

Predicting density and viscosity and identifying key features for biodiesel-alcohol blends based on explainable machine learning models

WANG Jianan¹, SHI Jiahao¹, BI Shengshan^{1,3,*}, AN Qingsong², GAO Kehui¹, CUI Junwei⁴

¹ Key Laboratory of Thermo-Fluid Science and Engineering, Ministry of Education, School of Energy and Power Engineering, Xi'an Jiaotong University, Xi'an 710049, China;

² Laboratory of Efficient Utilization of Low and Medium Grade Energy, School of Mechanical Engineering, Tianjin University, Tianjin 300350, China;

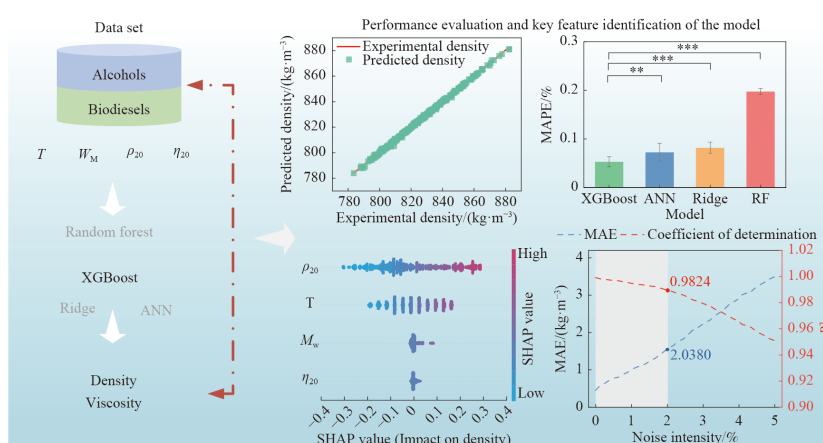
³ Shanxi Provincial Key Laboratory of High Efficiency Heat Storage and Low Carbon Heat Supply, Taiyuan 030021, China;

⁴ Key Laboratory of New Energy Development, Application and Automotive Energy Conservation in Shaanxi Province, Chang'an University, Xi'an 710064, China

Abstract: Accurate prediction of the thermophysical properties of biodiesel-alcohol blends is essential for spray combustion simulation when used as alternative marine fuels. To address the limitations of existing prediction models in terms of generalization capability and interpretability, this study systematically evaluated the predictive performance and applicability of Ridge regression, Artificial Neural Network

(ANN), Random Forest and XGBoost models for estimating the density and viscosity of the blends. The results indicated that the XGBoost model outperforms the other three models, achieving mean absolute percentage errors (MAPE) of 0.05% for density and 1.89% for viscosity on unseen test data. Statistical tests confirmed the significantly superior performance of this model. Furthermore, interpretability analysis based on SHapley Additive exPlanations (SHAP) disclosed that, in the XGBoost model, initial density (ρ_{20}) and experimental temperature (T) are the key inputs for density prediction, while T and initial viscosity (η_{20}) are the most influential for viscosity prediction. Finally, robustness testing demonstrated that the XGBoost model maintains excellent stability and generalization capability with input noise levels up to 2%, thereby providing high-precision data support for numerical simulation of in-cylinder spray characteristics in internal combustion engines.

Key words: data-driven; machine learning; biodiesel-alcohol blends; density; viscosity



Biodiesel-alcohol blends are regarded as one of the most promising alternative marine fuels before 2050^[1-2], due to their advantages in storage convenience, widespread availability, mature synthesis techniques, and outstanding

compatibility with existing engines. Previous researches have shown that blending biodiesel with diesel or alcohols enhances its atomization and combustion performance. For instance, Cho et al.^[3] summarized that adding alcohol

Received: 2025-12-03; Revised: 2026-01-10

Foundation items: Supported by National Natural Science Foundation of China (51776171), Shaanxi Qinchuangyuan 'Scientists + Engineers' Team Construction Project (2023KXJ-018) and Shaanxi Provincial Natural Science Foundation (General Project (Youth)) (2024JC-YBQN-0589).

* Corresponding author. E-mail: bss@mail.xjtu.edu.cn.

DOI: 10.1016/S1872-5813(26)60706-2 CSTR: 32157.14.2097-213X.2026.JFCT.0065 Associated Data DOI: 10.57760/sciencedb.j00124.00435

本文的英文电子版由 Elsevier 出版社在 ScienceDirect 上出版 (<http://www.sciencedirect.com/science/journal/18725813>)

to biodiesel blends improves fuel atomization, extends ignition delay, and increases premixed combustion efficiency, which greatly improves engine performance. As the critical processes affecting combustion efficiency and emission characteristics, the fuel injection and atomization depend on accurate density and viscosity, which are not only the basis for calculating atomization parameters (such as Reynolds, Weber, and Laplace numbers) but also the critical inputs for developing in-cylinder spray combustion models and evaluating pollutant generation^[4–5]. Mestila^[6] suggested that the density and viscosity of rapeseed oil-based biodiesel-bioethanol ternary blends are key determinants of atomization and combustion characteristics in diesel engines.

Regarding the properties of biodiesel-alcohol blends, existing studies on density have covered a wide range of temperature and pressure, while experimental research on viscosity remain mainly confined to atmospheric pressure conditions^[7–11], significantly hindering its application in high-pressure combustion simulation. Against the backdrop of rapid advancements in novel alternative fuel synthesis and artificial intelligence technologies, obtaining thermophysical property data for different fuels and their blend systems under wide ranges of temperature, pressure, and composition conditions through experimental methods requires considerable time and cost. Therefore, there is an urgent need to develop a machine learning prediction model with high computational accuracy for thermophysical properties to meet the requirements of applications such as simulation of high-pressure spray and combustion performance in internal combustion engine cylinders. Yadav et al.^[12] developed a density prediction model for biodiesel-ethanol blends using three machine learning methods (ANN, Support Vector Machine, and Radial basis function network), achieving accurate density predictions of the blend system under high pressure, and identified the main influence factors of the model through sensitivity analysis. Belmadani et al.^[13] used ANN to predict the density and viscosity, enabling high-accuracy predictions of thermophysical properties for novel blends without experimental process. Based on the investigations on the relationship between viscosity, composition, and temperature of diesel-biodiesel-bioethanol blends, Bietresato et al.^[14] developed a model exhibiting the significantly improved coefficient of determination (R^2) at high temperatures. Yesilova et al.^[15] suggested density and viscosity prediction models for diesel-biodiesel-alcohol blends by employing multiple machine learning methods, yet the temperature range of their research was confined to 283.15–313.15 K. The aforementioned models exhibit limitations in component diversity, temperature and pressure ranges and multi-output capabilities for biodiesel-alcohol blends.

