

# A new method for identifying lubricant additives: combining PCA and stack ensemble regression model with infrared spectroscopy

Nay Min Aung, Yanqiu Xia and Xin Feng

School of Energy, Power and Mechanical Engineering, North China Electric Power University, Beijing, China

## Abstract

**Purpose** – This study aims to develop an innovative technique for identifying and predicting lubricant components using Fourier-transform infrared spectroscopy (FTIR) spectroscopy data with Principal Component Analysis (PCA) and a Stacked Ensemble Regression Model Extreme Gradient Boosting (SER\_XGBRegressor). This technique is especially applicable to sectors such as wind energy, power generation, automotive and other similar industries, where the exact composition of lubricants plays a vital role in enhancing the performance and durability of mechanical components.

**Design/methodology/approach** – This study analyzed 35 lubricant samples comprising molybdenum dialkyl dithiocarbamate (MoDTC), zinc dialkyl dithiophosphate (ZnDDP), isoctyl acid phosphorodithioate amine salt (T308) and thiophosphoric ester amine salt (T310A). FTIR spectroscopy yielded extensive spectral data, in which PCA decreased the dimensionality while preserving more than 90% of data variability. Subsequently, an XGBRegressor-stacked ensemble regression model was used to accurately predict the lubricant components.

**Findings** – The PCA-SER Model attained high accuracy in predicting lubricant components ( $R^2$ : 0.983–0.996) for additives such as MoDTC, ZnDDP, T308 and T310A. PCA reduced dimensionality, preserving over 90% variance, while reducing errors to a mean absolute error of  $9.9 \times 10^{-5}$  and a mean squared error of  $1.6 \times 10^{-8}$ . These scores illustrate the ability of the model to precisely predict and classify lubricant components, even in complex and high-dimensional FTIR data sets.

**Originality/value** – This study presents a novel PCA and stacked ensemble learning framework for analyzing high-dimensional FTIR data, enhancing lubricant classification and prediction, optimizing formulation processes and ensuring quality control of lubricants.

**Peer review** – The peer review history for this article is available at: <https://publons.com/publon/10.1108/ILT-12-2024-0472/>

**Keywords** FTIR spectroscopy, Principal component analysis (PCA), Stacked ensemble regression model (XGBR), Lubricant additives

**Paper type** Research paper

## 1. Introduction

Lubricant oil is a specialized oil widely used in machinery and various industries to ensure the optimal performance of mechanical equipment. In addition, it protects moving components from undesirable effects, such as heat, pressure, corrosion, oxidation and contamination, among other critical activities (Hamnas and Unnikrishnan, 2023). Selecting an appropriate lubrication solution may increase the machine availability, productivity and energy efficiency. In general, the quality of lubricating oil depends on the variety and quality of the additives (Wen *et al.*, 2024).

Base oils, which serve as the foundation for lubricants, are generally classified into three categories based on their source and refining process: mineral, synthetic and vegetable oils. Additives, typically comprising 1%–10% of the lubricant by weight, are blended with these base oils to enhance their performance (Xia *et al.*, 2021). In this study, four key additives, molybdenum dialkyl dithiocarbamate (MoDTC),

zinc dialkyl dithiophosphate (ZnDDP), isoctyl acid phosphorodithioate amine salt (T308) and thiophosphoric ester amine salt (T310A), were investigated for their friction-reducing and anti-wear properties. Notably, ZnDDP and MoDTC also exhibit antioxidant properties (Xia *et al.*, 2024a, 2024b, 2024c). Given the importance of base oil compatibility, polyalphaolefin (PAO 40) was selected as the base oil due to its excellent thermal stability, high oxidation resistance and superior compatibility with performance-enhancing additives (Xia *et al.*, 2024a, 2024b, 2024c). These properties make PAO 40 highly suitable for tribological applications involving high loads and elevated temperatures. The effective analysis of lubricant quality is critical, as it ensures proper identification and usage, preventing mechanical inefficiencies and equipment damage caused by lubricant misuse. In addition, accurate lubricant analysis helps differentiate between fresh, contaminated and degraded oils, facilitating timely maintenance. Advanced predictive techniques enable rapid and accurate assessments of oil quality, which directly impact the performance and durability of equipment.

Existing data have confirmed that over 70% of mechanical equipment failures are caused by improper lubrication

The current issue and full text archive of this journal is available on Emerald Insight at: <https://www.emerald.com/insight/0036-8792.htm>



(Peng *et al.*, 2022). Therefore, choosing the appropriate lubricating oil is crucial for reducing the occurrence of mechanical equipment failures. Traditional lubricant identification methods, such as chromatographic analysis and chemical testing, are often time-consuming, labor-intensive and require a large number of samples, which is not conducive to the real-time detection of lubricants (Li *et al.*, 2024). Fourier-transform infrared (FTIR) spectroscopy is an efficient method for analyzing complex mixtures because it can provide precise information on molecular composition (Xia *et al.*, 2022; Chu *et al.*, 2024; Agulei *et al.*, 2023). However, FTIR spectroscopy alone faces challenges when analyzing complex lubricant mixtures owing to overlapping spectral features and high-dimensional data (Heen Blindheim and Ruwoldt, 2025; Awasthi *et al.*, 2024), necessitating advanced analytical approaches to effectively interpret the data (Xu *et al.*, 2023; Heinrich *et al.*, 2024).

Recent studies have used chemometric techniques (Agulei *et al.*, 2023; Xia *et al.*, 2020), such as classical least squares (CLS) (Maboudou-Tchao, 2020), inverse least squares (ILS) (Mayerhöfer *et al.*, 2022) and partial least squares (PLS) regression (Heen Blindheim and Ruwoldt, 2025; Perera *et al.*, 2021; Singh *et al.*, 2025), to enhance the analytical capabilities of FTIR. However, despite these advancements, traditional chemometric methods still exhibit limitations, especially in handling highly correlated variables and complex high-dimensional data sets (Zhu *et al.*, 2024; Xia *et al.*, 2024a, 2024b, 2024c).

This study proposes a novel analytical approach that integrates Principal Component Analysis (PCA) (Greenacre *et al.*, 2022) with a machine learning model called SER\_XGBR (Arachchilage *et al.*, 2024). PCA is used to reduce data dimensionality and isolate key spectral features, mitigating multicollinearity and complexity. The SER\_XGBR model then uses these optimized inputs to produce highly accurate predictions of lubricant composition. This integrated method significantly outperforms traditional chemometric techniques in both speed and accuracy, offering a powerful tool for lubricant quality control, formulation development and sustainable industrial practices.

## 2. Experimental

### 2.1 Analytical instruments and samples

To prepare for FTIR analysis, a standardized methodology was implemented for all lubricant and additive samples to ensure consistency and reproducibility. Initially, the base oil (PAO40) was warmed to 60°C and continuously stirred using a magnetic stirrer to ensure uniformity and homogeneity. Each additive (MoDTC, ZnDDP, T308 and T310A) was then separately blended into the base oil at predetermined concentrations, varying from 1% to 8% by weight, in accordance with specific formulation requirements. Following the addition of each additive, the mixtures were stirred continuously for approximately 30 min to ensure the thorough dispersion and even distribution of the additives throughout the lubricant matrix. Prepared samples were allowed to cool passively to ambient temperature before being analyzed using FTIR spectroscopy. This study used

analytical tools and samples, covering a total of 35 different types of lubricants, as listed in Table 1.

