure 8 presents corresponding RMSE values among SMR, CFLF and combining SMR - CFLF model.

As shown in Fig. 8, for analysis of the wind turbine gear oil indicators data set, the CFLF model alone did not achieve a better RMSE value than the SMR model, instead, its deviation was − 2.97%. In contrast, the RMSE value of the combined SMR-CFLF model achieved a relatively significant improvement with a deviation ratio of 35.73%. The RMSE results of the SMR-CFLF model were 35.73% better than those of the SMR model. At present, the combined model has achieved better results in the prototype test, which promoted research on gear wear evaluation; if it is later applied to the practice of wind turbine gear oil performance prediction on a large scale, it is hoped that the performance analysis of the wind turbine gear state and decision-making for maintenance will be greatly improved.

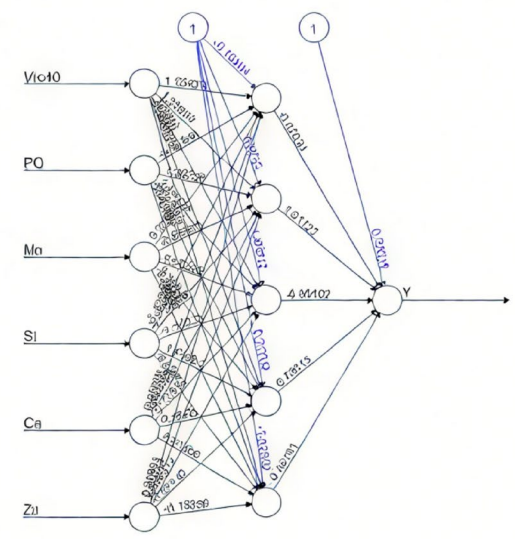
Conclusions

This study introduces a SMR modeling for gear oil diagnosis of wind turbines with a eleven-step optimization. By grouping oil indicators of a similar class into clusters, CFLF is introduced to address data heterogeneity issues and generate clustered models to assess gear oil performance, with the deviation ranges from 0.000001 to 0.006. A combined model of SMR and CFLF is proposed to assess the performance of gear oil. Actual data analysis of wind turbine gear oil showcase the superior performance of the proposed model over the single SMR model with higher prediction accuracy of 35.73%. The study provides significant promise in integrating SMR and CFLF into feature interaction learning for optimal diagnosis of wind gear oil.

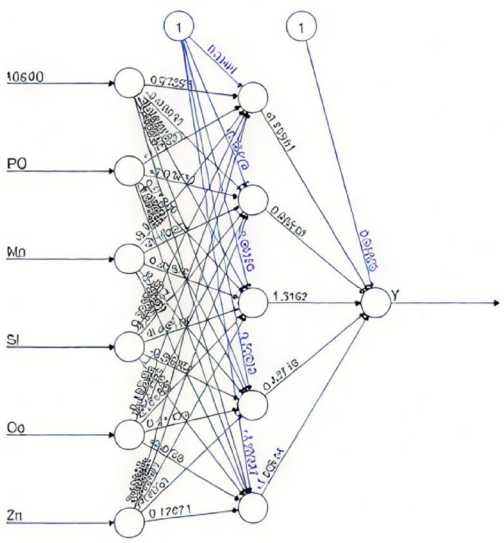


(a) Cluster 1, 2, 5 and 6

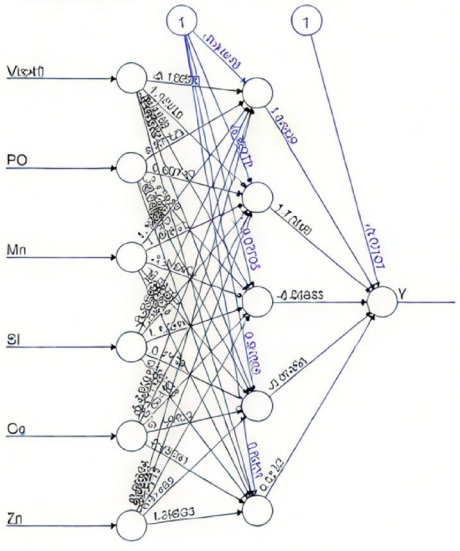
Fig. 7. Models of neural network learning after Clustering.