In this study, based on an extensive compilation of experimental data on the density and viscosity of biodiesel, diesel, and alcohol blends, four different machine learning models were first developed to estimate the density and

viscosity of biodiesel-alcohol blends. Afterwards, the predictive performance and computational efficiency of these four models were systematically assessed to determine the optimal model for predicting density and viscosity. Finally, the SHAP interpretability analysis method was introduced to identify the models' decision mechanisms and critical input features, and comprehensive evaluations of model robustness were thoroughly conducted using input noise interference tests and multi-fold cross-validation.

1 Modeling and evaluation methods

1.1 Data collection and processing

The precision of experimental thermophysical data has a direct impact on the accuracy of machine learning-based prediction models. This study compiled a total of 2262 experimental data points on the density and viscosity of biodiesel, diesel, and alcohol blends. Table 1 summarizes the temperature and pressure ranges, number of experimental points, and the uncertainties associated with the experimental methods, where the density and viscosity were determined using the vibrating tube method and the capillary method, respectively. To ensure the diversity and representativeness of data for model development, the compiled dataset contained key parameters including biodiesel composition, alcohol type, and alcohol molar fraction. Additionally, the Z-score method was applied to identify and remove outliers from the experimental data.

1.2 Input features

Accurate feature inputs are critical for elucidating the physical principles and intrinsic relationships within the data^[22]. For deep learning models predicting thermophysical properties at atmospheric pressure, experimental temperature (T) is an essential input parameter. Given the direct influence on molecular volume and intermolecular interactions, the molecular weight (M_w) of the blends and the molar fraction of alcohols were also included as model input variables. Moreover, to reduce the impact of isomers on model predictions, the initial density (ρ_{20}) and viscosity (η_{20}) of the blends at 20 °C and atmospheric pressure were provided as additional feature inputs. To prevent multicollinearity from compromising model stability and interpretability, Pearson correlation coefficients were used to examine the dependences among input features, as shown in Figure 1.

A strong negative correlation is observed between M_w and alcohol molar fraction (x) in Figure 1, where increasing x leads to a substantial decrease in M_w with this correlation strength reaching the threshold for high multicollinearity. Incorporating both x and M_w simultaneously as independent input variables may result in unreliable parameter estimates and decreasing model generalization capabilities.

Table 1 Experimental data on density and viscosity of biodiesel-alcohol blends

Composition of the blend		T/K	p/MPa	Uncertainty		Data
Fuel	Alcohol			$\rho/(\text{kg}\cdot\text{m}^{-3})$	$\eta/\%$	
Biodiesel	1-pentanol	288.15–338.15	0.1	0.6	1	330 ^[8]
	2-pentanol					
	2-methyl-1-butanol					
Methyl octanoate	1-propanol	293.15–333.15	0.1	0.6	1.3	234 ^[9]
	1-butanol					
	1-pentanol					
	1-hexanol					
	1-heptanol					
Diesel	1-butanol	288.15–323.15	0.1	0.8	0.35	288 ^[10]
	1-pentanol					
	1-octanol					
Biodiesel	1-butanol	293.15–333.15	0.1	0.6	1	270 ^[16]
	isobutyl alcohol					
	2-butanol					
Biodiesel	ethanol	293.15–323.15	0.1	0.8	0.35	47 ^[17]
Methyl myristate	1-propanol	303.15–333.15	0.1	0.8	2	140 ^[18]
	1-butanol					
	1-pentanol					
Biodiesel	1-hexanol	288.15–338.15	0.1	0.6	1	330 ^[19]
	4-methyl-2-pentanol					
	2-ethyl-1-butanol					
Biodiesel	1-octanol	288.15–338.15	0.1	0.6	1	363 ^[20]
Diesel	2-octanol	288.15–323.15	0.1	0.8	0.35	160 ^[21]
Diesel	2-ethyl-1-hexanol					
	<i>n</i> -butanol					
	<i>n</i> -pentanol					
diethylene glycol dimethyl ether						

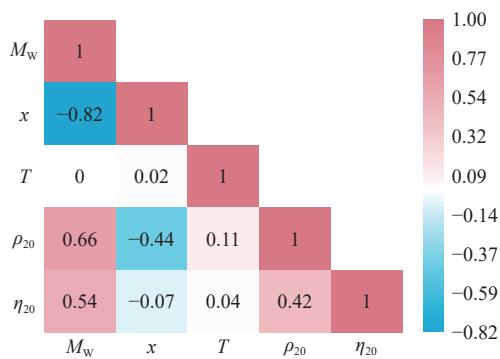


Fig. 1 Heatmap of input feature correlation matrix

The correlations among the remaining feature pairs range from moderate to low, consistent with the physical principles governing density and viscosity variations caused by increased M_w in biofuel blends, where no severe collinearity was observed in the correlation strength. Furthermore, the correlation coefficients between T and all other features are close to zero, demonstrating that temperature is statistically independent of other physical parameters. Accordingly, T , M_w , ρ_{20} and η_{20} were selected as feature inputs. All feature input data were compiled from the literature. For cases where ρ_{20} and η_{20} data were partially

unavailable from literature, these values were derived by fitting the available literature data to the Tait equation.

$$\rho(T, p) = \frac{\rho_0(T)}{1 - C \ln((B(T) + p)/(B(T) + 0.1 \text{ MPa}))} \quad (1)$$

where $\rho_0(T)$ and $B(T)$ are temperature-dependent fitting parameters, with their functional forms given as follows:

$$\rho_0(T) = A_0 + A_1 T + A_2 T^2 + A_3 T^3 \quad (2)$$

$$B(T) = B_0 + B_1 T + B_2 T^2 \quad (3)$$

where A_i , B_i and C are fitting parameters and the units for temperature, pressure, and density are K, MPa and kg/m^3 , respectively. The Levenberg-Marquardt optimization algorithm was employed for fitting process.