The four distinct varieties were as follows: PAO40 as the base oil was purchased from Kunlun Lubricating Oil Company (Beijing, China); MoDTC was provided by Nippon Oil Corporation; T310A and T308 were purchased from Shenyang Hualun Lubricant Additives Co., Ltd. (Liaoning, China) and ZnDDP was provided by PetroChina Lanzhou Lubrication Center, as shown in Table 2.

The principal analytical instrument used in this study was an FTIR spectrometer (Nicolet iS5, Thermo Scientific). The spectral features observed between 400 and 4,000  $\text{cm}^{-1}$  exhibited a transmission rate of above 92% and a data interval precision of 1.928  $\text{cm}^{-1}$ . These additives are frequently included in lubricant formulations to improve their performance.

### 2.2 PCA\_SER (XGBR regressor) model

PCA combined with a Stacked Ensemble Regressor, specifically the Xtreme Gradient Boosting Regressor (XGBR), is a sophisticated and efficient method for the predictive modeling of lubricant constituents based on FTIR spectroscopy data. The most informative features were extracted using PCA, which converts the original spectral data into a set of orthogonal, principal components (PCs):

$$\sum Z_{ij} = x_j^T V_i \quad (1)$$

where the eigenvectors  $V_1, V_2, \dots, V_k$  correspond to the largest  $k$  eigenvalues,  $Z_{ij}$  is the inner product of  $x_j^T$  with the  $i$ th eigenvector  $V_i$ , which yields the  $i$ th PC score for the  $j$ th observation  $x_j$ . These PC scores comprise the  $n \times k$  dimension-converted data matrix column  $Z$ . The optimal number of PCs was determined by combining empirical testing and standard analytical practices. In spectral analysis, retaining components that cumulatively explain approximately 90% of the variance is a common criterion used to balance dimensionality reduction and information retention. To validate this approach for our specific data set, preliminary analyses were conducted by assessing the variance explained by varying numbers of PCs. The Scree Plot [Figure 1(a)] indicated a significant drop in variance contribution after the fourth component, and the Cumulative Variance Plot [Figure 1(b)] further confirmed that using the first four PCs met the 90% variance threshold. Consequently, this combined approach ensured that the selected PCs adequately captured essential information while effectively reducing the data set complexity. After the transpose of the matrix containing the top  $k$  eigenvectors is multiplied by the lower-dimensional representation  $Z$ , the original data can be approximately reconstructed by adding back the mean that was subtracted during the centering process.

Once dimensionality reduction is complete, the XGBR uses a portion of the Stack Ensemble Regressor in the subsequent phase. XGBR builds an ensemble of decision trees (DTs) and trains each tree to generate predictions using subsets of the available data. Throughout the optimization phase, we calculated the first- and second-order gradients using the Hessian to tune the model parameters with higher accuracy. These formulas are (Iaousse *et al.*, 2020):

Table 1 Percentage additives of 35 types of lubricants

Sample_ID	MoDTC (%)	ZnDDP (%)	T308 (%)	T310A (%)	Sample_ID	MoDTC (%)	ZnDDP (%)	T308 (%)	T310A (%)
A_1	0.1	0.8	0.6	0.4	S_1	0.2	0.8	0.2	0.6
B_1	0.3	0.8	0.2	0.4	T_1	0.1	0.8	0.4	0.6
C_1	0.4	0.6	0.2	0.4	U_1	0.6	0.4	0.8	0.2
D_1	0.4	0.6	0.4	0.2	V_1	0.8	0.4	0.2	0.6
E_1	0.4	0.4	0.2	0.6	W_1	0.4	0.6	0.2	0.8
F_1	0.4	0.4	0.6	0.2	X_1	0.2	0.4	0.6	0.8
G_1	0.2	0.2	0.8	0.6	Y_1	0.4	0.6	0.2	0.8
H_1	0.1	0.4	0.6	0.8	Z_1	0.2	0.4	0.8	0.6
I_1	0.1	0.4	0.6	0.8	AA_1	0.6	0.4	0.2	0.8
J_1	0.2	0.6	0.2	0.8	BB_1	0.4	0.8	0.2	0.6
K_1	0.2	0.2	0.6	0.8	CC_1	0.2	0.6	0.4	0.8
L_1	0.1	0.4	0.8	0.6	DD_1	0.6	0.2	0.4	0.8
M_1	0.3	0.8	0.4	0.2	EE_1	0.8	0.4	0.2	0.6
N_1	0.3	0.2	0.8	0.4	FF_1	0.2	0.8	0.6	0.4
O_1	0.4	0.6	0.2	0.4	GG_1	0.6	0.2	0.8	0.4
P_1	0.4	0.6	0.2	0.4	HH_1	0.4	0.8	0.6	0.2
Q_1	0.4	0.4	0.6	0.2	II_1	0.6	0.2	0.4	0.8
R_1	0.4	0.6	0.2	0.4					

Source(s): Authors' own work

Table 2 Main typical properties of polyalphaolefin (PAO40)

Property	Limits
Appearance	Bright and clear
Kinematic viscosity @100°C, mm <sup>2</sup> /s	39–41
Kinematic viscosity @40°C, mm <sup>2</sup> /s	370–420
Viscosity index	147 typical value
Flash point (COC), °C	270 minimum
Pour point, °C	–35 maximum
Density @20°C, g/cm <sup>3</sup>	7.09 typical value

Source(s): Authors' own work

$$g_i = \frac{\partial(y_i, \hat{y}_i)}{\partial \hat{y}_i} \quad (2)$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \quad (3)$$

where  $g_i$  is first derivative (gradient) of the loss function,  $h_i$  Second derivative (Hessian),  $\partial$  represents a partial derivative,  $L$  represents the loss function,  $y_i$  are the actual values and  $\hat{y}_i$  are the predicted values. Figure 2 shows a flowchart of the PCA\_Stack Ensemble Regressor (XGBR) model.

AdaBoost regressors (ABR) (Jin *et al.*, 2020), XGBR (Pristyanto *et al.*, 2023) and DT (Sun *et al.*, 2024) are a few examples of base models trained independently on PCA-transformed data in a stacked ensemble framework. These foundational models produce forecasts that are used as input features by a meta-learner, such as XGBR. Specifically, the data not used during the training of the base models were used as input to the models. Together with the predicted outcomes, these predictions were used to create the input and output pairs of the training data set used to train the meta-model. This complementary combination enables the development of an