Cluster 3 Error: 0.005492 Etops: 23



Cluster 4 Error: 0.004112 Steps: 40



Cluster 7 Error: 0.005122 Stape: 41

No.	Error	Steps
Cluster 1	0.000489	33
Cluster 2	0.001505	23
Cluster 3	0.005492	23
Cluster 4	0.004112	40
Cluster 5	0.002001	32
Cluster 6	1e-05	17
Cluster 7	0.005122	41

(b) Cluster 3, 4, 7 and summary information

Fig. 7. (continued)

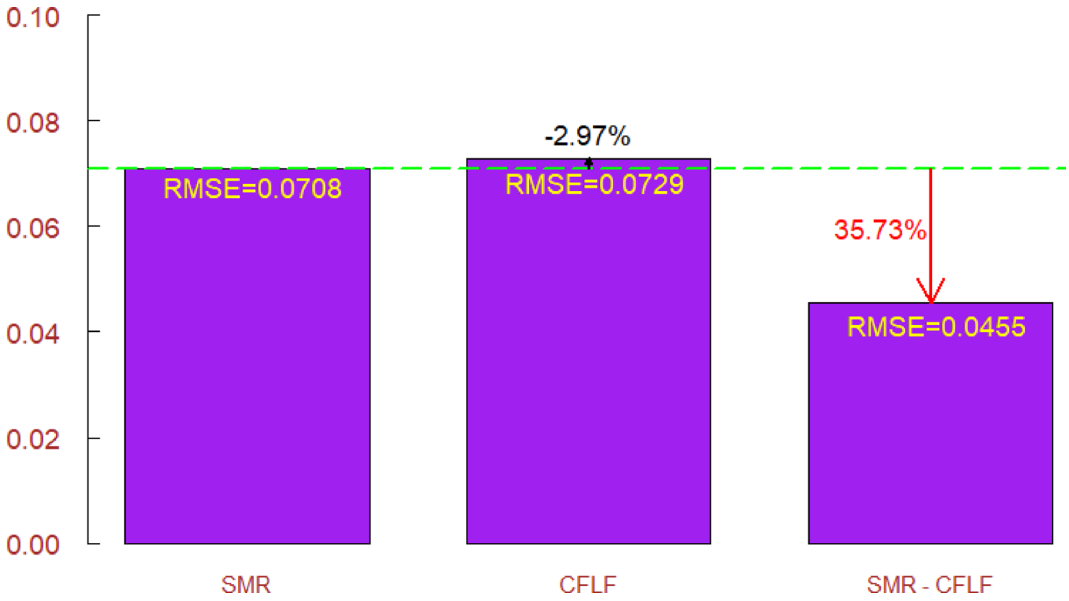


Fig. 8. RMSE comparison among SMR, CFLF and combining SMR - CFLF model.

Data availability

Data will be made available on request and can be contacted at corresponding author (Yuchun Li, liych@csust.edu.cn).

Received: 20 December 2024; Accepted: 11 June 2025
Published online: 02 July 2025