1.3 Modeling approach

Biodiesel-alcohol blends are composed of components with substantial differences in chemical structure and polarity. The complex coupling of intermolecular forces, such as van der Waals interactions and hydrogen bonding, leads to density and viscosity behaviors characterized by

pronounced nonlinearity, strong compositional dependence, and significant temperature sensitivity. Due to their powerful fitting capability and data-driven nature, machine learning methods show remarkable advantages for such prediction tasks. To evaluate the adaptability of different machine learning models in predicting the density and viscosity of biodiesel-alcohol blends, this study employed four models representing linear, neural network, and decision tree architectures: Ridge Regression, Random Forest, ANN and XGBoost. The modeling workflow is illustrated in Figure 2.

1.3.1 Ridge model

Ridge regression is a technique that introduces L2 regularization into least squares estimation, effectively mitigating collinearity in data. It is particularly useful for addressing the common multicollinearity issues encountered in biodiesel blend studies. The interpretability of its coefficients plays a key role in understanding the physical property formation mechanisms of such blends.

1.3.2 ANN model

ANN is a computational model inspired by the structure and functioning of biological neural systems. It is capable of capturing highly complex nonlinear relationships through interconnected layers of processing units^[23]. The architecture of the model typically comprises an input layer, one or more hidden layers, and an output layer. In this study, the hidden layers employ the hyperbolic tangent (tanh) activation function, and training is performed using the Adam optimization algorithm. To mitigate overfitting and improve generalization, regularization techniques and an early stopping mechanism are incorporated into the training process.

1.3.3 Random Forest model

Random Forest is an ensemble learning method that operates by constructing multiple decision trees^[24]. It introduces randomness through feature selection during the tree-building process, which effectively reduces model

variance. This allows the method to capture complex feature interactions within blends and to identify the key factors influencing the prediction of thermophysical properties.

1.3.4 XGBoost model

XGBoost builds upon the framework of gradient-boosting decision trees^[25]. To handle the dual-output regression task in this study, a “single-output single-model” strategy was implemented. Under this approach, separate XGBoost models were trained for each output variable. Each model sequentially constructs an ensemble of weak learners, which are combined linearly to form a strong learner capable of achieving accurate predictions.

For hyperparameter tuning across the aforementioned four models, a Bayesian optimization approach based on the Optuna framework was employed. The core of this method lies in constructing a probabilistic model of the objective function to guide the search process, enabling an intelligent balance between exploration and exploitation for efficiently navigating complex parameter spaces. In this study, training strategies including early stopping and dropout regularization were incorporated into different machine learning models to enhance the stability and efficiency of the optimization process. Hyperparameter tuning for all models was aimed to optimize the objective function on the validation set. Based on literature references and empirical random search, preliminary hyperparameter experiments were conducted to define the tuning ranges for each model, as illustrated in Figure 3.

1.4 Evaluation method of model

1.4.1 SHAP explainability analysis

With the increasing complexity of machine learning models, interpretable machine learning has become essential for uncovering internal decision logic and establishing transparent relationships between model predictions and input features. The SHAP method is a technique for interpreting predictions from machine learning models, particularly effective for “black-box” models.

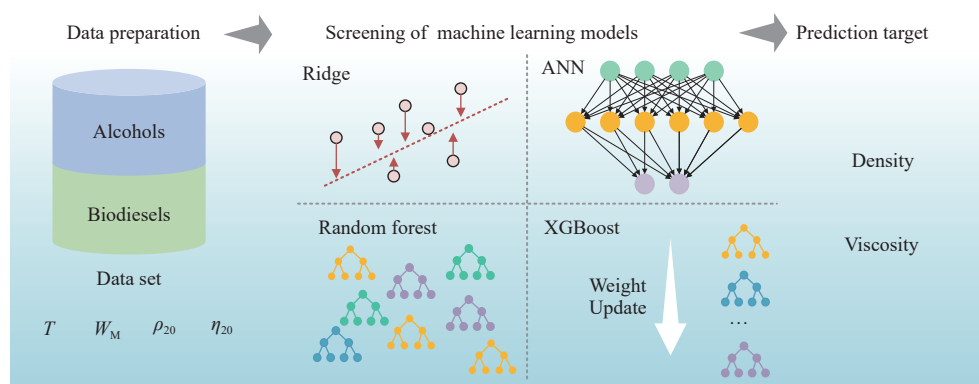


Fig. 2 Modeling process of machine learning models

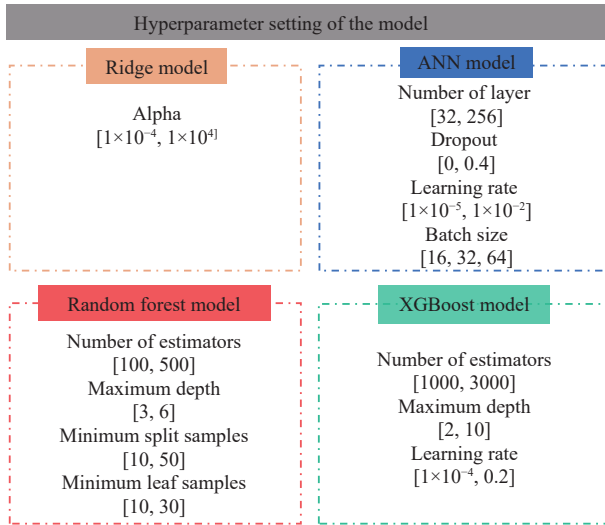


Fig. 3 Hyperparameter settings for the model

Initially proposed by mathematician Shapley in 1953^[26], its core concept is based on the Shapley value from cooperative game theory, which has been adapted to quantify the contribution of each feature to the model's predictions. For a given feature set S and a model $f(X)$, where X represents the input features and S denotes a subset containing certain features, the Shapley value for each feature is calculated by considering all possible combinations of features. The contribution of each feature to the model output across these combinations is quantified by Equation (4).

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S|)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (4)$$

where $\phi_i(v)$ represents the Shap values of feature i , $v(S)$ is the model's prediction result, $|S|$ denotes the size of set S , $|N|$ represents the size of the full feature set N and $S \subseteq N \setminus \{i\}$ indicates the combinations of features excluding feature i .