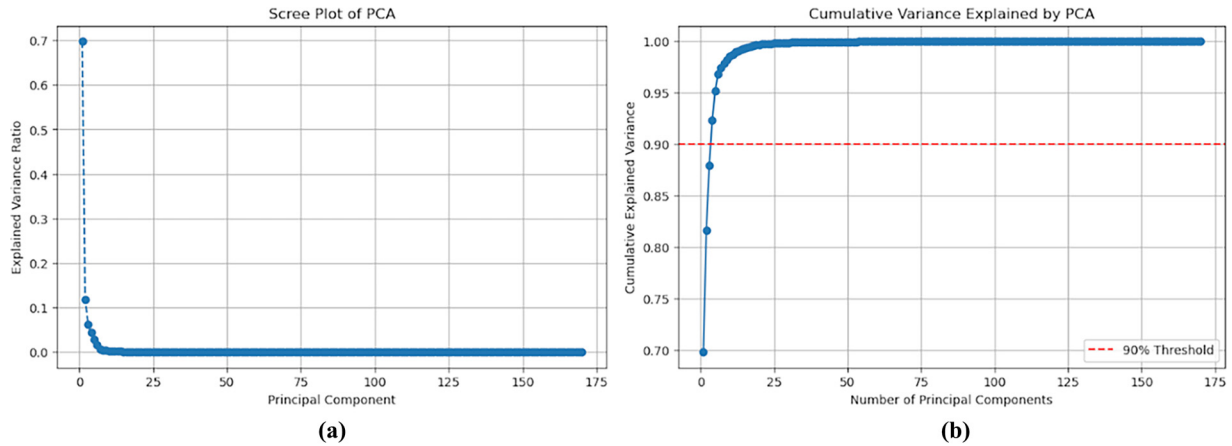
effective predictive model capable of reliably identifying and determining lubricating ingredients using complex-wavelength data.

### 2.3 An examination of the Fourier-transform infrared bands

FTIR spectroscopy determines certain chemical features by passing infrared light through a sample and analyzing the absorption and transmission of the infrared radiation. FTIR is highly regarded for its ability to provide comprehensive insights into molecular structures and their associated functional groups. These bands are predominantly detected in the IR spectra in the range of 3,300–600 cm<sup>-1</sup> and are essential for elucidating the structural characteristics of the chemicals. Figure 3 depicts the spectrum fluctuations, emphasizing the correlation between the unique FTIR features and the particular attributes of the examined lubricant samples.

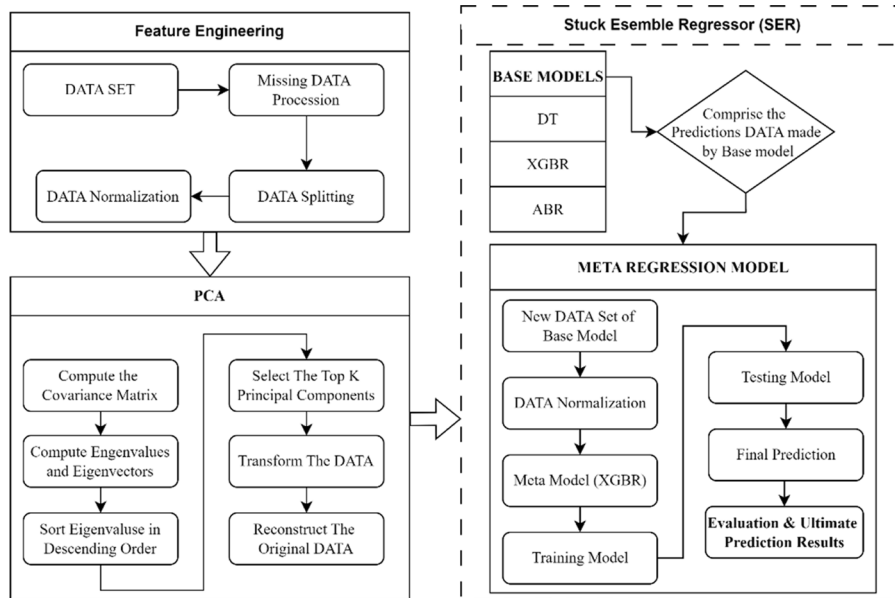
The MoDTC additive is very complex owing to the possibility of multiple vibration modes (Garcia *et al.*, 2021; Wang *et al.*, 2025). MoDTC is expected to show characteristic absorption peaks associated with carbamate groups and sulfur-containing compounds; C = O stretching vibrations around 1,700–1,750 cm<sup>-1</sup> and N-H stretching near 1,550–1,650 cm<sup>-1</sup>. Sulfur-related stretches, particularly those involving Mo-S bonds, appeared in regions below 650 cm<sup>-1</sup>. ZnDDP is known for its  $\nu$  = O stretching vibration at 1,000–1,200 cm<sup>-1</sup>, P-S stretching near 900–1,000 cm<sup>-1</sup> and possibly O-P-O and S-P-S stretching vibrations in the same region (Shin *et al.*, 2022). For T308, the amine salt functional group is highly polar, with a positively charged nitrogen atom balanced by a negatively charged acid anion. N-H stretching vibrations were observed at approximately 3,300 cm<sup>-1</sup> and 1,600 cm<sup>-1</sup>, respectively. For T310A, the ester band was observed at C = O stretching vibrations around 1,700–1,750 cm<sup>-1</sup> and 1,100–1,300 cm<sup>-1</sup>.

Figure 1 (a) Scree and (b) cumulative variance of PCA



Source: Authors' own work

Figure 2 Flow chart diagram of PCA\_SER model (XGBRegressor)



Source: Authors' own work

### 3. Result and discussion

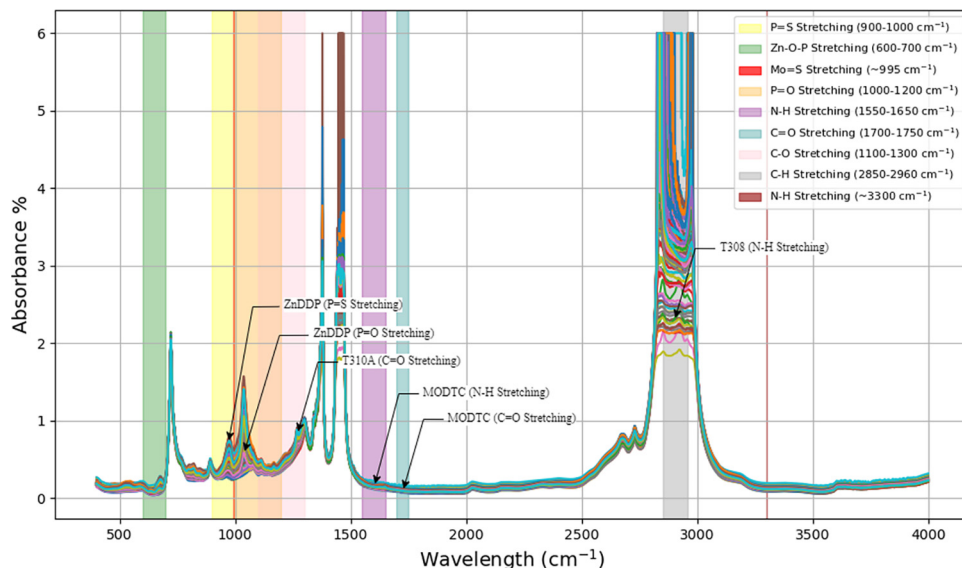
#### 3.1 Analysis of the principal component analysis variance

Figure 4 illustrates the PCA explained variance graphs for four different components: MoDTC [Figure 4(a)], ZnDDP [Figure 4(b)], T308 [Figure 4(c)] and T310A [Figure 4(d)]. The charts show the total percentage of variance explained by the consecutive PCs, offering insights into how well the PCA reduces the dimensionality of complex data sets while maintaining the crucial information.

