Applied Artificial Intelligence Research

INNOVIAIR

Research on Performance Prediction Model of Wind Turbine Gearbox Lubricating Oil Based on Deep Learning

Tongwei Xie[a], Xikun Zhang[b], Xiaodong Liu[c], Xuan Zhang[d,*]

[a]*Zhengzhou University of Technology, Zhengzhou, Henan 450044, China*

[b]*Tianjin University,Tianjin 300072, China*

[c]*Ningxia Baofeng Energy Group Co., Ltd., Yinchuan, Ningxia 750001, China*

[d]*Tsinghua University, Beijing 100000, China*

ARTICLEINFO

Keywords:
Gearbox Lubricating Oil
Antioxidant Performance Prediction
Data Preprocessing
Feature Extraction
Machine Learning
SNV-PCA-BP Model
Rapid Detection

ABSTRACT

Gearbox lubricating oil oxidation degradation severely impairs the operational stability of wind turbines and increases maintenance costs. Conventional detection methods (e.g., Rotating Pressure Vessel Oxidation Test (RPVOT), Pressure Differential Scanning Calorimetry (PDSC)) have high instrument dependence and long testing cycles, and may not meet on-site rapid detection demands. This study took in-service industrial gear oil samples as the research object, selected five physicochemical indices as input variables, and systematically optimizes data preprocessing (Standard Normal Variate (SNV)), feature extraction (Principal Component Analysis (PCA)), and machine learning algorithms (Back Propagation Neural Network (BPNN)/Support Vector Machine (SVM)/Random Forest (RF)). The proposed SNV-PCA-BP hybrid model achieved excellent predictive performance with a R2 of 0.9960 and Residual Prediction Deviation (RPD) of 6.1247, which is 360 times more efficient (defined as single-sample detection time) than traditional RPVOT/PDSC methods based on parallel tests of 62 samples. This model provides a low-cost and reliable technical support for the predictive maintenance of wind turbine gearboxes.

1. Introduction

As a critical component of engines, the gear relies on effective lubrication to ensure stable operation, which can impact the operational efficiency and maintenance costs of the overall mechanical system. However, the gear oil is susceptible to oxidative degradation induced by multiple factors, including high temperature, high pressure, oxygen exposure, and metal catalysis. This can result in viscosity anomalies and increasing acid level, ultimately leading to equipment failures such as gear wear and bearing malfunction. Therefore, accurately assessing the decay patterns of gear oil antioxidant performance is important for optimizing oil change intervals and mitigating maintenance risks.

Conventional methods for assessing gear oil antioxidant performance, such as the Rotating Pressure Vessel Oxidation Test (RPVOT) and Pressure Differential Scanning Calorimetry (PDSC), require specialized instrumentation and involve lengthy detection cycles up to days for processing[1]. When applied with complex operational procedures, these methods are unable to meet the practical demands for rapid condition assessment. Recent research showed a correlation between lubricant physicochemical properties and antioxidant performance, which allowed machine learning models to study the physicochemical properties and find trend to predict future decay of antioxidant performance in lubricant oil[2,3]. However, existing literature predominantly focused on engine oils, while specialized research concerning gear oils remains relatively scarce[4]. Notably, engine oils are formulated to accommodate high-temperature, low-load engine operations, while gearbox oils (i.e., gear oils) are designed to endure heavy loads, shock, and scuffing in gear transmission systems, resulting in distinct performance requirements and physicochemical characteristics. Although some studies have attempted to establish predictive models, they are frequently compromised by issues such as high complexity and insufficient accuracy, suggesting a need to improve for practical engineering requirements[5].

* Corresponding author.
E-mail address: zws1127@126.com.
https://doi.org/10.65455/2s6npy06
Received 18 January 2026; Received in revised form 26 January 2026; Accepted 28 January 2026; Available online 20 March 2026

To address these limitations, this study investigated in-service gear oil from Olefin Plant No. 1 of Ningxia Baofeng Energy Group (China). Five categories of indices—particle analysis, physicochemical properties, elemental analysis, wear indices, and contamination indices—are selected as input variables for preprocessing. This paper then compared various data preprocessing methods, feature extraction algorithms, and machine learning models. Herein, an accurate and robust model for the rapid prediction of gear oil antioxidant performance, demonstrating a firm basis with high practical significance for smart lubricant condition monitoring and maintenance. The five selected indices are based on preliminary experimental data of 62 samples, with Pearson correlation coefficients of 0.48-0.89 with oxidation value and no severe multicollinearity (VIF<3), and detailed screening data are available from the corresponding author.

2.Materials and methods

2.1.Experimental materials

The gear oil samples used in this study were obtained from critical transmission equipment—including compressors, extruders, and reactors—currently in service at the Olefin Plant No. 1 of Ningxia Baofeng Energy Group Co., Ltd. (China). The specific application scenarios encompassed polypropylene (PP) reactor agitators and the main reduction gearboxes of both PP and polyethylene (PE) extruders. The gear oil samples, obtained from industrial equipment under actual operating conditions, consisted of various industrial gear oils (Shell Omala S2 G series) spanning ISO viscosity grades (VG) 150–320(including partial extended grade VG 460 for individual samples). The physicochemical properties across different oil types were sufficient for the sampling requirements for antioxidant performance prediction. The actual service duration of the oils was approximately two years. The monitoring period extended from November 2024 to July 2025, yielding a total of 62 valid samples. All collected data exhibited high continuity and integrity, comprehensively reflecting the performance characteristics of gear oil throughout actual service.

2.2.Detection indices and methodology

To comprehensively characterize the performance parameters of the gear oil, five core indices were systematically evaluated across all 62 samples in strict adherence to international standards and industry specifications, thereby ensuring the accuracy and reliability of the experimental data. Specifically, wear particle analysis was conducted via analytical ferrography (SH/T 0573), utilizing a ferrograph microscope for qualitative grading and a PQ indexer for quantification[6]. Elemental concentrations in both in-service and fresh lubricants were determined using Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES) following ASTM D5185[7]. The contamination index was quantified utilizing a light-extinction Automatic Particle Counter (APC) in accordance with DL/T 432, with calibration executed per ISO 11171[8]. Furthermore, the wear index (PQ Index), reflecting ferrous debris content in in-service fluids, was measured using a Particle Quantifier in compliance with ASTM D8184[9]. Key physicochemical indices—including kinematic viscosity (at 40°C and 100°C), closed cup flash point, and foaming characteristics—were determined following Chinese National Standards GB/T 265, GB/T 3535, and GB/T 12579, respectively[10].