1.4.2 Predictive performance evaluation metrics

Model evaluation metrics were used to quantify the predictive performance of a model on a given dataset, facilitating the assessment of its effectiveness and generalizability. For the regression tasks, root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), Bias, and coefficient of determination (R^2) were employed as evaluation metrics. each metric is defined as follows.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{i,\text{exp}} - y_{i,\text{pre}})^2} \quad (5)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_{i,\text{exp}} - y_{i,\text{pre}}| \quad (6)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{(y_{i,\text{exp}} - y_{i,\text{pre}})}{y_{i,\text{exp}}} \right| \times 100\% \quad (7)$$

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^N \frac{(y_{i,\text{exp}} - y_{i,\text{pre}})}{y_{i,\text{exp}}} \times 100\% \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_{i,\text{exp}} - y_{i,\text{pre}})^2}{\sum_{i=1}^N (y_{i,\text{exp}} - \bar{y})^2} \quad (9)$$

where $y_{i,\text{exp}}$ is the experimental values for density and viscosity, while $y_{i,\text{pre}}$ is the predicted values for density and viscosity.

2 Results and discussion

2.1 Evaluation of models

2.1.1 Analysis of prediction accuracy

The experimental density and viscosity data compiled in Table 1 were used to train and test the different machine learning models. To ensure training stability and effectiveness, all input features and output variables were normalized and scaled to the range [0, 1], thereby eliminating the influence of dimensional differences on the training process. To comprehensively evaluate the predictive performance of different models on the density and viscosity of the blended systems, the data from References [9] and [21] in Table 1 were reserved as an independent test set. Using experimental temperature as the stratification variable, the remaining dataset was divided into nine mutually exclusive strata and subsequently split into a training set (70%) and a validation set (30%) where the latter one was used to monitor early stopping criteria. The test set, which was not involved in either model training or validation, served as completely unseen data for final performance assessment. During the full training process of the four machine learning models, a stratified sampling mechanism was adopted for the dataset. In the hyperparameter optimization phase, a stratified K-fold cross-validation approach was employed to ensure that the model parameters were independent of the specific data split. The optimal hyperparameters obtained after model training are presented in Table 2.

To accurately evaluate the predictive performance of the models on the thermophysical property data, the prediction results of density and viscosity predictions of the four models for biodiesel-alcohol blends on the test set are presented in Figures 4 and 5.

As shown in Figure 4, the XGBoost model demonstrates high predictive accuracy across all datasets, exhibiting both strong generalization capability and precision for density prediction of the blends. In contrast, the prediction accuracy of the ANN and Ridge models deteriorates for densities

below 800 kg/m^3 . The Random Forest model shows the poorest among the four models, with noticeable deviations between its predicted values and experimental data on the test set, indicating limited generalization ability. Overall, the XGBoost model exhibits the best performance and is well-suited for predicting the density of biodiesel-alcohol blends.

Figure 5 compares the viscosity prediction performance of the four models. The predicted values from the XGBoost model align closely with the experimental data, demonstrating superior predictive capability and significantly outperforming the other models. While the ANN and RF models capture the overall trends in reasonable agreement with the experimental values, their generalization ability exhibits some instability. The Ridge model, however, fails to adequately capture the nonlinear characteristics of viscosity, resulting in systematic prediction bias. A comprehensive comparison of the evaluation metrics for density and viscosity predictions is summarized in Table 3, where the XGBoost model consistently achieves the best performance across the training, validation, and test sets.

2.1.2 Statistical significance analysis

A paired *t*-test was employed to quantitatively evaluate the differences in predictive performance and statistical significance among different models. The results of the significance tests for the MAPE and MAE metrics of each model are shown in Figure 6.

Model	Optimal parameters
Ridge	Alpha=1.001
Random Forest	N_estimators=250 Max_depth=6 Min_samples_split=19 Min_samples_leaf=10
ANN	Layers=[204] Dropout=0.0185 Learning rate=0.001478 Batch size=32
XGBoost	Density Learning rate=0.1227 Max depth=3 N estimators=2800 Viscosity Learning rate=0.1170 Max depth=3 N estimators=2400

As shown in Figure 6, the XGBoost model achieves significantly lower MAPE and MAE values than the ANN, Ridge, and RF models for both density and viscosity predictions, demonstrating its optimal predictive accuracy across the evaluation metrics. Furthermore, the performance differences between XGBoost and the other models are highly statistically significant, indicating that XGBoost strikes an excellent balance between predictive performance and computational efficiency.