In the case of MoDTC (Figure 3a), the curve initially demonstrated a pronounced ascent, with the first two PCs accounting for more than 90% of the total variation in the data set. The curve progressively stabilized and approached 100%, indicating that the new components yielded diminishing

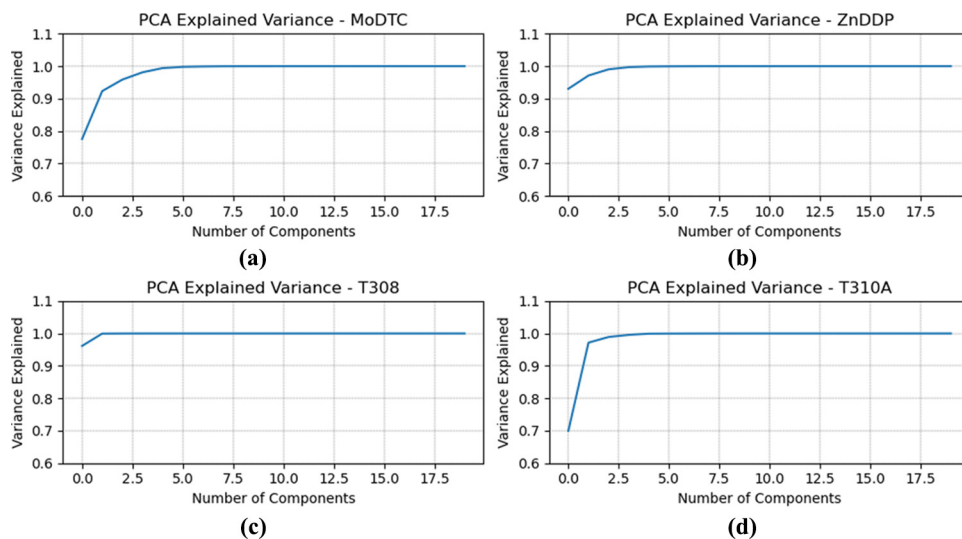
contributions to overall variation. PCA efficiently identified the most critical data features, as the first two components accounted for 90% of ZnDDP [Figure 3(b)]. The curve increased rapidly, indicating that the primary sources of variability have been identified. The PCA results for T308 [Figure 3(c)] further corroborate this tendency. The first two components represented approximately 90% of the total variance, confirming the efficacy of the PCA in identifying the most significant patterns in the data set. This indicates that the supplementary material provide only negligible enhancements to the overall variability, which is typically linked to modest and localized fluctuations or noise. The graph for T310A [Figure 3(d)] shows a trend similar to that of Tm. The first two components accounted for almost 90% of the overall variation, exhibiting a sharp ascent before leveling off when additional components were added.

Figure 3 FTIR spectra of lubricating oil components and their characteristic absorption bands



Source: Authors' own work

Figure 4 PCA Explained variance for MoDTC(a), ZnDDP(b), T308(c), and T310A(d)



Source: Authors' own work

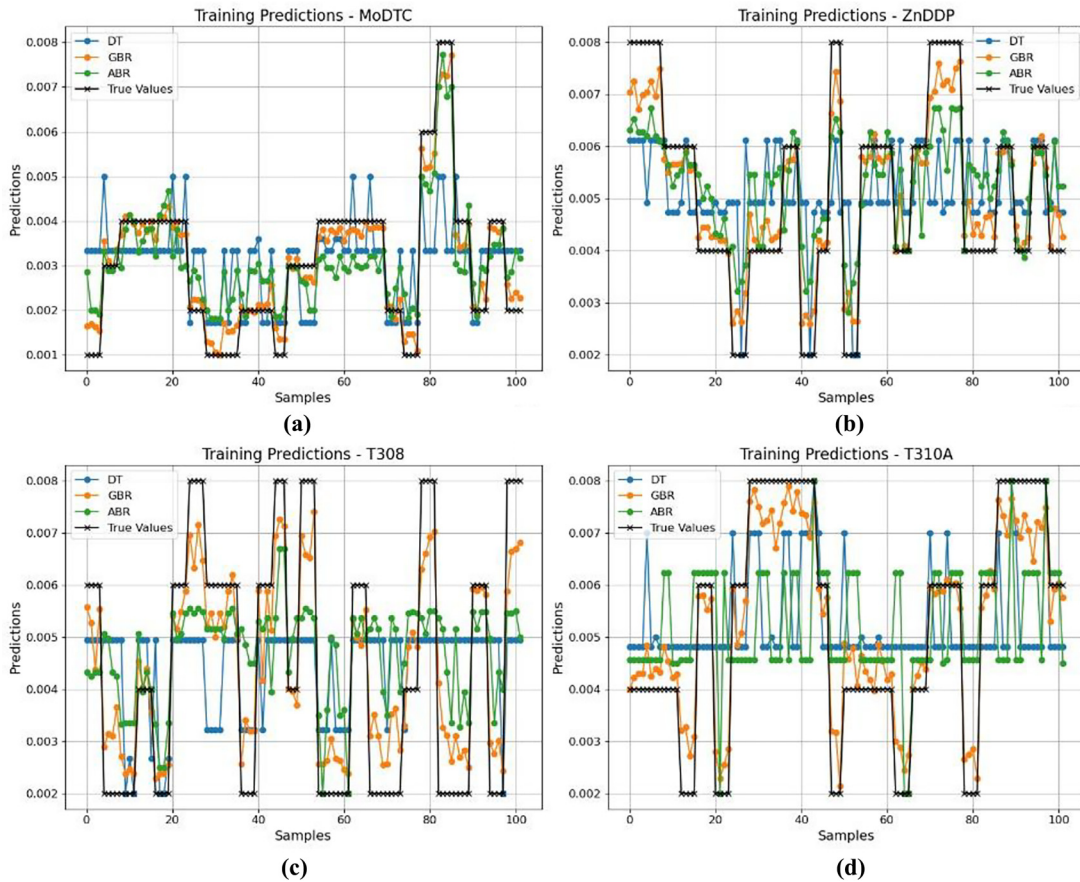
The MoDTC, ZnDDP, T308 and T310A PCA results were consistent across all four data sets. In each instance, multiple major components, usually the first two, accounted for more than 90% of the cumulative variation. This discovery underscores the efficacy of PCA as a method for dimensionality reduction, particularly in data sets in which a limited number of underlying factors govern variability.

### 3.2 Analysis of prediction variance of the stack ensemble regression (XGBR)

The test of model performance on MoDTC, ZnDDP, T308 and T310A showed that the three regression models used (DT, XGBR and ABR) had different levels of success. The

predictions of the training set demonstrated that the XGBR model regularly tracked the actual values closely, suggesting its resilience in capturing the trends. In contrast, the ABR model exhibited higher variability, which was particularly evident for MoDTC and T308 values. This implies that ABR are more susceptible to interference or fluctuations in the training data. Conversely, DT has a moderate prediction accuracy but shows variations in data sets, such as T310A. The accurate prediction of the ensemble learning technique from the training data in Figure 5 shows a tight alignment between the meta-model predictions and the actual values.