The oxidation value, serving as the primary metric for evaluating gear oil antioxidant performance, was quantitatively analyzed using the peak area method via Fourier Transform Infrared Spectroscopy (FTIR) in accordance with ASTM E2412-04[11]. Due to data availability, the oxidation stability was used in substitution of antioxidant content, which was derived via the Rotating Pressure Vessel Oxidation Test (RPVOT), same as in[12]. Specifically, the acid value (as a key physicochemical index) and oxidation value (as the core antioxidant performance metric) showed a strong correlation (r=0.89), while the oxidation stability (substitute for antioxidant content) and oxidation value also exhibited a significant correlation (r=0.87), which collectively confirmed the validity of this substitution. To prevent model bias, these specific samples were uniformly distributed into calibration and prediction sets using the Sample set Partitioning based on joint X-Y distances (SPXY) algorithm[13]. To minimize experimental error, all measurements were repeated three times, with the arithmetic mean recorded as the final result.

2.3.Data processing and modeling

The computational framework used in this study consisted of four critical stages: sample set partitioning, data preprocessing, feature extraction, and machine learning modeling. All computational procedures were executed using MATLAB 2023b.

2.3.1.Sample set partitioning

Sample set partitioning was based on joint X-Y distances (SPXY) algorithm to ensure robust model training efficacy and generalization capability. The SPXY algorithm offered a distinct advantage by simultaneously accounting for the distribution characteristics of both the input variables (five physicochemical indices) and the response variable (oxidation value), effectively mitigating model bias caused by uneven sample distribution[13]. Based on a 7:3 ratio, the dataset was separated into a calibration set comprising 43 samples and a prediction set comprising 19 samples. The calibration set was used for model building and training, while the prediction set served to validate the model performance, ensuring representativeness and distributional consistency across both subsets.

2.3.2.Data preprocessing

Raw experimental data were often compromised by instrument and system errors, environmental interference, and dimensional heterogeneity among indices[14]. To optimize modeling data quality, five standard preprocessing techniques were applied to the raw dataset (Raw): Multiplicative Scatter Correction (MSC) was employed to eliminate interference from particle scattering and baseline drift; SNV transformation was utilized to reduce data heteroscedasticity

through normalization; Savitzky-Golay (SG) smoothing was applied to effectively suppress random noise while preserving intrinsic data trends; First Derivative (FD) was used to enhance spectral resolution and amplify disparities between samples; and Second Derivative (SD) was implemented to further improve feature peak separation and data discrimination[15]. To identify the optimal preprocessing protocol, the raw data and the datasets processed by these five methods were individually combined with Partial Least Squares (PLS) regression models[16], and derive the optimal method.

2.3.3.Feature extraction

To effectively reduce data dimensionality and mitigate the impact of information redundancy on model performance, two classical feature extraction methodologies were used to optimize the preprocessed dataset. Firstly, the Principal Component Analysis (PCA) was used to project high-dimensional physicochemical index data onto a low-dimensional subspace through linear transformation[17]. By retaining the principal components with the highest variance contributions, PCA achieves significant dimensionality reduction while maximizing the preservation of critical information, thus simplifying the analytical complexity. Secondly, the Successive Projections Algorithm (SPA) was applied to select variables characterized by low correlation and complementary information, effectively attenuating the adverse effects of multicollinearity[18]. By using Multiple Linear Regression (MLR) models for candidate subsets and evaluating the Root Mean Square Error (RMSE), the subset yielding the minimum RMSE was identified as the optimal feature set, providing high-quality input variables for the subsequent machine learning phase[19].

2.3.4.Machine learning methods

To establish a robust predictive framework for gear oil antioxidant performance and identify the optimal strategy, this study used three classical machine learning algorithms renowned for their wide applicabilities: Back Propagation Neural Network (BPNN), Support Vector Machine (SVM), and Random Forest (RF)[20]. All model hyperparameters were optimized via the Grid Search method[21]. The BPNN model utilized a standard three-layer feedforward architecture, where the number of input nodes corresponds to the post-extraction feature variables, and the single output node represents the oxidation value[22]. The number of hidden layer nodes was determined to be 10 based on minimizing training error through 5-fold cross-validation; strictly defined training parameters included a maximum iteration count of 1000, a learning rate of 0.01, and a convergence threshold of 10-5 (details in Table 1).

Table 1. Hyperparameter settings for the BPNN model.

Parameter	Value
Number of hidden layer nodes	10
Maximum iterations	1000
Learning rate	0.01
Convergence threshold	10^{-5}

Table 2. Hyperparameter settings for the SVM model.

Parameter	Setting
Kernel function	Radial Basis Function (RBF)
Penalty coefficient (C)	100
Kernel coefficient (γ)	0.002

Table 3. Hyperparameter settings for the RF model.

Parameter	Value
Number of trees	200
Maximum tree depth	80
Feature sampling ratio	0.5

For the SVM regression that searches for an optimal hyperplane to minimize the deviation between training data and the fitting function, the Radial Basis Function (RBF) kernel was selected for its robust generalization properties[23]. Following grid search optimization (range: penalty coefficient $C \in [1,1000]$, kernel coefficient $\gamma \in [0.001,0.1]$, the optimal parameters were established as $C=100$ and $\gamma=0.002$ (details in Table 2). The Random Forest algorithm, with Bootstrap resampling to build multiple independent decision trees, led to the final prediction from the ensemble mean of the constituent trees[24]. The core parameters, comprising the number of trees, maximum depth, and feature sampling ratio, were optimized to 200, 80, and 0.5, respectively (details in Table 3). All model training and validation procedures were executed in MATLAB 2023b; technical documentation, including core functions, custom scripts, and hyperparameter optimization logs, is available from the corresponding author upon reasonable request.

2.3.5.Implementation of core modeling code

All modeling procedures in this study were executed within the MATLAB 2023b environment. The core code implementation primarily targets the end-to-end realization of the optimal combination model (SNV-PCA-BP), while the computational logic for comparative models (SVM, RF) and SPA feature extraction can be replicated through the modular substitution of the respective core functions. The implementation strictly adheres to a standard machine learning pipeline comprising data standardization, feature dimensionality reduction, model training, and performance validation. Representative segments of the core code are presented below:

Algorithm 1. Core implementation code of the SNV-PCA-BP model

```
%% Data Import and SPXY Partitioning
data = xlsread('GearOil_Oxidation_Dataset.xlsx');
X = data(:,1:5);
Y = data(:,6);
[cal_idx, pre_idx] = spxy(X, Y, 0.7);
X_cal = X(cal_idx,:); Y_cal = Y(cal_idx);
X_pre = X(pre_idx,:); Y_pre = Y(pre_idx);
%% SNV Preprocessing
X_cal_snv = snv_transform(X_cal);
X_pre_snv = snv_transform(X_pre);
%% PCA Feature Extraction
[coeff, score, ~, ~, explained] = pca(X_cal_snv);
cum_explained = cumsum(explained);
pc_num = find(cum_explained >= 97.68, 1);
X_cal_pca = score(:,1:pc_num);
```

```
    X_pre_pca = (X_pre_snv - mean(X_cal_snv)) *
coeff(:,1:pc_num);
    %% BP Neural Network Modeling
    net = feedforwardnet(10);
    net.trainParam.epochs = 1000;
    net.trainParam.lr = 0.01;
    net.trainParam.goal = 1e-5;
    net.trainFcn = 'trainlm';
    net.divideFcn = '';
    net = train(net, X_cal_pca', Y_cal');
    %% Prediction and Performance Evaluation
    Y_cal_pred = net(X_cal_pca')';
    Y_pre_pred = net(X_pre_pca')';
    cal_metrics = model_evaluation(Y_cal, Y_cal_pred);
    pre_metrics = model_evaluation(Y_pre, Y_pre_pred);
    fprintf('Calibration Set: R^2=%.4f, RMSE=%.4f,
MAE=%.4f, RPD=%.4f\n',...
    cal_metrics(1), cal_metrics(2), cal_metrics(3),
cal_metrics(4));
    fprintf('Prediction Set: R^2=%.4f, RMSE=%.4f,
MAE=%.4f, RPD=%.4f\n',...
    pre_metrics(1), pre_metrics(2), pre_metrics(3),
pre_metrics(4));
    %% Plotting
    figure;
    subplot(1,2,1);
    plot(Y_cal, Y_cal_pred, 'bo'); hold on;
    plot(Y_cal, Y_cal, 'r-');
    xlabel('Measured'); ylabel('Predicted');
    title('Calibration Set');
    legend('Predicted','Ideal Fit','Location','best');
    grid on;
    subplot(1,2,2);
    plot(Y_pre, Y_pre_pred, 'go'); hold on;
    plot(Y_pre, Y_pre, 'r-');
    xlabel('Measured'); ylabel('Predicted');
    title('Prediction Set');
    legend('Predicted','Ideal Fit','Location','best');
    grid on;
    %% Local Functions
    function [cal_idx, pre_idx] = spxy(X, Y, ratio)
    [n, ~] = size(X);
    Dx = pdist(X); Dx = squareform(Dx);
    Dy = pdist(Y); Dy = squareform(Dy);
    D = Dx/max(Dx(:)) + Dy/max(Dy(:));
    [~, idx] = min(D(:));
    [r, c] = ind2sub(size(D), idx);
    cal_idx = [r, c]; pre_idx = [];
    while length(cal_idx) < round(n*ratio)
    d_cal = min(D(cal_idx, :), [], 1);
    d_pre = min(D(pre_idx, :), [], 1);
    d_pre(isnan(d_pre)) = 0;
    [~, new_idx] = max(d_cal - d_pre);
    cal_idx = [cal_idx, new_idx];
    pre_idx = setdiff(1:n, cal_idx);
    end
    end
    function X_snv = snv_transform(X)
    [n, p] = size(X);
    X_snv = zeros(n, p);
    for i = 1:n
    mu = mean(X(i,:));
    sigma = std(X(i,:));
    X_snv(i,:) = (X(i,:) - mu) / sigma;
    end
    end
    function [metrics] = model_evaluation(Y_true, Y_pred)
    SSE = sum((Y_true - Y_pred).^2);
    SST = sum((Y_true - mean(Y_true)).^2);
    R2 = 1 - SSE/SST;
    RMSE = sqrt(mean((Y_true - Y_pred).^2));
    MAE = mean(abs(Y_true - Y_pred));
    RPD = std(Y_true) / RMSE;
    metrics = [R2, RMSE, MAE, RPD];
    end
```

2.4.Model evaluation

To objectively assess the training efficacy, predictive accuracy, and generalization capability of the proposed models, and to identify the optimal practical configuration, four internationally recognized metrics were used as comprehensive evaluation parameters: the Coefficient of Determination R2, RMSE, Mean Absolute Error (MAE), and Residual Prediction Deviation (RPD)[25]. Specifically, R2 quantifies the degree of linear correlation between predicted values and actual measurements, where a value approaching unity indicates the goodness-of-fit. RMSE, calculated as the square root of the mean squared errors, penalizes larger errors, and effectively reflects prediction uncertainties; meanwhile, MAE represents the arithmetic mean of absolute deviations, providing an intuitive measurement of average prediction bias. Lower values for both RMSE and MAE represent enhanced prediction accuracy and model stability. RPD, calculated from the ratio of the standard deviation of the prediction set to the RMSE, serves as a primary indicator of practical predictive utility. In accordance with widely accepted industry standards, an RPD > 2.5 indicates excellent predictive performance, a value between 1.5 and 2.5 suggests suitability for quantitative analysis, while an RPD < 1.5 implies the model lacks sufficient predictive capability[26]. All evaluation metrics were computed using MATLAB 2023b, with calculation procedures strictly adhering to relevant international standards to ensure the reliability and comparability of the results.

3.Results

3.1.Correlation analysis of detection indices

To analyze the relationship between various detection indices and the antioxidant performance of gear oil, a quantitative analysis was conducted using Pearson correlation coefficients. As illustrated in the correlation heatmap (Figure 1), there were variations in the degree of association between different indices and the oxidation values. The Acid Number exhibited the most robust correlation (r=0.89). This finding is consistent with the chemical mechanism of oxidative degradation, where the oxidation process generates carboxylic acid byproducts, leading to an increase in acid number

proportional to the degree of oxidation[27]. Kinematic viscosity at 40°C (r=0.73) and water content (r=0.65) also demonstrated significant positive correlations; the former was attributed to the formation of high-molecular-weight polymeric products during oxidation, while the latter reflected the role of water in accelerating the oxidation reaction. Conversely, the correlation between kinematic viscosity at 100°C and the oxidation value was relatively weaker (r=0.61), which could be attributed to the reduced sensitivity of viscosity to oxidation products at elevated temperatures. Flash point displayed the lowest correlation (r=0.48), indicating that it was less influenced by the oxidation. The correlation analysis provided an important reference for feature variable selection, justifying the selection of core variables for model building[28].Based on the degradation mechanism, the indices are categorized as follows:

I. Direct correlation indices

High-Temperature Oxidation Particles (Wear Particle Analysis): These particles serve as direct indicators of oxidation product formation under thermal stress.

Black Iron Oxide (Wear Particle Analysis): Identified as Fe3O4, this is a product of oxidative corrosion on metal surfaces. Its presence suggests surface oxidation of machinery components.

Sludge (Wear Particle Analysis): Sludge and carbonaceous deposits are prototypical byproducts of lubricant oxidation. Their deposition within the oil can significantly reduce lubricant's quality.

II. Indirect correlation indices

Foaming Characteristics (Physicochemical Indices): Deterioration in high-temperature foam is a critical precursor to oxidative degradation, as oxidation byproducts tend to stabilize foam, rendering it difficult to dissipate.

Kinematic Viscosity at 40°C and 100°C (Physicochemical Indices): Continuous monitoring of viscosity value is paramount for antioxidant management. Oxidation can generate polymerization in the oil, resulting in an increase in viscosity.