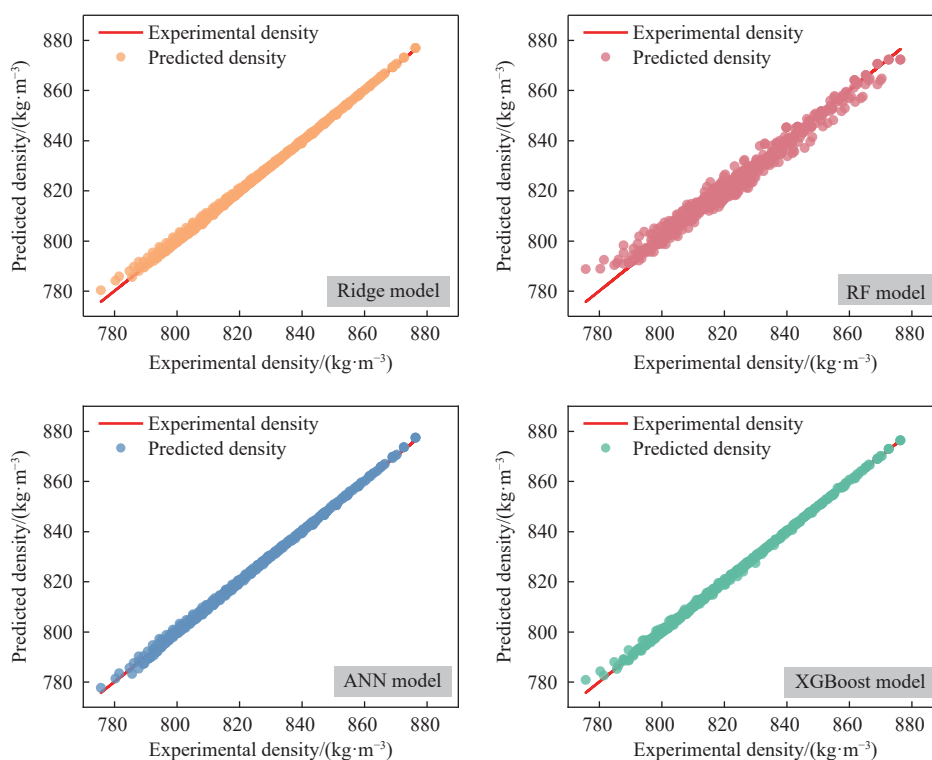


Fig. 4 Prediction deviation of density at various models

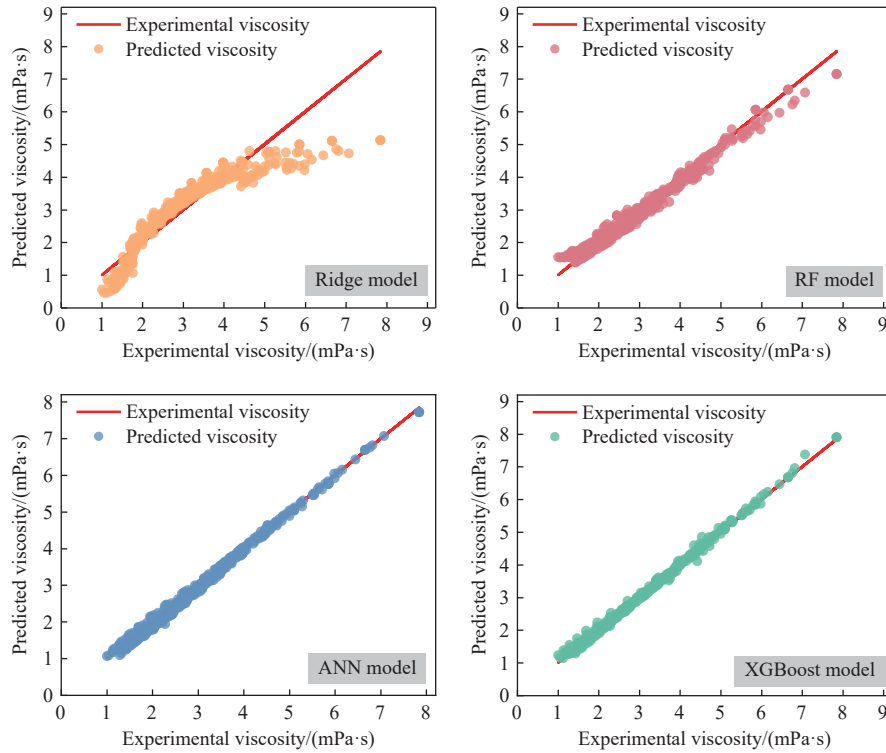


Fig. 5 Prediction deviation of viscosity at various models

Table 3 Evaluation results of models on various datasets

Evaluation Metrics	Dataset	Ridge	RF	ANN	XGBoost
Density					
RMSE/(kg·m ⁻³)	Training	0.98	3.17	0.72	0.43
	Validation	0.82	3.12	0.87	0.47
	Test	1.09	3.71	0.71	0.64
R^2	Training	0.998	0.974	0.999	0.999
	Validation	0.998	0.975	0.998	0.999
	Test	0.997	0.967	0.999	0.999
MAE/(kg·m ⁻³)	Training	0.63	2.38	0.45	0.31
	Validation	0.53	2.29	0.54	0.33
	Test	0.67	2.69	0.51	0.38
MAPE/%	Training	0.08	0.28	0.05	0.04
	Validation	0.06	0.27	0.07	0.04
	Test	0.08	0.33	0.06	0.05
Viscosity					
RMSE/(mPa·s)	Training	0.436	0.230	0.08	0.047
	Validation	0.411	0.230	0.08	0.050
	Test	0.369	0.230	0.11	0.067
R^2	Training	0.890	0.967	0.996	0.999
	Validation	0.890	0.969	0.996	0.999
	Test	0.911	0.967	0.992	0.997
MAE/(mPa·s)	Training	0.303	0.146	0.053	0.033
	Validation	0.302	0.151	0.053	0.037
	Test	0.273	0.170	0.091	0.045
MAPE/%	Training	14.29	6.69	2.65	1.446
	Validation	13.99	6.75	2.64	1.63
	Test	13.04	6.31	3.75	1.89

Based on the comprehensive performance evaluation of the four models with distinct architectures, the XGBoost model demonstrates clear advantages over the other three models in predicting the density and viscosity of biodiesel-alcohol blends as follows: (1) its decision-tree architecture effectively captures nonlinear relationships and interactions among features, while the sequential tree-building process further reduces prediction bias; (2) the regularization mechanism integrated into its objective function helps control model complexity and effectively mitigates overfitting. Consequently, owing to its strong generalization capability, stability, and robustness, the XGBoost model emerges as the preferred modeling approach for predicting the thermophysical properties of biodiesel-alcohol blended fuels. Subsequent work will focus on the interpretability analysis of key influential features for the XGBoost model.

2.2 Interpretability analysis of the XGBoost model

2.2.1 SHAP global influence analysis of input features

The SHAP method was employed to analyze the impact of input features on the predictions of the XGBoost model, as shown in Figure 7.

Figure 7 illustrates the decision logic and feature dependences of the XGBoost model in predicting density and viscosity. For density prediction, ρ_{20} and T are identified as the dominant features. Their SHAP values exhibit a wide distribution range with distinct positive and negative regions, indicating that these variables exert significant positive or negative effects on the predicted

density depending on their specific values. For viscosity prediction, T and η_{20} are identified as the key input features. The SHAP values for temperature are largely concentrated within the range of -0.2 to 0.3 , which aligns with the physical principle of fluid viscosity varying with temperature. Meanwhile, η_{20} exhibits a strong positive driving influence on the predictions.

2.2.2 SHAP Importance analysis of input features

Figure 8 presents a quantitative analysis of the average SHAP values for each feature of the XGBoost model in density prediction.

As shown in Figure 8, for density prediction, ρ_{20} emerges as the most critical input feature, with a mean SHAP value of 0.1149 . T follows closely behind and still plays a substantial role in density prediction. Regarding viscosity prediction, both T and η_{20} show pronounced influences on

the viscosity output, whereas M_w of blends and initial density contribute less. To further clarify the significant contribution of the initial density and viscosity inputs, a comparative experiment was conducted using only T and M_w of blends as input features. The corresponding model predictions on the test set are presented in Figure 9.

As illustrated in Figure 9, when T and M_w of blends were used as the sole input features, the prediction accuracy of the XGBoost model for both density and viscosity decrease significantly, with substantial deviations observed for certain blended systems. The results indicate the critical roles of initial density and viscosity in model generalization. These features not only represent known physical properties of the target systems but also encode essential information for distinguishing different alcohol isomers, which further validates the accuracy of the SHAP importance analysis and the rationality of the feature selection.