From the histogram on the diagonal of Figure 6, we can observe the distribution characteristics of each target variable

**Figure 5** Training and test set of base model and meta-model predictions for (a) MoDTC, (b) ZnDDP, (c) T308 and (d) T310A

Source: Authors' own work

(MoDTC, ZnDDP, T308 and T310A). The distribution of MoDTC was relatively scattered but was mainly concentrated in the range of 0.002–0.006, showing a specific right-biased distribution. The distribution of ZnDDP data was more uniform, and the values were mainly distributed in the range of 0.004–0.007, indicating that the variable has a high mean characteristic. For T308 and T310A, the distributions of both are more concentrated, especially the data value of T310A, which is mainly concentrated in the range of 0.0025–0.006, indicating that the data fluctuates less and has an intense concentration.

The paired scatter plots show the pairwise relationships between target variables. From the scatter distribution, it can be observed that there is almost no apparent linear correlation between MoDTC and the other three variables (ZnDDP, T308 and T310A), and the scatter distribution is relatively random, indicating a low correlation between them. This independence characteristic indicates that in the modeling process, the use of dimensionality reduction methods can effectively retain the main features and eliminate redundant information, which helps improve model performance and prediction accuracy.

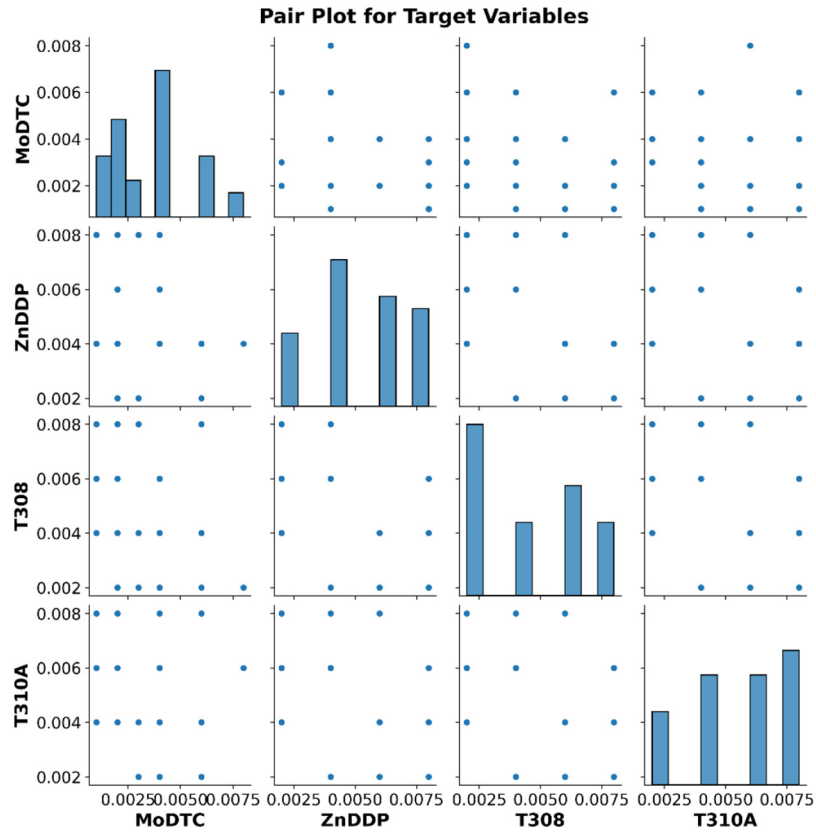
### 3.3 Performance analysis of PCA\_SER(XGBR) models

Three distinct machine learning algorithms, PCA\_SER(XGBR), SER\_XGBR and partial least squares regression (PLSR)\_SER(XGBR), were used to identify the lubricant

additives MoDTC, ZnDDP, T308 and T310A. The predictions from the training set indicated that the SER\_XGBR model, which lacked dimensionality reduction, exhibited considerable fluctuations and discrepancies compared to the actual values. The training results of the MoDTC data set show a comparison between the model prediction and true values (Figure 7). The red solid line represents the true value, the blue dashed line represents the XGBR model prediction and the yellow and green dashed lines represent the PCA\_SER(XGBR) and PLSR\_SER(XGBR) models, respectively. As shown in Figure 6, the true value fluctuated in the range of 0.003–0.008 m. The predicted value of the XGBR model near the early (0–20) data points was significantly higher than the true value, indicating a certain degree of overfitting. As the number of data points increased, the predicted value of the XGBR model gradually approached the true value; however, there was still a large-amplitude fluctuation in the predicted value. In contrast, the PCA\_SER(XGBR) and PLSR\_SER(XGBR) models performed more smoothly on the training set, especially the PCA\_SER(XGBR) model, which better captured the true values of the trend.

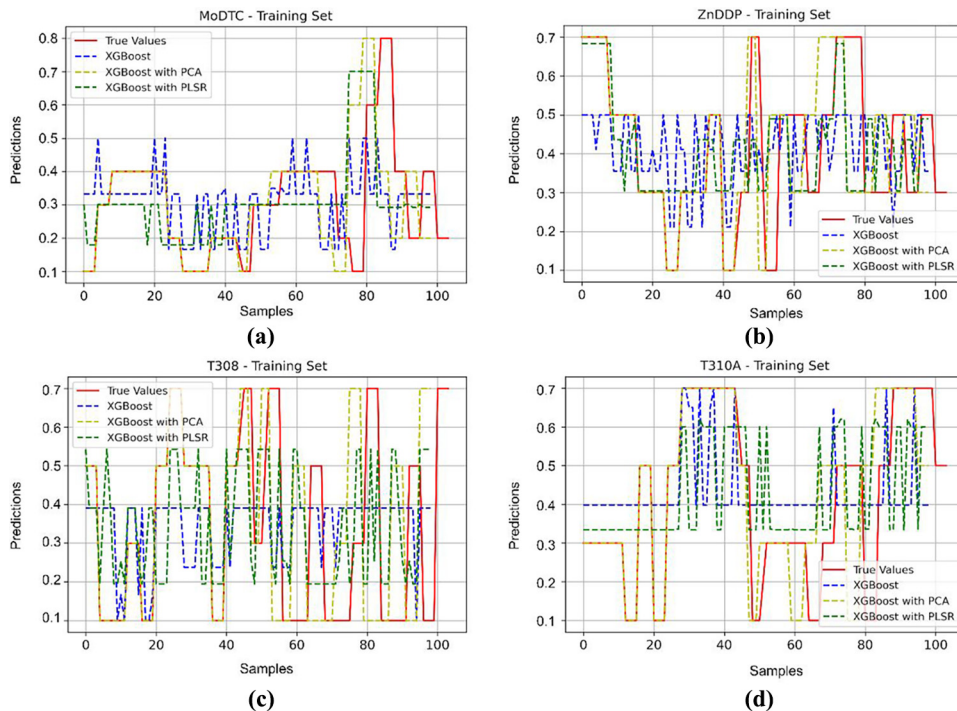
In the ZnDDP data set, the actual value curve showed a relatively stable range of fluctuations (0.004–0.008), whereas the model-predicted performance varied. The XGBR model fluctuated significantly in the early data points (0–40), and

Figure 6 The pair plot of the visualization tool for exploring the relationship between multiple variables



Source: Authors' own work

Figure 7 Performance analysis of predictive models for (a) MoDTC, (b) ZnDDP, (c) T308 and (d) T310A



Source: Authors' own work

some areas deviated significantly from the true value. In contrast, the PCA\_SER(XGBR) and PLSR\_SER(XGBR) models can get closer to the true value for most data points, particularly in the later period (60–100), where the predictions are more accurate. In addition, the T308 data set exhibited complex and irregular fluctuations in the true values, ranging from 0.002–0.008. Among all models, the XGBR model prediction curve fluctuated more and did not follow the true value. The performance of the PLSR\_SER(XGBR) model was slightly improved, and some areas overlapped well with the actual values. PCA\_SER(XGBR) exhibited a smoother trend, better capturing the overall fluctuation characteristics and improving the model's ability to fit complex and fluctuating data sets.

