

Research Article

Open Access



Exploring generic fractions of multi-origin asphalts and revisiting the linkage to their bulk properties via machine learning

Jin Li^{1,2}, Jie Ma³, Jie Wu^{1,*}, Wentao He^{1,*}, Qian Xiang¹, Jian-Min Ma⁴, Mingjun Hu⁵

¹Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai 201804, China.

²Lyles School of Civil and Construction Engineering, Purdue University, West Lafayette, IN 47907, USA.

³China Academy of Transportation Science, Beijing 100029, China.

⁴Department of Chemistry, Queen's University, Kingston K7L 3N6, Ontario, Canada.

⁵School of Infrastructure Engineering, Dalian University of Technology, Dalian 116024, Liaoning, China.

*Correspondence to: Jie Wu, Wentao He, Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, 4800 Cao'an Highway, Jiading District, Shanghai 201804, China. E-mail: 2211371@tongji.edu.cn; 2111227@tongji.edu.cn

How to cite this article: Li, J.; Ma, J.; Wu, J.; He, W.; Xiang, Q.; Ma, J. M.; Hu, M. Exploring generic fractions of multi-origin asphalts and revisiting the linkage to their bulk properties via machine learning. *J. Mater. Inf.* **2025**, *5*, 37. <https://dx.doi.org/10.20517/jmi.2025.14>

Received: 17 Mar 2025 **First Decision:** 16 Apr 2025 **Revised:** 30 Apr 2025 **Accepted:** 12 May 2025 **Published:** 29 May 2025

Academic Editor: Rika Kobayashi **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Significant efforts have been made to investigate the relationship between generic fractions and bulk properties of asphalt, as the most used binder in road pavement engineering. However, due to limited data availability, advanced data mining techniques, such as machine learning (ML), have rarely been applied in the field. This study aimed to collect extensive data on asphalt generic fractions and bulk properties and to explore their underlying linkage using ML methods. A total of 800 datasets for asphalt fractions of the saturate, aromatic, resin, and asphaltene (SARA) were collected and analyzed across various asphalt types. The generic fractions and derived indices were used as input variables in ML models to predict key asphalt properties, including penetration, softening point, rutting factor, and rotational viscosity. The contribution of different generic fractions, derived indices, and additional variables (e.g., asphalt type and geographical origin) to these properties was quantified using the SHapley Additive exPlanations (SHAP) technique. Among the ML models evaluated, adaptive boosting (AdaBoost) showed the best predictive performance, while the support vector machine demonstrated greater robustness. SHAP analysis revealed that penetration was primarily influenced by the proportions of asphaltenes and saturates, while asphaltene content and the asphaltenes index were the most significant predictors for other properties, such as



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



softening point, rutting factor, and rotational viscosity. Including asphalt type and geographical origin as categorical variables in the models further improved prediction accuracy. This study highlights the potential of ML techniques in uncovering complex relationships between asphalt fractions and their bulk properties, surpassing conventional statistical approaches, though challenges remain.

Keywords: Paving asphalt, generic fractions, fraction-property linkage, machine learning, model interpretation

INTRODUCTION

Petroleum asphalt has been widely used as the paving binder in the construction of road infrastructures all over the world, of which the performance is highly dominated by chemical components and structures^[1]. Due to the complexity of asphalt molecules, the most practical way to characterize the chemical composition of asphalts is by separating them into several generic fractions based on polarity. In this term, four generic fractions include saturates, aromatics, resins, and asphaltenes, generally known as saturate, aromatic, resin, and asphaltene (SARA) fractions^[2-6].

Regarding the asphalt binder as the colloid, a widespread colloid model has been proposed to describe the distribution of SARA fractions in asphalt, as presented in Figure 1^[7-9]. In general, the fractions with relatively high molecular weight (as colloidal particles) are dispersed in the liquid phase composed of fractions with relatively low molecular weight (as dispersing agent)^[10]. In the asphalt colloid system, the colloidal particles mainly consist of asphaltenes that are surrounded by a stabilizing solvation layer consisting of resins, also known as micelles. In addition, the micelles are distributed in the dispersing agent composed of saturates and aromatics.

At present, the linkage between SARA fractions and bulk physical properties of asphalts has been widely examined and reported in previous studies^[11,12]. For example, Hofko *et al.* found that the asphaltenes proportion played a dominant role in determining the stiffness and elasticity of asphalts^[13]. Sakib *et al.* statistically examined the relationships and observed that the high-temperature modulus was highly dependent on asphaltenes while the low-temperature creep stiffness was strongly affected by aromatics and resins^[14]. Besides, Xu *et al.* found that the dynamic modulus of asphalts increased with asphaltenes or resins while decreasing with saturates or aromatics^[15]. Saturates had the greatest effect on rheological properties. Weigel and Stephan also reported the strong impact of SARA fractions on the stiffness, viscosity and temperature susceptibility of asphalts^[16]. Additionally, the aging resistance was mostly dependent on asphaltenes. Wang *et al.* focused on the effect of SARA fractions on low-temperature properties and reported that the glass transition temperature of asphalt binders was primarily determined by saturates and aromatics^[17]. Furthermore, Wang *et al.* established several correlation models between SARA fractions and viscoelastic master curves of asphalt binders using the multiple linear regression method^[18].