Elemental Metal Content (Fe, Cu) (Elemental Analysis): Metal ions, particularly Iron (Fe) and Copper (Cu), act as catalysts for oxidation reactions. An increase in their concentration accelerates the oxidative degradation of the lubricant.

Water Content (Physicochemical Analysis): The presence of water significantly accelerates oxidation reactions and promotes the formation of acidic substances. Maintaining low water content is essential for controlling oxidation rates.

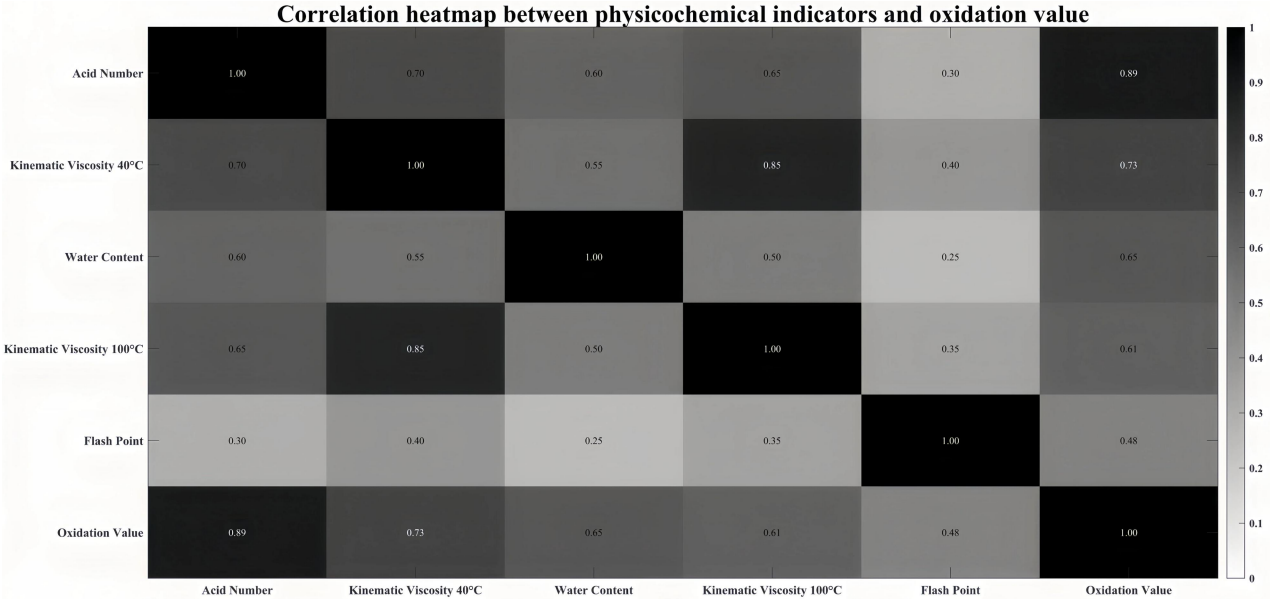

Fig 1. Correlation heatmap between physicochemical indicators and oxidation value

3.2.Results of data preprocessing

Data preprocessing is an important process to enhance modeling precision, and minimize instrumental errors, environmental interferences, and dimensional heterogeneity. In this study, the raw dataset (Raw) was optimized with five standard preprocessing techniques: MSC, Standard Normal Variate (SNV), Savitzky-Golay (SG) smoothing, First Derivative (FD), and Second Derivative (SD). The efficacy of these methods was validated via Partial Least Squares (PLS) regression, with comparative results presented in Table 4. The tabulated data revealed significant disparities in model performance across different preprocessing protocols. The model with raw data yielded a Prediction R2 of 0.9895 and a Prediction RMSE of 4.8762 for the prediction set. With SNV preprocessing, the model performance exhibited a substantial improvement: the prediction set R2 increased to 0.9952 while the RMSE decreased to 2.4318. The calibration set achieved R2 of 0.9985 with an RMSE of merely 0.6873. These metrics demonstrated that SNV preprocess effectively reduced data heteroscedasticity through normalization, mitigated the noise interference, and preserved essential information. On the contrary, the application of MSC resulted in an elevated prediction set RMSE of 7.5429. This may be due to the over-correction of MSC, which led to a data-distortion. Among all preprocessing methods, SNV was identified as the optimal preprocessing strategy for maximizing data quality and was selected for the subsequent processing.

Table 4. PLS modeling results under different preprocessing methods.

Preprocessing method	Calibration R^2	Calibration RMSE	Prediction R^2	Prediction RMSE
Raw data	0.9963	1.3257	0.9895	4.8762
MSC	0.9971	0.9436	0.9887	7.5429
SNV	0.9985	0.6873	0.9952	2.4318
SG smoothing	0.9958	1.5428	0.9901	3.5674
FD	0.9961	1.2845	0.9903	3.4219
SD	0.9967	1.1326	0.9912	2.9847

3.3.Feature extraction results

3.3.1.PCA feature extraction

Following the application of PCA to the SNV-preprocessed dataset, the distribution of variance contribution rate for the first eight principal components (PCs) was analyzed, as illustrated in Figure 2.

The variance contribution rate of PC2, PC3, PC4, and subsequent minor components are 12.47%, 6.15%, 3.86%, and 16.88%, respectively. The cumulative variance contribution of the first four principal components (with minor components included to reach the threshold) reaches 97.68%. The individual variance contribution of each component exceeds the 1% threshold, satisfying the standard criterion for principal component selection. These results demonstrate that the first four PCs sufficiently characterize the critical information inherent in the original five physicochemical indices. Consequently, PCA-based dimensionality reduction effectively eliminates multicollinearity among indices and simplifies the model architecture without compromising information integrity. Therefore, the first four principal components were selected as the feature variables for subsequent machine learning modeling.