The training results of the 310A data set also showed performance differences between the different models. The fluctuation range of the true value was 0.004–0.008, which was relatively stable in the early stage (0–40) and fluctuated significantly in the later stage. PCA\_SER(XGBR) and PLSR\_SER(XGBR) performed more stably in the last stage, especially the PCA\_SER(XGBR) method, which effectively followed the fluctuation trend of the true values. Based on the results of the four data sets, the XGBR model was prone to overfitting or large prediction fluctuations in complex data areas. Simultaneously, the improved version combining PCA can effectively improve the fitting effect of the model, particularly on data sets with high feature complexity.

Figure 8 provides a comprehensive comparison of the predictive efficacy of each model, demonstrating the extent of correspondence between the predicted and actual experimental results for each lubricant component. The predictions from the test set indicate that the PCA\_SER(XGBR) model has a superior alignment with the actual values compared to the other two models. The PCA\_SER(XGBR) model mitigates noise and removes superfluous features by converting the input data into a condensed collection of PCs, emphasizing the most

critical patterns and connections. Comparable trends were observed for ZnDDP, T308 and T310A, where the predictions from the PCA\_SER(XGBR) model demonstrated improved stability and greater alignment with the actual values.

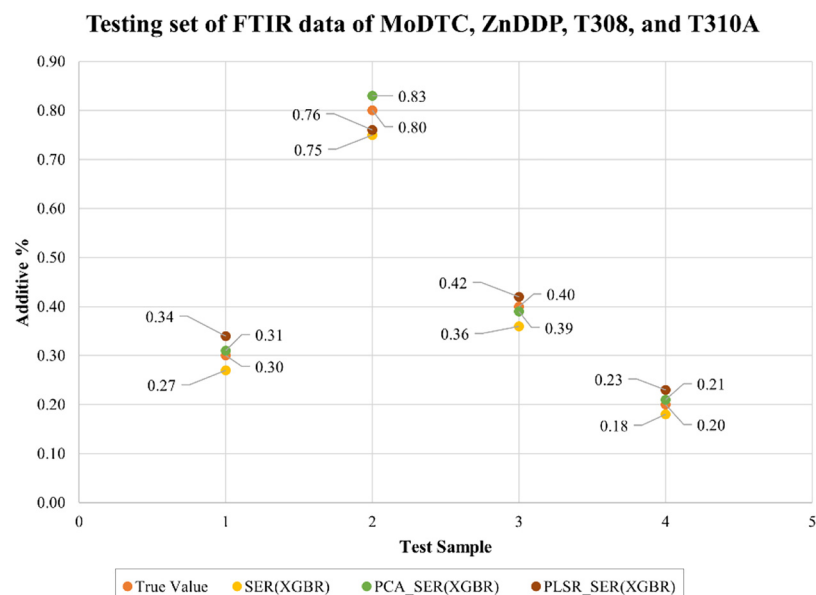
Figure 8 shows the FTIR spectra of MoDTC, ZnDDP, T308 and T310A additives. The graph contrasts the actual values of these additives with the forecasts produced by the three models:

- 1 SER(XGBR);
- 2 PCA\_SER(XGBR); and
- 3 PLSR\_SER(XGBR).

In the MoDTC data set, the PCA\_SER(XGBR) model predictions (0.0031) were closely aligned with the true value (0.0030), indicating a strong predictive accuracy. Similarly, the PLSR\_SER(XGBR) model prediction (0.0034) was slightly higher but remained close to the actual prediction. However, the SER(XGBR) prediction (0.0027) was slightly lower, indicating minor underestimation. For the ZnDDP data set, predictions from the PCA\_SER(XGBR) model (0.0083) were slightly higher than the true value (0.0080), whereas the PLSR\_SER(XGBR) model prediction (0.0076) was closely aligned with the underestimation. The SER(XGBR) model prediction (0.0075) was marginally lower than that of the

For T308, the SER(XGBR) model (0.0036) exhibited a modest underestimation. The PCA\_SER(XGBR) and PLSR\_SER(XGBR) models demonstrated similar accuracies, with values of 0.0039 and 0.0040, respectively, which were closely aligned with the true value (0.004). For T310A, the PCA\_SER(XGBR) model prediction (0.0021) and the PLSR\_SER(XGBR) prediction (0.0023) were both very close to the true value (0.0020). The SER(XGBR) prediction (0.0018) again indicates a slight underestimation. The PCA\_SER(XGBR) model demonstrated superior stability and accuracy, effectively capturing the critical patterns in the data after dimensionality reduction.

Figure 8 Testing set of unknown ingredients FTIR data of MoDTC, ZnDDP, T308 and T310A



Source: Authors' own work

In the T308 and T310A results, the prediction results of the XGBR model were significantly lower than the true values, indicating that the original model was underfitting this data set and failed to effectively capture the feature distribution in the data Table 3.

The PCA\_XGBR model improved the prediction accuracy, was almost consistent with the true value and showed superior adaptability to data sets with high feature complexity.

Figure 9 shows that on the MoDTC data set, the performances of the original XGBR model and the model

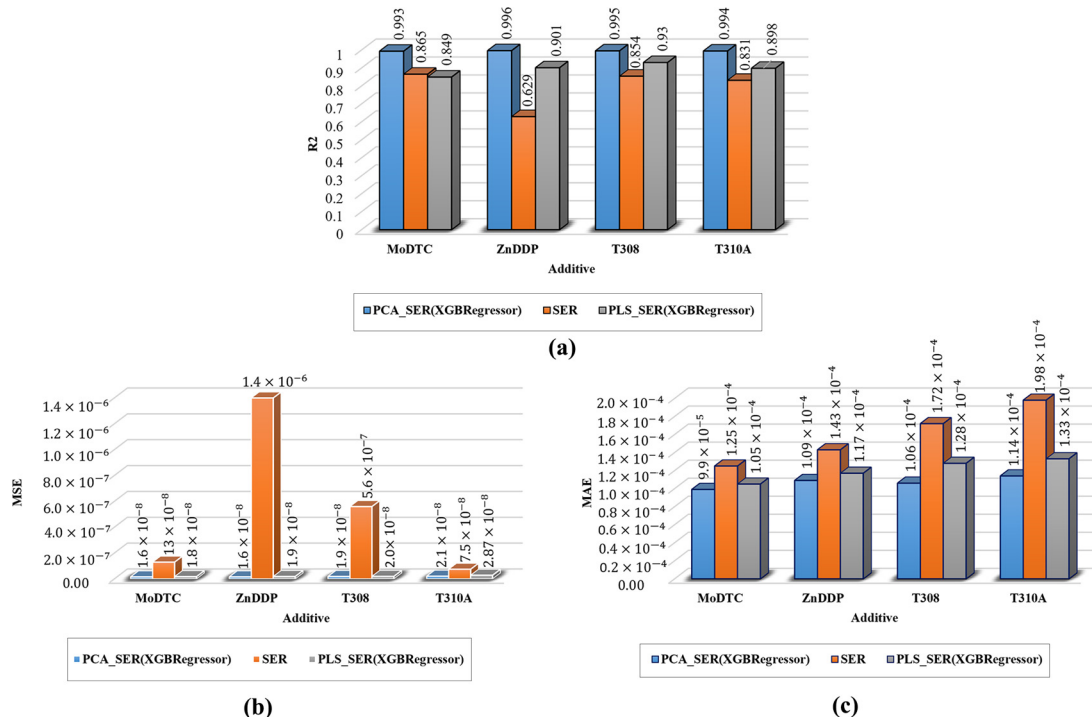
combined with PCA were highly consistent, both of which were close to the true value of 0.003. This shows that PCA's dimensionality reduction processing of the XGBR model by PCA under more data features does not cause information loss but helps the model maintain stable performance. Next, we observed the ZnDDP data set. The true and predicted values of each model were concentrated at 0.0052. All models can accurately predict, but XGBR combined with PCA is closer to the true value than the original XGBR model. Comparatively, the SER model exhibited more significant  $R^2$ , mean absolute