References

1. Tian, Z. & Wang, J. Variable frequency wind speed trend prediction system based on combined neural network and improved multi-objective optimization algorithm. *Energy* **254**, 124249. https://doi.org/10.1016/j.energy.2022.124249 (2022).
2. Liu, H. & Chen, C. Data processing strategies in wind energy forecasting models and applications: a comprehensive review. *Appl. Energy*. **249**, 392–408. https://doi.org/10.1016/j.apenergy.2019.04.188 (2022).
3. Meng, A., Ge, J., Yin, H. & Chen, S. Wind speed forecasting based on wavelet packet decomposition and artificial neural networks trained by crisscross optimization algorithm. *Energy Convers. Manag.* **114**, 75–88. https://doi.org/10.1016/j.enconman.2016.02.013 (2016).
4. Li, Z. et al. Friction-reducing and anti-wear performance of SiO_2-Coated TiN nanoparticles in gear oil. *Wear Volumes* 538–539, 15 February 2024, 205219 https://doi.org/10.1016/j.wear.2023.205219
5. Bibliometric analysis was. performed using the citexs website (https://www.citexs.com/).
6. Ayvat, P. & Omeroglu, S. Mortality Estimation using APACHE and CT scores with Stepwise linear regression method in COVID-19 intensive care unit: A retrospective study. *Clin. Imaging Volume.* **88**, 4:8. https://doi.org/10.1016/j.clinimag.2022.04.017 (August 2022).
7. Ozmen, A. Multi-objective regression modeling for natural gas prediction with ridge regression and CMARS. *Int. J. Optim. Control Theor. Appl. (IJOCTA).* **12** (1), 56–65. https://doi.org/10.11121/ijocta.2022.1084 (2022).
8. Du, X. et al. Regression analysis and prediction of monthly wind and solar power generation in China. *Energy Rep.* **12**, 1385–1402. https://doi.org/10.1016/j.egyr.2024.07.027 (2024).
9. Du, X. et al. Forecasting models for economically driven provincial gas loads in China. *Oil Gas Storage Transp.* **42**, 1184–1192. https://doi.org/10.2991/978-94-6463-042-8_169 (2023).
10. Liu, Z. et al. Clinical significance of serum lactate dehydrogenase combined with a multivariate model for predicting the near-term outcome of primary nasopharyngeal carcinoma. *Life Sci. Volume.* **351**, 122856. https://doi.org/10.1016/j.lfs.2024.122856 (August 2024).
11. Fan, H. J. et al. Fluctuation pattern recognition based ultra-short-term wind power probabilistic forecasting method. *Energy* **266** https://doi.org/10.1016/j.energy.2022.126420 (2023).
12. Ullah, N. et al. Enhanced fault diagnosis in milling machines using CWT image augmentation and ant colony optimized AlexNet. *Sensors* **24**, 7466. https://doi.org/10.3390/s24237466 (2024).
13. Saleem, F. et al. Acoustic Emission-Based pipeline leak detection and size identification using a customized One-Dimensional densenet. *Sensors* **25**, 1112. https://doi.org/10.3390/s25041112 (2025).
14. Zhou, W. et al. A shale gas production prediction model based on masked convolutional neural network. *Appl. Energy*. **353** https://doi.org/10.1016/j.apenergy.2023.122092 (2024).
15. ElRobrini, S. et al. Federated learning and non-federated learning based power forecasting of photovoltaic/wind power energy systems: A systematic review. Energy and AI. December 18, 100438. (2024). https://doi.org/10.1016/j.egyai.2024.100438
16. Chen, Y. et al. Error revision during morning period for deep learning and multi-variable historical data-based day-ahead solar irradiance forecast: towards a more accurate daytime forecast. *Earth Sci. Inf. 1865 – 0473*, **16** 2261–2283, https://doi.org/10.1007/s12145-023-01026-3. (2023).
17. Helbing, G. & Ritter, M. Deep learning for fault detection in wind turbines. *Renew. Sustain. Energy Rev.* **98**, 189–198 (2018).
18. Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol.* **10**(2), 12. https://doi.org/10.1145/3298981 (2019).
19. Umar, M. et al. Milling machine fault diagnosis using acoustic emission and hybrid deep learning with feature optimization. *Appl. Sci.* **14**, 10404. https://doi.org/10.3390/app142210404 (2024).