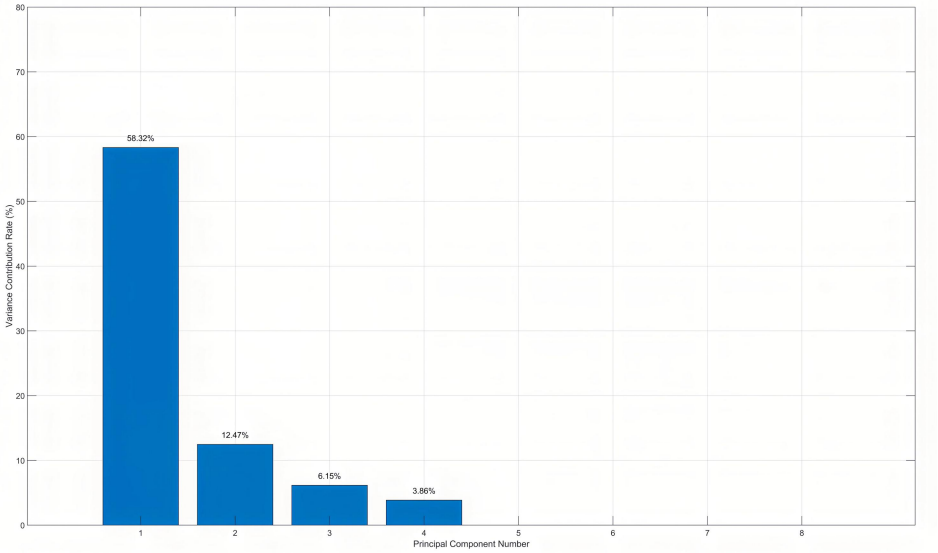

Fig 2. Variance contribution distribution of principal components for SNV-preprocessed gear oil physicochemical data

3.3.2.SPA feature extraction

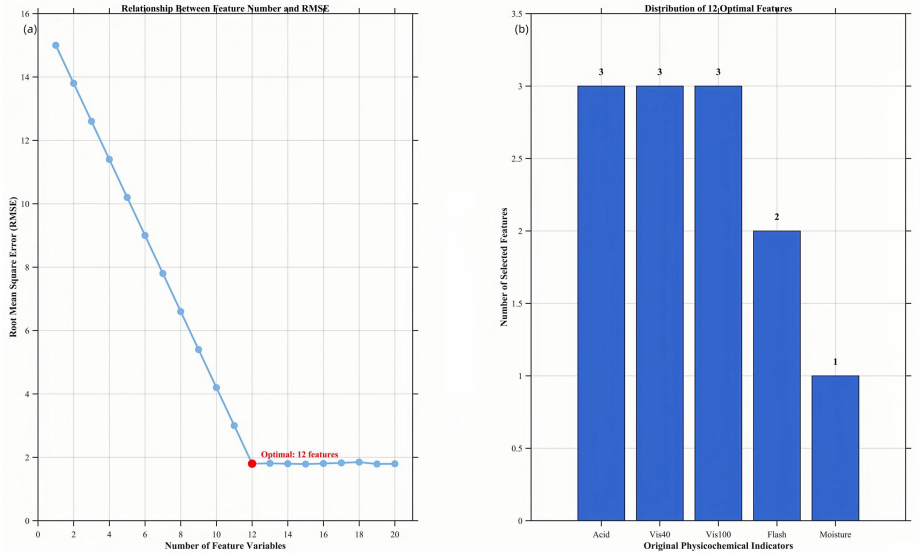

Fig 3. SPA feature extraction results for gear oil physicochemical indices: (a) RMSE vs. number of feature variables; (b) Distribution of optimal feature subset (n=12).

The Successive Projections Algorithm (SPA) was used to extract variables from the SNV-preprocessed dataset, which can minimize data redundancy by identifying variables with low correlation. The extraction results are presented in Figure 3.

Figure 3a depicts the relationship between the number of variables and the model's RMSE. The RMSE reached a minimum and achieved stability when the number of selected variables was 12. Figure 3b illustrates the specific distribution of these 12 variables (e.g., 3 features for Acid, 3 for Visc40, etc.), which corresponded to distinct combinational features derived from the original five physicochemical indices. However, a comparative assessment of the subsequent modeling revealed that the predictive efficacy of variables extracted via SPA is significantly less than PCA. It could be linked with the limited number of physicochemical indices. Under this condition, SPA could lead to the loss of valid information. Therefore, PCA was used in the current study for feature extraction.

3.4.Prediction model results and comparison

Based on the SNV-preprocessed dataset, six combinatorial prediction models were proposed as in, PCA-BP, PCA-SVM, PCA-RF, SPA-BP, SPA-SVM, and SPA-RF. These models integrated two extraction methodologies (PCA and SPA) with three machine learning algorithms (BPNN, SVM, and RF). The performance metrics for each model are detailed in Table 5, while the fitting efficacy for the calibration and prediction sets is visualized in Figure 4 and Figure 5, respectively.

An analysis of Table 5 and Figures 4–5 revealed distinct differentiation in the predictive performance between the six models:

(1)Impact of Feature Extraction

Models based on PCA feature extraction (PCA-BP, PCA-SVM, PCA-RF) demonstrated superior overall performance compared to their SPA-based counterparts. Specifically, PCA-associated models yielded prediction set R2 values over 0.86 and Residual Prediction Deviation (RPD) values greater than 2.18, satisfying the requisites for quantitative analysis. Conversely, SPA-associated models achieved a maximum prediction set R2 of only 0.8096, with RPD values consistently below 2.06. Notably, certain models (e.g., SPA-SVM) exhibited an RPD<1.5, indicating a lack of quantitative predictive capability. This disparity further demonstrated the performance of PCA for feature extraction in this study, attributed to its ability to retain core information through effective dimensionality reduction while reducing effects of multicollinearity[29].

(2)Comparison of Machine Learning Algorithms

With the same feature extraction, BPNN exhibited optimal performance as Machine Learning Algorithm. Taking PCA feature extraction as the baseline, the PCA-BP model achieved a prediction set R2 of 0.9960, with an RMSE of 2.0125 and an MAE of 3.3872. The resulting RPD of 6.1247 (>2.5) is indicative of excellent predictive capability. The PCA-SVM model followed, with a prediction set R2 of 0.9872. However, its significantly higher RMSE (6.8945) suggested that SVM has weaker fitting capabilities when processing small-sample nonlinear data compared to BPNN. The PCA-RF model yielded relatively poor predictive results (R2 = 0.8679, RMSE = 13.7654), likely due to the tendency of decision tree ensemble algorithms to overfit on small datasets[30].

Figure 4-5 revealed that data points for the PCA-BP model collapsed with the ideal regression line in both calibration and prediction sets (Figure 4a, Figure 5a), suggesting a better consistency between the predicted and observed values. In contrast, SPA-associated models exhibit greater dispersion (Figure 4d–f, Figure 5d–f), showing their insufficient predictive accuracy. Furthermore, the PCA-BP model demonstrated consistent prediction accuracy (R2=0.987) for samples with off-spec indices (e.g.,flashpoint<40◦ C, watercontent 0.18%) relative to normal samples. This shows that SNV preprocessing can effectively mitigate interference from outliers, ensuring robust model adaptability.