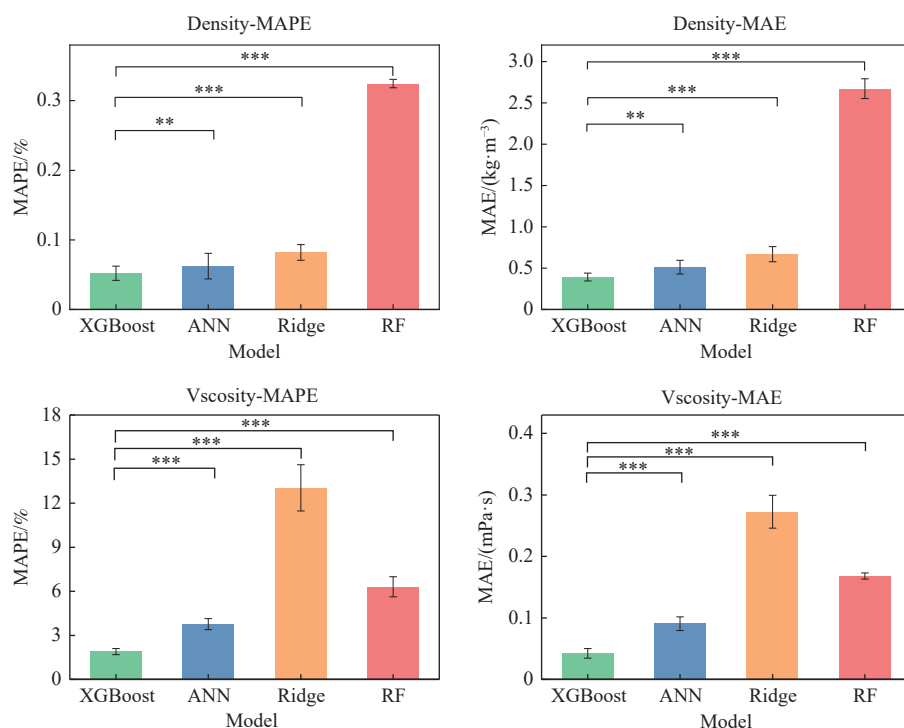


Fig. 6 Significance test for the prediction of different models

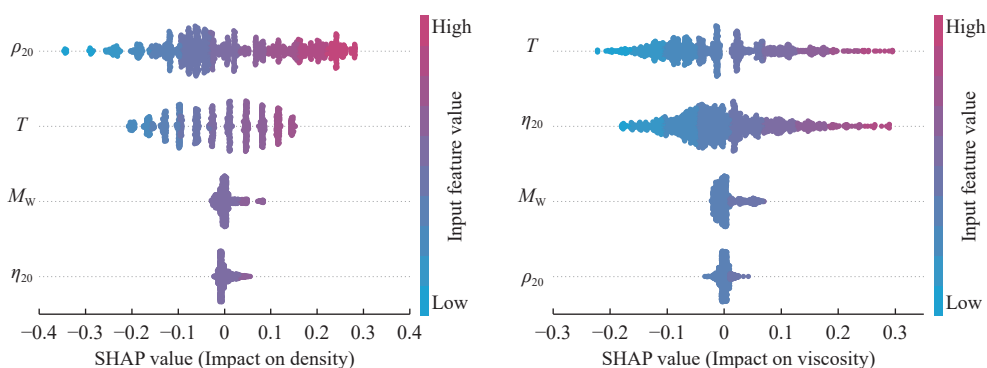


Fig. 7 SHAP analysis of the XGBoost model

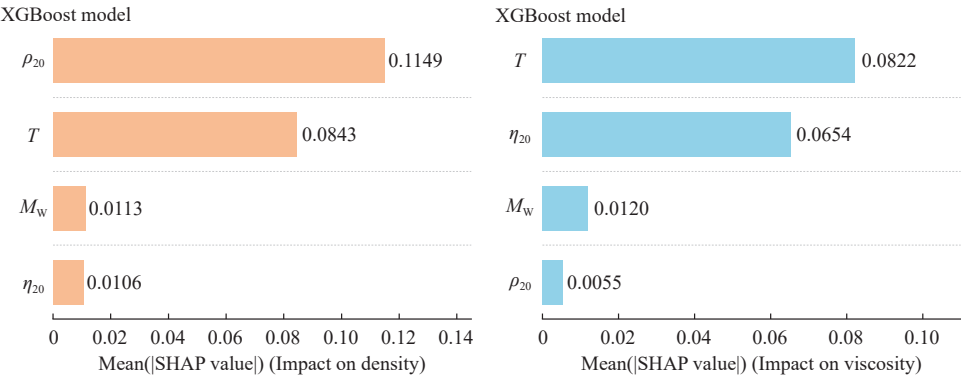


Fig. 8 Mean SHAP values for various feature outputs

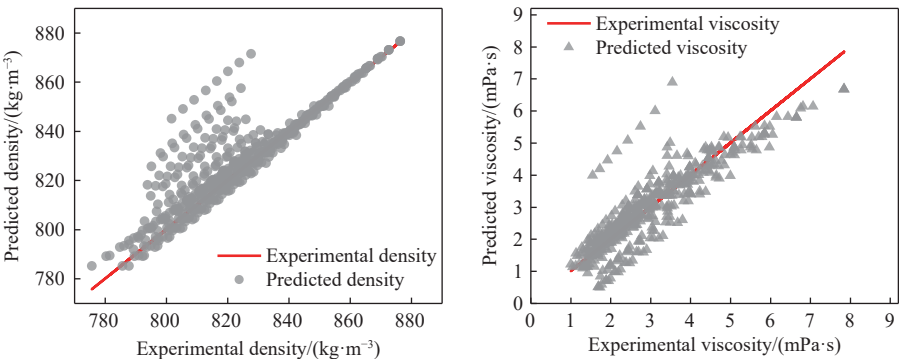


Fig. 9 Density and viscosity results of the XGBoost model with dual feature inputs

2.3 Robustness and applicability analysis of the XGBoost model

2.3.1 Analysis of resistance to input noise interference

The reliability of a model in practical applications is directly influenced by the accuracy of its input features. Considering the maximum uncertainty of 2% in the experimental measurement of density and viscosity, this study examines the impact of input noise intensity (0–5%) on prediction performance using a noise-robustness test. The input noise was introduced as zero-mean Gaussian noise, with its standard deviation set as a percentage of the corresponding input variable. The results of the noise-

robustness analysis are shown in Figure 10.