Table 3 Performance metrics of predictive models for different ingredients

Ingredients	MoDTC	ZnDDP	T308	T310A
<b>PCA_SER(XGBRegressor)</b>				
$R^2$	0.993	0.996	0.995	0.994
MSE	$1.6 \times 10^{-8}$	$1.6 \times 10^{-8}$	$1.9 \times 10^{-8}$	$2.1 \times 10^{-8}$
MAE	$9.9 \times 10^{-5}$	$1.09 \times 10^{-4}$	$1.06 \times 10^{-4}$	$1.14 \times 10^{-4}$
<b>SER</b>				
$R^2$	0.865	0.629	0.854	0.831
MSE	$1.3 \times 10^{-8}$	$1.4 \times 10^{-6}$	$5.6 \times 10^{-7}$	$7.5 \times 10^{-8}$
MAE	$1.25 \times 10^{-4}$	$1.43 \times 10^{-4}$	$1.72 \times 10^{-4}$	$1.98 \times 10^{-4}$
<b>PLS_SER(XGBRegressor)</b>				
$R^2$	0.849	0.901	0.930	0.898
MSE	$1.8 \times 10^{-8}$	$1.9 \times 10^{-8}$	$2.0 \times 10^{-8}$	$2.87 \times 10^{-8}$
MAE	$1.05 \times 10^{-4}$	$1.17 \times 10^{-4}$	$1.28 \times 10^{-4}$	$1.33 \times 10^{-4}$

Source(s): Authors' own work

Figure 9 Performance metrics of predictive models



Note(s): (a)  $R^2$ ; (b) MSE; (c) MAE

Source: Authors' own work

error (MAE) and mean squared error (MSE) values than the PCA\_SER(XGBR) model.

This indicates that the predictions made by the SER model were not accurate. The PLS\_SER (XGBR) model performs well in Figure 8, with  $R^2$  values ranging from 0.898 to 0.930 and somewhat higher MSE and MAE values than the PCA\_SER(XGBR) model. However, the PCA\_SER (XGBR) model surpassed its accuracy. The PCA\_SER(XGBR) model can capture complex information structures, providing more exact and accurate predictions.

#### 4. Conclusion

This study proposes a novel approach for identifying and forecasting lubricant additives by combining PCA with a stacked ensemble regression model (XGBR). By applying advanced machine learning algorithms and dimensionality reduction technology, this approach effectively uses FTIR spectroscopy data to resolve the complexity of high-dimensional data sets, offering a novel approach for studying lubricant additives. In wind power applications, this method can optimize the performance and durability of wind turbine mechanical equipment by ensuring the proper lubricant composition, reducing friction and preventing unexpected failure:

- The results show that PCA can effectively reduce the dimensionality of the data while retaining more than 90% of the data variance, thereby significantly reducing redundant information and noise. This feature not only improves the interpretability of the data but also enhances the efficiency and accuracy of the subsequent modeling process. This method successfully captures the nonlinear characteristics of the data by building multiple basic regression models and using meta-learners to optimize the prediction results.
- The PCA\_XGBR model performed excellently in predicting lubricant additives (including MoDTC, ZnDDP, T308 and T310A). The  $R^2$  values of this model both exceed 0.98 and the MSE and MAE are significantly lower than those of traditional methods, indicating their superior performance on complex datasets.
- The integration of PCA and Stacked Ensemble Regression (XGBR) provides an advanced and effective approach for identifying and predicting lubricant components using FTIR spectroscopy. The PCA component efficiently reduces the dimensionality of the data, addressing high dimensionality and multicollinearity challenges while preserving the essential variance.

#### References

- Agulei, K.D., Githaiga, J.T., Dulo, B. and Nganyi, E.O. (2023), "Identification of bioactive compounds from onion (*Allium burdickii*) bulb using Raman, and FTIR spectroscopy", *Research Journal of Textile and Apparel*, doi: [10.1108/RJTA-07-2023-0070](https://doi.org/10.1108/RJTA-07-2023-0070).
- Arachchilage, C.B., Huang, G., Zhao, J., Fan, C. and Liu, W.V. (2024), "Hybrid extreme gradient boosting regressor models for the multi-objective mixture design optimization of cementitious mixtures incorporating mine tailings as fine aggregates", *Cement and Concrete Composites*, Vol. 154, p. 105787.
- Awasthi, M., Joshi, V., Upadhyay, R., Kukrety, A., Verma, A.K. and Kumar, P. (2024), "Development of petroleum-derived polymeric additive to enhance the bituminous properties with the use of a machine-learning model", *Sustainable Chemistry for the Environment*, Vol. 8, p. 100186.
- Chu, L., Guo, C., Zhang, Q., Wang, Q., Ge, Y. and Hao, M. (2024), "Differentiation of multilayered automotive coatings with Fourier transform infrared spectroscopy, Raman spectroscopy and scanning electron microscope/energy dispersive xray spectrometer", *Pigment & Resin Technology*, Vol. 53 No. 1, pp. 36-43.
- Garcia, C.E., Ueda, M., Spikes, H. and Wong, J.S. (2021), "Temperature dependence of molybdenum dialkyl dithiocarbamate (MoDTC) tribofilms via time-resolved Raman spectroscopy", *Scientific Reports*, Vol. 11 No. 1, p. 3621.
- Greenacre, M., Groenen, P.J.F., Hastie, T., D'Enza, A.I., Markos, A. and Tuzhilina, E. (2022), "Principal component analysis", *Nature Reviews Methods Primers*, Vol. 2 No. 1, p. 100.
- Hamnas, A. and Unnikrishnan, G. (2023), "Bio-lubricants from vegetable oils: characterization, modifications, applications and challenges – review", *Renewable and Sustainable Energy Reviews*, Vol. 182, p. 113413.
- Heen Blindheim, F. and Ruwoldt, J. (2025), "Quantifying the abundance of alkane moieties in lignins with FTIR spectroscopy and PLS regression; estimating grafting degree of esterification", *ChemSusChem*, Vol. 18 No. 3, p. e202400938.
- Heinrich, F., Ghaeni, H., Erz, R., Schmidt, C., Esch-Letica, E. v. d. and Moll, S. (2024), "A predictive maintenance concept for lubricant oil usage", *Proc. IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society*, 3-6 Nov. 2024 2024.
- Jaoussse, M., Elhadri, Z., Hanafi, M., Dolce, P. and Elkettani, Y. (2020), "Properties of the correlation matrix implied by a recursive path model using the finite iterative method", *Electronic Journal of Applied Statistical Analysis*, Vol. 13 No. 2, pp. 369-375.
- Jin, X., Li, S., Zhang, W., Zhu, J. and Sun, J. (2020), "Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms", *Applied Sciences*, Vol. 10 No. 4, p. 1520.
- Li, J., Guo, X., Panchal, B., Wang, J., Guo, W. and Liu, B. (2024), "Quantitative analysis of molecular structure characterization of different liptinite-rich coals using FTIR spectroscopy", *Infrared Physics & Technology*, Vol. 141, p. 105458.
- Maboudou-Tchao, E.M. (2020), "Change detection using least squares one-class classification control chart", *Quality Technology & Quantitative Management*, Vol. 17 No. 5, pp. 609-626.
- Mayerhöfer, T.G., Ilchenko, O., Kutsyk, A. and Popp, J. (2022), "Infrared spectroscopy of quasi-ideal binary liquid mixtures: the challenges of conventional chemometric regression", *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 280, p. 121518.
- Peng, H., Zhang, H., Shangguan, L. and Fan, Y. (2022), "Review of tribological failure analysis and lubrication technology research of wind power bearings", *Polymers*, Vol. 14 No. 15, p. 3041.
- Perera, K.D.C., Weragoda, G.K., Haputhanthri, R. and Rodrigo, S.K. (2021), "Study of concentration dependent curcumin interaction with serum biomolecules using ATR-FTIR spectroscopy combined with principal component