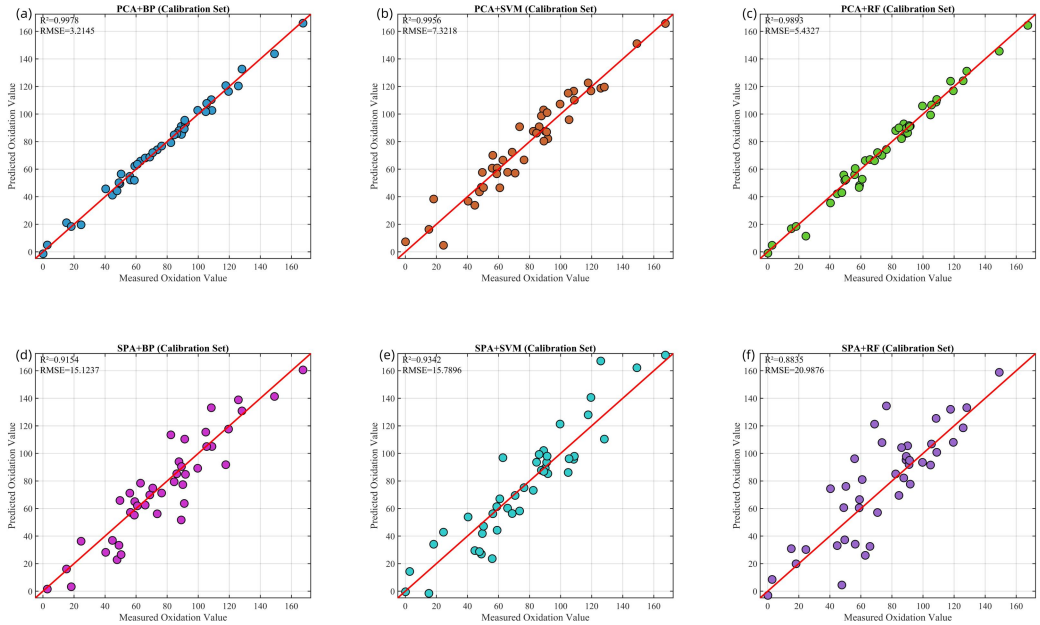

Fig 4. Measured vs. predicted oxidation value fitting plots for calibration set of six combinatorial prediction models (PCA/SPA + BP/SVM/RF)

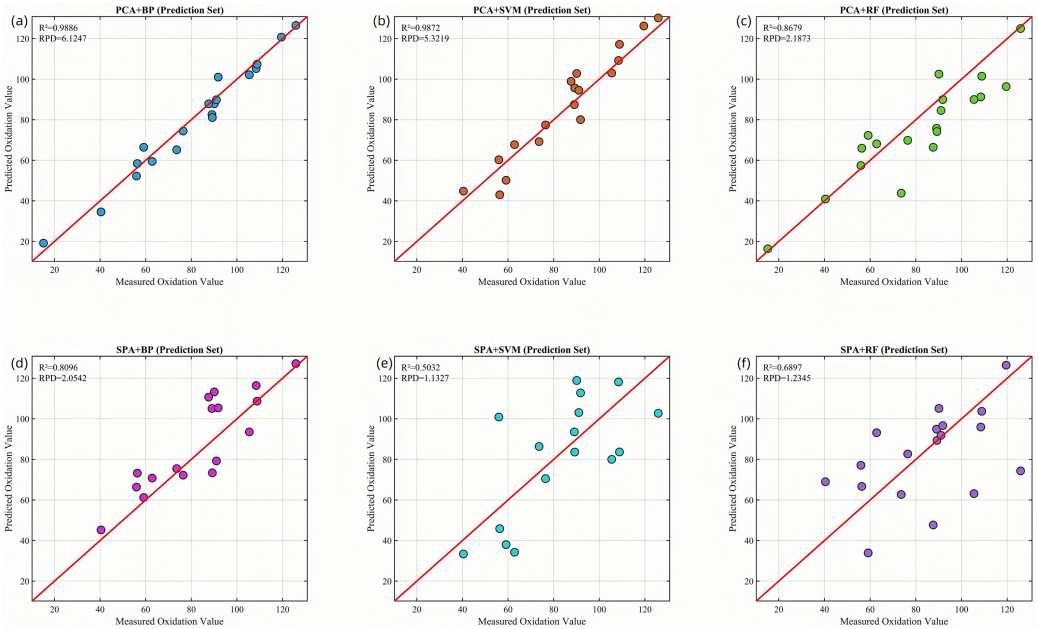

Fig 5. Measured vs. predicted oxidation value validation plots for prediction set of six combinatorial prediction models (PCA/SPA + BP/SVM/RF)

Table 5. Performance evaluation metrics of different combinatorial models.

Model	Calibration Set				Prediction Set			
	R^2	RMSE	MAE	RPD	R^2	RMSE	MAE	RPD
PCA+BP	0.9978	3.2145	2.1873	13.6824	0.9960	2.0125	3.3872	6.1247
PCA+SVM	0.9956	7.3218	6.0125	5.6438	0.9872	6.8945	5.9873	5.3219
PCA+RF	0.9893	5.4327	4.2186	7.9856	0.8679	13.7654	10.9876	2.1873
SPA+BP	0.9154	15.1237	9.8762	2.0875	0.8096	12.1345	12.4568	2.0542
SPA+SVM	0.9342	15.7896	12.9873	2.6541	0.5032	19.1234	15.4321	1.1327
SPA+RF	0.8835	20.9876	16.8945	2.0134	0.6897	21.5678	16.7895	1.2345

4.Discussion

This study focused on the prediction of antioxidant performance in gear lubricants. Through a systematic exploration of algorithms, comprising correlation analysis of detection indices, data preprocessing, variables extraction, and machine learning modeling, a series of results with significant engineering implications were obtained. By integrating these findings with existing literature and industrial application scenarios, the underlying mechanisms and practical value of this research are analyzed below.

(1)Correlation Analysis and Mechanistic Interpretation: The correlation analysis between detection indices and antioxidant performance reveals that the Acid Number exhibited the most robust correlation with the oxidation value (r=0.89). This finding is reflective of the chemical nature of oxidative degradation in gear oils[27]. Under conditions of high temperature, high pressure, and metal catalysis, base oil molecules undergo chain oxidation reactions, progressively generating oxidation products such as carboxylic acids and ketones. The accumulation of carboxylic species directly leads to a rise in Acid Number, in agreement with previous research[28]. The strong positive correlations observed for kinematic viscosity at 40°C (r=0.73) and water content (r=0.65) further demonstrated the synergistic effect between polymer formation and water-accelerated oxidation. Specifically, high-molecular-weight polymers formed during oxidation can increase oil viscosity, while water, acting as a polar substance, can disrupt the colloidal stability of the lubricant and diminish the activity of antioxidant additives, accelerating the oxidation process. Conversely, the flash point demonstrated a weaker correlation (r=0.48). This is attributed to the fact that flash point primarily reflects volatility of the fluid, and the alteration of volatile components during gear oil oxidation occurs at a rate significantly lower than the generation of oxidation byproducts. Consequently, the flash point exhibited low sensitivity to the degree of oxidation. These results provided a definitive basis for further understanding towards performance decay of lubricants, justifying the prioritization of high-correlation indices, such as Acid Number and kinematic viscosity (40°C), as core modeling variables.