A noticeable degradation of the MAE and R^2 values was observed for both density and viscosity predictions when the noise intensity increases from 0 to 5%, indicating that the model is highly sensitive to noise interference in the input features. However, since the experimental measurement errors for biodiesel-alcohol blends are typically below 2%, the model maintained an R^2 exceeding 0.98 within this range, demonstrating the excellent robustness of the XGBoost model under low-noise conditions. Nevertheless, the strong dependence of the model on input accuracy highlights the need for rigorous data quality assessment in practical applications to ensure reliable predictions.

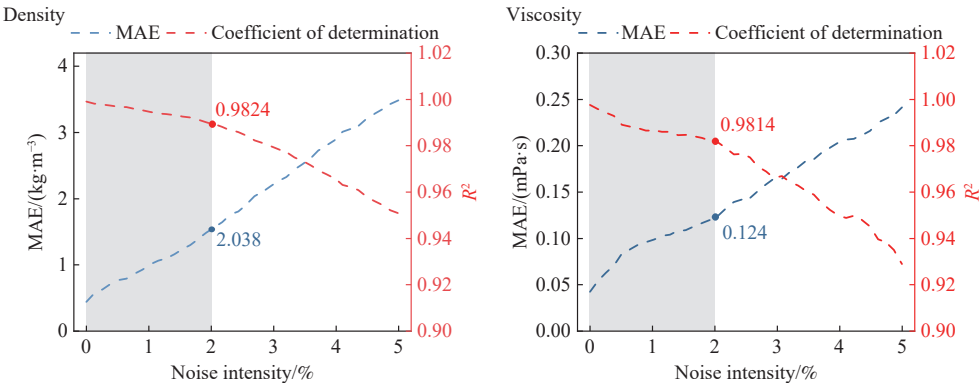


Fig. 10 Results of resistance to input noise disturbances for XGBoost model

2.3.2 Analysis of XGBoost model predictive stability

A 5-fold cross-validation was employed to systematically evaluate the stability of the XGBoost model across different data partitions. Table 4 summarizes the key performance metrics for the prediction task across each fold.

As shown in Table 4, the mean R^2 and MAE for density prediction are 0.9988 and 0.4703 kg/m³, respectively, while those for viscosity prediction are 0.9950 and 0.0505 mPa·s. The consistently high performance across various datasets, coupled with the small performance variation across cross-validation folds, demonstrates that the regularization strategies and ensemble learning mechanism effectively suppress overfitting, thereby ensuring robust model stability.

Table 4 Results of 5-fold cross-validation for the XGBoost model

Fold	Density		Viscosity	
	R^2	MAE/(kg·m ⁻³)	R^2	MAE/(mPa·s)
1	0.9986	0.5076	0.9939	0.0534
2	0.9989	0.4572	0.9932	0.0575
3	0.9986	0.5004	0.9967	0.0478
4	0.9989	0.4574	0.9966	0.0480
5	0.9991	0.4287	0.9944	0.0459
Mean	0.9988	0.4703	0.9950	0.0505
±Std	0.0002	0.0330	0.0016	0.0048

3 Conclusions

This study systematically investigated the modeling and performance evaluation of four machine learning models including Ridge regression, ANN, Random Forest, for density and viscosity prediction of biodiesel-alcohol blends. The key findings are summarized as follows:

The XGBoost model achieved a MAPE of 0.05% and 1.89% for density and viscosity prediction, respectively, with corresponding MAE of 0.38 kg/m³ and 0.045 mPa·s. Statistical significance analysis confirms that XGBoost exhibits significantly superior overall performance to the other three models, making it the most suitable choice for predicting the thermophysical properties of biodiesel-alcohol blends.

The SHAP method reveals that the decision logic of the XGBoost model aligns closely with the physical principles governing the thermophysical properties of fluids. It identifies ρ_{20} and T as the key inputs for density prediction, and T and η_{20} as the dominant inputs for viscosity prediction, while the influence of M_w remains relatively limited—a finding that enhances both the credibility and interpretability of the XGBoost model.

The XGBoost model maintains an R^2 above 0.98 under input noise levels not exceeding 2%, demonstrating noise resistance. Furthermore, the model exhibits consistently high performance across cross-validation folds, indicating its excellent robustness and stability in the prediction task. Nevertheless, given the limitations of the available

experimental data, further high-pressure experimental data are required to broaden the applicability of the XGBoost model.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

- [1] LI Z, WANG K, LIANG H, et al. Marine alternative fuels for shipping decarbonization: Technologies, applications and challenges[J]. Energy Convers Manage, 2025, 329: 119641–119679.
- [2] VOGT E T C, WECKHUYSEN B M. The refinery of the future[J]. Nature, 2024, 629(8011): 295–306.
- [3] CHO H M. Exploring the effects of synergistic combustion of alcohols and biodiesel on combustion performance and emissions of diesel engines: A review[J]. Energies, 2024, 17(24): 6274–6297.
- [4] ZHONG W J, HUANG X, GUO H, et al. Spray-evaporation characteristics of n-pentanol/n-dodecane binary fuel at ultra-high injection pressure[J]. Renewable Energy, 2023, 219: 119505–119515.
- [5] ELEBI Y, CENGİZ M, HUSEYİN A. A comprehensive utilization of pentanol and its blends as diesel engine fuel: A review[J]. Biomass Bioenergy, 2025, 193: 107602–107637.
- [6] MESTILA M. Performance of diesel engines fueled by ternary blends containing biodiesel and alcohol[J]. Fuel, 2019, 253: 1380–1390.
- [7] ALVES A A A, COSTA M F L D, MEDEIROS L H G D, et al. Investigating the thermophysical properties of the 1-Butanol+Biodiesel System: The impact of pressure on volumetric characteristics[J]. Fuel, 2024, 371: 132073.
- [8] VARGAS-IBANEZ L T, IGLESIAS-SLIVA G A, CANO-GOMEZ J J, et al. Densities and viscosities for binary liquid mixtures of biodiesel + 1-Pentanol, 2-Pentanol, or 2-Methyl-1-Butanol from (288.15 to 338.15) K at 0.1 MPa[J]. J Chem Eng Data, 2018, 63: 2438–2450.
- [9] WANG X J, MA Y J, HUO J N, et al. Viscosity measurement and correlation research of one biodiesel component (methyl octanoate) with five 1-alcohols[J]. Fuel, 2022, 324: 124441–124451.
- [10] OSMAN S, MERT G, STEFANIU A. Density, viscosity, and refractive index variations in diesel fuel+higher alcohols blends at various temperatures[J]. Int J Thermophys, 2025, 46(2): 20–26.
- [11] AITBELALE R, ACHAK M, CHHITI Y, et al. Effects of 2-Butanol addition on waste cooking oil biodiesel density: An updated experimental measurement and thermodynamic modeling study[J]. J Chem Eng Data, 2021, 66: 3123–3141.
- [12] YADAV K, A, MANSOUR S, ALKWAİ LM, et al. Reliable models for density in waste cooking oil biodiesel-alcohol mixtures using advanced learning techniques[J]. Ind Crop Prod, 2025, 234: 121602–121614.
- [13] BELMADANI S, HAMADACHE M, SI-MOUSSA C, et al. Artificial neural network models for prediction of density and kinematic viscosity of different systems of biofuels and their blends with diesel fuel[J]. Kem Ind, 2020, 69: 527–540.
- [14] BIETRESATO M, BOLLA A, CALIGIURI C, et al. The kinematic viscosity of conventional and bio-based fuel blends as a key parameter to indirectly estimate the