20. Zhang, B., Wang, W. & He, Y. A hybrid approach combining deep learning and signal processing for bearing fault diagnosis under imbalanced samples and multiple operating conditions. *Sci. Rep.* **15** (1). https://doi.org/10.1038/s41598-025-98138-1 (2025).
21. Zaman, W. et al. Hybrid deep learning model for fault diagnosis in centrifugal pumps: A comparative study of VGG16, ResNet50, and wavelet coherence analysis. *MACHINES* **12** (12), 905. https://doi.org/10.3390/machines12120905 (2024).
22. Zaman, W. et al. A new dual-input CNN for multimodal fault classification using acoustic emission and vibration signals. *Eng. Fail. Anal.* https://doi.org/10.1016/j.engfailanal.2025.109787 (2025).
23. R Core Team. _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. (2024). https://www.R-project.org/
24. Shahid, F., Zameer, A. & Muneeb, M. A novel genetic LSTM model for wind power forecast. *Energy* **223**, 120069. https://doi.org/10.1016/j.energy.2021.120069 (2021).
25. Barjasteh, A., Ghafouri, S. H. & Hashemi, M. A hybrid model based on discrete wavelet transform (DWT) and bidirectional recurrent neural networks for wind speed prediction. *Eng. Appl. Artif. Intell.* **127**, 107340. https://doi.org/10.1016/j.engappai.2023.107340 (2024).
26. Yu, R. et al. LSTM-EFG for wind power forecasting based on sequential correlation features. *Future Generation Comput. Syst.* **93**, 33–42. https://doi.org/10.1016/j.future.2018.09.054 (2019).
27. Liu, H., Mi, X. & Li, Y. Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network. *Energy. Conv. Manag.* **156**, 498–514 (2018).
28. Liu, Z., Jiang, P., Zhang, L. & Niu, X. A combined forecasting model for time series: application to short-term wind speed forecasting. *Appl. Energy*. **259**, 114137 (2020).
29. Meilisa, M. & Otok, B. Purnomo. Estimation curve of multivariate adaptive biresponse fuzzy clustering means regression splines approach to stunting and wasting cases. *Southeast. Sulawesi MethodsX*. **12**, 102775. https://doi.org/10.1016/j.mex.2024.102775 (2024).
30. Diani, C., Galimberti, G. & Soffritti, G. Multivariate cluster-weighted models based on seemingly unrelated linear regression. *Comput. Stat. Data Anal.* **171**, 107451. https://doi.org/10.1016/j.csda.2022.107451 (2022).
31. Vats, F. Basu. A novel cluster-based framework for developing correlation model and its implementation for spectral acceleration. *Soil Dyn. Earthq. Eng.* **188A**, 109056. https://doi.org/10.1016/j.soildyn.2024.109056 (2025).
32. Shen, J., Liu, Q. & Feng, X. Hourly PM2.5 concentration prediction for dry bulk Port clusters considering Spatiotemporal correlation: A novel deep learning blending ensemble model. *J. Environ. Manage.* **370**, 122703. https://doi.org/10.1016/j.jenvman.2024.122703 (2024).
33. Du, R. et al. 3DTCN-CBAM-LSTM short-term power multi-step prediction model for offshore wind power based on data space and multi-field cluster spatio-temporal correlation. *Appl. Energy* **376A**, 124169, https://doi.org/10.1016/j.apenergy.2024.124169. (2024).
34. Lu, Z. et al. RSC-based differential model with correlation removal for improving multi-omics clustering. *J. Theor. Biol.* **556**, 111328. https://doi.org/10.1016/j.jtbi.2022.111328 (2023).
35. Yang, H. et al. K-PCD: A new clustering algorithm for Building energy consumption time series analysis and predicting model accuracy improvement. *Appl. Energy*. **377C**, 124584. https://doi.org/10.1016/j.apenergy.2024.124584 (2025).
36. Ding, Z. et al. Clustering driven incremental learning surrogate model-assisted evolution for structural condition assessment. *Mech. Syst. Signal Process.* **224**, 112146. https://doi.org/10.1016/j.ymssp.2024.112146 (2025).

Author contributions

Huihui Han, Ye Zhao and Yuchun Li wrote the main manuscript text and Hao Jiang, Muxin Chen prepared Figs. 1, 2 and 3. Song Zhou, Zihan Lin developed R scripts and Xin Wang, Boman Mao, Xinyue Yang prepared other figures. Yuchun Li reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.