(2)Optimization of Data Preprocessing: In the data preprocessing phase, SNV transformation yielded optimal performance. The model processed via SNV achieved a prediction set R2 of 0.9952 and an RMSE of 2.4318, significantly outperforming other methods such as MSC and Savitzky-Golay (SG) smoothing. The main reason lies in SNV's ability to effectively eliminate data heteroscedasticity and dimensional heterogeneity through normalization, while maximizing the preservation of the intrinsic correlations between physicochemical indices and oxidation values[31]. In contrast, MSC preprocessing resulted in a lower prediction accuracy (RMSE increased to 7.5429), likely due to over-correction causing distortion of valid information within the raw data. This aligned with findings in the literature suggesting that MSC is prone to overfitting when applied to low-dimensional data[32]. The marginal improvements

provided by SG smoothing and derivative methods indicated that the raw data in this study were of high quality, which were minimally affected by random noise or baseline drift, rendering complex smoothing unnecessary. This observation suggested a valid and defendable preprocessing approach in rapid industrial detection.

(3)Efficacy of Feature Extraction: Feature extraction results indicated that Principal Component Analysis (PCA) significantly outperformed the SPA. The first four principal components accounted for a cumulative variance contribution of 97.68%, sufficiently characterizing the critical information of the original five physicochemical indices. This advantage originated from PCA's effective mitigation of multicollinearity. Because physicochemical indices of gear oil are not entirely independent (e.g., the correlation between viscosity and flash point), PCA retained the core information contributing to maximum variance while simplifying the model structure[33]. On the contrary, the aggressive screening logic of SPA likely resulted in the loss of valid information. This study suggested that for low-dimensional feature extraction, methods like PCA are preferable to avoid information loss caused by excessive dimensionality reduction.

(4)Performance of Machine Learning Models: Regarding machine learning modeling, the PCA-BP ensemble model demonstrated better efficacy, with a prediction set R2 of 0.9960 and an RPD of 6.1247, significantly surpassing the PCA-SVM and PCA-RF models. This was the result of the integration of PCA's dimensionality reduction and the Back Propagation Neural Network's (BPNN) non-linear mapping capabilities. PCA reduced model complexity and overfitting risks and the BPNN dealt with the complex non-linear relationships between physicochemical indices and oxidation values. This was consistent with the established adaptability of BPNNs in non-linear regression tasks[34]. In comparison, the slightly lower accuracy of the SVM model (RMSE=6.8945) may be due to the limited generalization capability of the Radial Basis Function kernel on small datasets. The relatively poor performance of the Random Forest (RF) model is linked to the characteristics of decision tree ensemble algorithms, which are susceptible to overfitting when sample sizes are limited, leading to diminished predictive capacity for new samples[35].

(5)Innovations, Limitations, and Future Directions: Compared to existing research on wind turbine gear oil antioxidant prediction, there are three major novelties in the current study[36]. First, the samples were derived from actual in-service equipment in an industrial setting, covering a broad spectrum of industrial lubricants (including Shell Omala S2 G series, L-TSA 46#, CPI-1507-68#) across ISO VG 150–320 viscosity grades, with service durations spanning 1–2 years. This ensured realistic data authenticity and representativeness. Second, the SNV-PCA-BP model achieved high-precision prediction using only five routine physicochemical indices, eliminating reliance on complex techniques like infrared spectroscopy or mass spectrometry, lowering the barrier for on-site application. Third, the prediction efficiency of the model was substantially improved compared to traditional detection methods, satisfying the demand for rapid industrial maintenance detection. However, certain limitations must be acknowledged: the dataset currently covers only gear oils within a 1–2 year cycle, lacking long-term service samples (e.g., >3 years) with deep oxidation, and the influence of extreme environmental factors (e.g., low temperature, high humidity) was not considered, which could be a meaningful extensive of the current research.

Future research should focus on three areas: i) expanding the scope to include long-term service oils and those from extreme environments to enhance model generalization, ii) incorporating additional features and variables to optimize model inputs, such as antioxidant additive content and elemental metal concentrations, and iii) integrating these models with online monitoring technologies to develop real-time prediction systems.

5.Conclusions

Focusing on in-service gear oils from industrial field operations, this study systematically investigated the impact of data preprocessing, feature extraction, and machine learning algorithms on antioxidant performance prediction models. A rapid prediction methodology based on routine physicochemical indices was established through comparison. The main conclusions are summarized as follows:

(1)Optimization of Preprocessing Methods: Among the evaluated data preprocessing techniques, SNV transformation exhibited superior performance. It effectively mitigated the influence of instrumental errors, environmental interference, and data heteroscedasticity, while retaining core information.

(2)Efficacy of Feature Extraction Methods: Comparative analysis revealed that PCA significantly outperformed the SPA in dimensionality reduction. PCA effectively eliminated multicollinearity among physicochemical indices and simplified the model structure without compromising information integrity.

(3)Algorithm Performance: Of the three machine learning algorithms assessed, the BPNN demonstrated the most robust non-linear fitting capability, enabling the modelling of complex intrinsic correlations between physicochemical indices and oxidation values. Consequently, the combined model of BPNN with PCA feature extraction yielded optimal performance.

(4)Model Utility and Engineering Value: The final SNV-PCA-BP hybrid model is characterized by high precision and efficiency, achieving a prediction set coefficient of determination R2 of 0.9960 and a Residual Prediction Deviation (RPD) of 6.1247. This enabled the rapid and accurate prediction of wind turbine gear oil antioxidant performance. Requiring only five routine physicochemical indices as inputs, the model is able to conduct rapid assessments and conclude detection within seconds. Furthermore, the model was fully established on real-life in-service lubricants, across ISO VG 150–320 viscosity grades (including Shell Omala S2 G series, L-TSA 46#, and CPI-1507-68#) over a 1–2 year service cycle. This provided a low-cost and convenient approach for oil condition assessment in industrial maintenance, allowing customized oil change strategies and reducing both maintenance costs and failure risks.

References

[1] DÍAZ-DÍAZ A M, LÓPEZ-BECEIRÓ J J, DÍAZ R P A, et al. Oxidative stability of edible oils: linking rancimat to PDSC results. Journal of Food Measurement and Characterization, 2026, 20:1139-1150.

[2] JIANG S, XIAO Y, LI Q, et al. SERS technology in virus Detection: Advances, challenges, and future perspectives. Biosensors and Bioelectronics, 2025, 290: 117902.

[3] ABDELFATTAH W, ABOSAODA M K, DOSHI H, et al. Development of data driven models to accurately estimate density of fatty acid ethyl esters. Scientific Reports, 2025, 15: 30961.

[4] SAGRALOFF N, DOBLER A, TOBIE T, et al. Development of an Oil Free Water-Based Lubricant for Gear Applications. Lubricants, 2019, 7(4): 33.