- analysis (PCA) and partial least square regression (PLS-R)", *Vibrational Spectroscopy*, Vol. 116, p. 103288.
- Pristyanto, Y., Mukarabiman, Z. and Nugraha, A.F. (2023), "Extreme gradient boosting algorithm to improve machine learning model performance on multiclass imbalanced dataset", *JOIV: International Journal on Informatics Visualization*, Vol. 7 No. 3, pp. 710-715.
- Shin, K.-S., Kim, E.-Y., Lee, S.-J., Lee, D.-S. and Paeng, K.-J. (2022), "A method for the determination of ZnDDP in lubricant oils by applying solid-phase extraction with mixed resin", *Journal of Analytical Science and Technology*, Vol. 13 No. 1, pp. 17-29.
- Singh, S., Chaubey, D.S., Raj, R., Kumar, V., Paliwal, M. and Mahlawat, S. (2025), "Social media communication, consumer attitude and purchase intention in lifestyle category products: a PLS-SEM modeling", *Marketing Intelligence & Planning*, Vol. 43 No. 2, pp. 272-296.
- Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M. and Liang, X. (2024), "An improved random Forest based on the classification accuracy and correlation measurement of decision trees", *Expert Systems with Applications*, Vol. 237, p. 121549.
- Wang, L., Wang, X., Li, L., Yang, C. and Zhu, Y. (2025), "Research on the tribological properties of kaolin/MoDDP composite lubricant additives under heavy load and impact conditions", *Industrial Lubrication and Tribology*, Vol. 77 No. 2, pp. 300-308.
- Wen, G., Liu, W., Wen, X., Wei, P., Cao, H. and Bai, P. (2024), "Effective tribological performance-oriented concentration optimization of lubricant additives based on a machine learning approach", *Tribology International*, Vol. 197, p. 109770.
- Xia, Y., Cao, Y. and Feng, X. (2021), "A comparative study on the electrical and tribological characteristic of magnetron sputtered Ag, Cu and Al films under current-carrying friction", *Industrial Lubrication and Tribology*, Vol. 73 No. 10, pp. 1219-1225.
- Xia, Y., Chen, W., Chang, X. and Feng, X. (2024a), "Study on the tribological properties of acid-doped polypyrrole (PPY) as conductive grease additive", *Tribology Transactions*, Vol. 67 No. 4, pp. 754-764.
- Xia, Y., Chen, W., Zhang, Y., Yang, K. and Yang, H. (2024b), "Enhanced tribological properties of sliding contacts through the synergistic effect of PTFE film and PSAIL 2280", *Industrial Lubrication and Tribology*, Vol. 76 No. 10, pp. 1236-1245.
- Xia, Y., Wang, C. and Feng, X. (2022), "GA-BPSO hybrid optimization of Middle infrared spectrum feature band selection of lubricating oil additive type identification technology", *Mocaxue Xuebao/Tribology*, Vol. 42 No. 1, pp. 142-152.
- Xia, Y., Xu, D., Feng, X. and Cai, M. (2020), "Identification and content prediction of lubricating oil additives based on extreme learning machine", *Tribology*, Vol. 40 No. 1, pp. 97-106.
- Xia, Y., Zou, S., Xie, P. and Feng, X. (2024c), "A kind of multi-dot ensemble regression AI detector for lubricating oil additive content based on lambert-beer law", *Spectrochimica*

- Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 318, p. 124436.
- Xu, J., Liu, S., Gao, M. and Zuo, Y. (2023), "Classification of lubricating oil types using mid-infrared spectroscopy combined with linear discriminant analysis-support vector machine algorithm", *Lubricants*, Vol. 11 No. 6, pp. 268-271.
- Zhu, M., Wu, T., Chen, Y. and Zhu, H. (2024), "Filter-processing-assisted Fourier transform Raman spectroscopy for forensic discrimination lubricant machinery oils", *Microchemical Journal*, Vol. 204, p. 111011.

## Further reading

- Xia, Y., Wang, C., Feng, X., *et al.* (2023a), "Prediction of friction and wear performance of lubricating oil based on GRNN optimized by GWO", *Tribology*, Vol. 43 No. 8, pp. 947-955.
- Xia, Y., Wang, Y., Feng, X., *et al.* (2023b), "Optimization efficiency of swarm intelligence search in base oil performance prediction model", *Tribology*, Vol. 43 No. 4, pp. 429-438.

## Supplementary material

The supplementary material for this article can be found online.

## About the authors



**Nay Min Aung** received his master's degree from the North China Electric Power University, School of Energy Power and Mechanical Engineering, in 2023. He is now studying for a PhD at North China Electric Power University, mainly in current-carrying friction research.



**Yanqiu Xia** received his PhD degree in mechanical engineering from Northeastern University, China, in 1999 and was selected as a professor in 2007. He joined the School of Energy Power and Mechanical Engineering at North China Electric Power University in 2010. His current position is as a professor. His research areas cover mechanical and electrical equipment tribology, oil monitoring and artificial intelligence. Yanqiu Xia is the corresponding author and can be contacted at: [jve@ncepu.edu.cn](mailto:jve@ncepu.edu.cn)



**Xin Feng**: She received her PhD degree in systems engineering from Northeastern University, China, in 2007. She joined the School of Energy Power and Mechanical Engineering at North China Electric Power University, China, in 2010. She is currently an Associate Professor. Her research areas cover tribology and artificial intelligence.