- performance of compression-ignition engines for agricultural purposes[J]. Fuel, 2021, 298: 120817.
- [15] YESILOVA K, YUCEL O, ERGAN B T. Modeling prediction of physical properties in sustainable biodiesel-diesel-alcohol blends via experimental methods and machine learning[J]. Processes, 2025, 13: 2310.
- [16] CANO-GOMEZ JJ, IGLESIAS-SILVA GA, RIVAS P, et al. Densities and viscosities for binary liquid mixtures of biodiesel+1-Butanol, + Isobutyl Alcohol, or+2-Butanol from 293.15 to 333.15 K at 0.1 MPa[J]. J Chem Eng Data, 2017, 6: 3391–3400.
- [17] PARK S H, YOON S H, SUH H K, et al. Effect of the temperature variation on properties of biodiesel and biodiesel-ethanol blends fuels[J]. Oil Gas Sci Technol, 2008, 63(6): 737–745.
- [18] YANG FX, WANG X, TAN H, et al. Experimental investigations on the thermophysical properties of methyl myristate in alcoholic solutions[J]. Fuel, 2018, 215: 187–195.
- [19] VARGAS-IBANEZ L T, CANO-GOMEZ J J, IGLESIAS-SILVA G A, et al. Physical properties of biodiesel blended with hexanol isomers at different temperatures: Surface tension, density, viscosity, and refractive index[J]. J Chem Eng Data, 2020, 65(7): 3706–3727.
- [20] VARGAS-IBANEZ L T, CANO-GOMEZ J J, IGLESIAS-SILVA G A, et al. Thermophysical and excess properties of diesel+biodiesel with octanol isomers at different temperatures[J]. J Mol Liq, 2022, 363: 119779–119790.
- [21] OSMAN S, STEFANIU A. Density, viscosity, and distillation temperatures of binary blends of diesel fuel mixed with oxygenated components at different temperatures[J]. Sustainability-Basel, 2023, 15(21): 15460–15480.
- [22] JIRASEK F, HASSE H. Combining machine learning with physical knowledge in thermodynamic modeling of fluid mixtures[J]. Annu Rev Chem Biomol, 2023, 14: 31–51.
- [23] MING ZY, LIU HF, WANG QL, et al. Supercritical thermophysical properties prediction of multi-component hydrocarbon fuels based on artificial neural network models[J]. Sci Tech S, 2021, 65(4): 903–919.
- [24] SINGH S, GHOSH S K. A unique artificial intelligence approach and mathematical model to accurately evaluate viscosity and density of several nanofluids from experimental data[J]. Colloids Surf Physicochem Eng Asp, 2022, 640: 128389–128398.
- [25] LU X, ZHAO C, TU H, et al. Research on prediction of energy density and power density of biomass carbon-based supercapacitors based on machine learning[J]. Sustainable Mater Techno, 2025, 44: e01309.
- [26] LUNBERG S M, LEE S I. A unified approach to interpreting model predictions[J]. Adv Neural Inf Process Syst, 2017, 2017(3): 4766–4775.

基于可解释性机器学习模型的生物柴油-醇混合物密度与黏度预测及关键特征辨识

王建安¹, 施家豪¹, 毕胜山^{1,3,*}, 安青松², 高克慧¹, 崔军卫⁴

(1. 西安交通大学 能源与动力工程学院 热流科学与工程教育部重点实验室, 陕西 西安 710049;

2. 天津大学 机械工程学院 中低温热能高效利用教育部重点实验室, 天津 300350;

3. 高效储热与低碳供热山西省重点实验室, 山西 太原 030021;

4. 长安大学 交通新能源开发、应用与汽车节能陕西省重点实验室, 陕西 西安 710064)

摘要: 准确预测生物柴油-醇混合物的热物性对其作为船用替代燃料的喷雾燃烧仿真至关重要。针对现有预测模型泛化能力与可解释性不足的问题, 本研究评估了岭回归(Ridge)、人工神经网络(Artificial Neural Network)、随机森林(Random Forest)和XGboost模型对混合物密度与黏度的预测性能与适用性。结果表明, XGBoost模型在四种模型中表现最佳, 其在测试集的全新数据上预测密度与黏度的平均绝对百分比误差(MAPE)分别为0.05%与1.89%, 利用统计检验证实了该模型高度显著的性能优势; 基于SHAP的可解释性分析揭示了XGBoost模型中初始密度(ρ_{20})与实验温度(T)为密度预测中的关键输入, 实验温度(T)和初始黏度(η_{20})为黏度预测中的关键输入。模型鲁棒性测试表明, 在输入噪声强度不高于2%时, 模型仍能保持优异的稳定性与泛化能力, 为内燃机缸内喷雾特性的数值计算提供高精度的数据支撑。

关键词: 数据驱动; 机器学习; 生物柴油-醇混合物; 密度; 黏度

中图分类号: TK428.9

文献标识码: A

作者简介:



王建安(1996), 博士研究生, 研究方向为极端工况低碳燃料热物性测量与智能预测模型开发, 真实热物性影响下燃料喷雾特性, wangjianan0423@126.com。



毕胜山(1978), 博士, 教授, 研究方向为热力学及节能、极端条件下燃油热物性, bss@mail.xjtu.edu.cn。