[5] GRIBOK A, HINES J W, URMANOV A M, et al. Heuristic, systematic, and informational regularization for process monitoring. International Journal of Intelligent Systems, 2002, 17.

[6] ROYLANCE B J, POCOCK G. Wear studies through particle size distribution I: Application of the Weibull distribution to ferrography. Wear, 1983, 90(1): 113–136.

[7] GONCALVES I M, MURILLO M, GONZÁLEZ A M. Determination of metals in used lubricating oils by AAS using emulsified samples. Talanta, 1998, 47(4): 1033–1042.

[8] KROGSØE K, ERIKSEN R L, HENNEBERG M. Performance of a light extinction based wear particle counter under various contamination levels. Sensors and Actuators A: Physical, 2021, 331: 112956.

[9] ABBASI S, JANSSON A, OLANDER L, et al. A pin-on-disc study of the rate of airborne wear particle emissions from railway braking materials. Wear, 2012, 284–285: 18–29.

[10] AHMED H E, HASSAN M, NOUR M, et al. Lubricant Oils as a Certified Reference Material for Cleveland Open Cup Flash Point Testers. MAPAN, 2017, 32: 215–222.

[11] MACIÁN V, TORMOS B, GÓMEZ Y, et al. Proposal of an FTIR Methodology to Monitor Oxidation Level in Used Engine Oils: Effects of Thermal Degradation and Fuel Dilution. Tribology Transactions, 2012, 69.

[12] ESQUIVEL M M, RIBEIRO M A, BERNARDO-GIL M G. Relations between Oxidative Stability and Antioxidant Content in Vegetable Oils Using an Accelerated Oxidation Test - Rancimat. 2009, 4(4).

[13] GAO T, HU L, JIA Z, et al. SPXYE: an improved method for partitioning training and validation sets. Cluster Computing, 2019, 22: 3069–3078.

[14] RAWAT S, CUI H, XIE Y, et al. An improved framework for multi-objective optimization of cementitious composites using Taguchi-TOPSIS approach. Expert Systems with Applications, 2025, 272: 126732.

[15] DHANOA M, LÓPEZ S, SANDERSON R, et al. Methodology adjusting for least squares regression slope in the application of multiplicative scatter correction to near-infrared spectra of forage feed samples. Journal of Chemometrics, 2023, 37(11):e3511.

[16] SHRESTHA B, POSOM J, SIRISOMBOON P, et al. Comprehensive Assessment of Biomass Properties for Energy Usage Using Near-Infrared Spectroscopy and Spectral Multi-Preprocessing Techniques. Energies, 2023, 16(14): 5351.

[17] TALEBANPOUR BAYAT E, HEMMATEENEJAD B, AKHOND M, et al. On the dependency between principal components: Application to determine the rank of a matrix in an evolutionary process. Journal of Chemometrics, 2018, 33: e3102.

[18] SOARES A, GALVÃO FILHO A, GALVÃO R. Improving the computational efficiency of the successive projections algorithm by using a sequential regression implementation: a case study involving nir spectrometric analysis of wheat samples. Journal of the Brazilian Chemical Society, 2010, 21(4): 760–763.

[19] YARAHMADI M N, MIRHASSANI S A, HOOSHMAND F. Handling the significance of regression coefficients via optimization. Expert Systems with Applications, 2024, 238: 121910.

[20] BAGHOLIZADEH M, NASAJPOUR-ESFAHANI N, PIRMORADIAN M, et al. Using different machine learning algorithms to predict the rheological behavior of oil SAE40-based nano-lubricant in the presence of MWCNT and MgO nanoparticles. Tribology International, 2023, 187: 108759.

[21] BARBERO JIMÉNEZ Á, LÓPEZ LÁZARO J, DORRONSORO J R. Finding optimal model parameters by deterministic and annealed focused grid search. Neurocomputing, 2009, 72(13–15): 2824–2832.

[22] MA Y, HUANG M, WAN J, et al. Prediction model of DnBP degradation based on BP neural network in AAO system. Bioresource Technology, 2011, 102(5): 4410–4415.

[23] PANG H X, DONG W D, XU Z H, et al. Novel Linear Search for Support Vector Machine Parameter Selection. Journal of Zhejiang University-SCIENCE C (Computers & Electronics), 2011, 12: 885–896.

[24] WAINER J, FONSECA P. How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms. Artificial Intelligence Review, 2021, 54: 4771–4797.

[25] LIU Y, ZHOU Y, WEN S, et al. A Strategy on Selecting Performance Metrics for Classifier Evaluation. International Journal of Mobile Computing and Multimedia Communications, 2014, 6(4): 20–35.

[26] CHICCO D, WARRENS M J, JURMAN G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 2021, 7: e623.

[27] MUKHIN A A, SKRYABINA A E, FADEEEV V K, et al. Express method for determining acid number of lubricating oils for gas-pumping units. Chemistry and Technology of Fuels and Oils, 2013, 49: 359–361.

[28] LACROIX-ANDRIVET O, HUBERT M, LOUTELIER-BOURHIS C, et al. Characterization of Base Oil and Additive Oxidation Products from Formulated Lubricant by Ultra-High Resolution Mass Spectrometry. Lubricants, 2023, 11(8): 345.

[29] SALEM N, HUSSEIN S. Data dimensional reduction and principal components analysis. Procedia Computer Science, 2019, 163: 292–299.

[30] DOS SANTOS E M, SABOURIN R, MAUPIN P. Overfitting cautious selection of classifier ensembles with genetic algorithms. Information Fusion, 2009, 10(2): 150–162.

[31] GRISANTI E, TOTSKA M, HUBER S, et al. Dynamic localized SNV, Peak SNV, and partial peak SNV: Novel standardization methods for preprocessing of spectroscopic data used in predictive modeling. Journal of Spectroscopy, 2018, 2018: 5037572.

[32] UMPRECHT A, FONSECA DIAZ V, HÜPFL B, et al. Unsupervised optimization of spectral pre-processing selection to achieve transfer of Raman calibration models. Measurement, 2025, 255: 117906.

[33] HATHOUT R M. Using principal component analysis in studying the transdermal delivery of a lipophilic drug from soft nano-colloidal carriers to develop a quantitative composition effect permeability relationship. Pharmaceutical Development and Technology, 2014, 19(5): 598–604.

[34] ZHANG J, CAO J, WANG L. Robust Bayesian functional principal component analysis. Statistics and Computing, 2025, 35: 46.

[35] LONG J, LI T, YANG M, et al. Hybrid Strategy Integrating Variable Selection and a Neural Network for Fluid Catalytic Cracking Modeling. Industrial & Engineering Chemistry Research, 2019, 58(1): 247–258.

[36] BURKHART C, JOHANSSON J, UKONSAARI J, et al. Performance of lubricating oils for wind turbine gear boxes and bearings. Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology, 2017, 232.