Extensive studies have investigated the relationship between asphalt fractions and their properties. However, the following three critical research gaps remain: (1) Current literature demonstrates insufficient consideration of asphalt origins, with limited crude oil sources considered in regional investigations, thereby limiting the generalizability of findings and failing to comprehensively characterize the variation in fractions of various asphalts due to geological differences; (2) Conventional correlation models exhibit excessive reliance on Pearson's linear correlation analysis and multivariate linear regression (MLR) methodologies, which fundamentally treat SARA fractions as four independent variables while presuming linear relationships. Those approaches ignore the chemically inherent, potentially nonlinear interactions within asphalt components. In contrast, advanced data mining techniques, such as machine learning (ML),

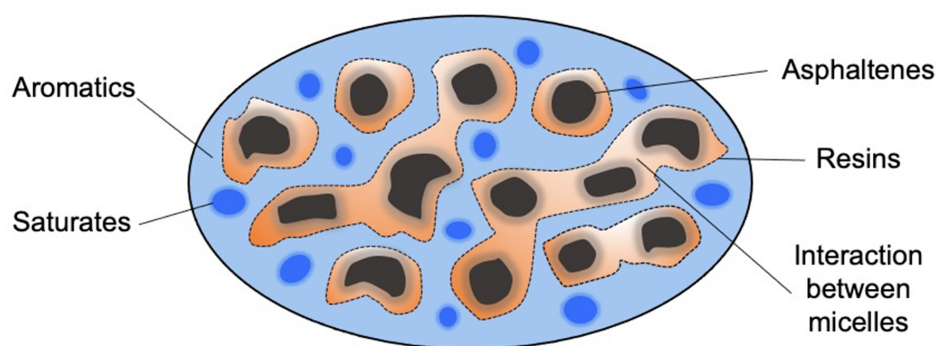


Figure 1. Colloid model of asphalt considering SARA fractions. SARA: Saturate, aromatic, resin, and asphaltene.

overcome these methodological constraints by capturing complex nonlinear dependencies and variable interactions without presuming linearity or independence, thereby achieving superior predictive accuracy than linear models and providing deeper explanation of mechanism through feature analysis and SHapley Additive exPlanations (SHAP) interpretation; (3) Although prior research has accumulated experimental data, the relatively limited dataset has restricted application of advanced data mining techniques.

To address these gaps, this study aims to collect extensive data on asphalt fractions and properties from diverse global origins through literature-based data extraction. Advanced ML modeling and interpretation techniques are then employed to explore the relationships between generic fractions and bulk properties of asphalts across multiple origins.

MATERIALS AND METHODS

Data extraction

In this study, a total of approximately 800 asphalt samples are collected from 130 papers. The data on SARA fractions, asphalt geographical origins, asphalt types and asphalt properties are extracted and processed, respectively. The raw data and corresponding information of selected literature can be found in the [Supplementary Materials](#). It is noted that the raw data summarized in the file needs to be cleaned by removing outliers for the subsequent data analysis and modeling. Specifically, outliers are identified using a combination of statistical thresholds and domain-specific rules. The interquartile range (IQR) method is applied, where data points falling outside $1.5 \times \text{IQR}$ below the first quartile or above the third quartile for key asphalt properties are flagged as potential outliers. Additionally, domain-specific rules were used to exclude data points that deviated significantly from expected physical or chemical behavior. Five types of asphalt are considered, including virgin asphalts (unaged and unmodified asphalts, VG), polymer-modified asphalts (PMA), aged asphalts (AG), recovered asphalts from RAP (RAP), and rejuvenated asphalts (RJ).

It is well known that asphalt consists of a large number of compounds, and its composition varies with the geographical origin of the crude oil. [Figure 2](#) shows the main origins of asphalts in our dataset. Based on the origin and distribution density of the materials, the origins of asphalts could be classified into 11 regions, including: (1) Western Europe; (2) Eastern Europe; (3) North Africa; (4) Sub-Saharan Africa; (5) Middle East; (6) South and Southeast Asia; (7) East and Central Asia; (8) Pacific/Oceania; (9) North America; (10) Central America and Caribbean; (11) South America.

Data analysis

ML algorithms

Given the same data quantity and number of input variables, the ML model's accuracy and performance are

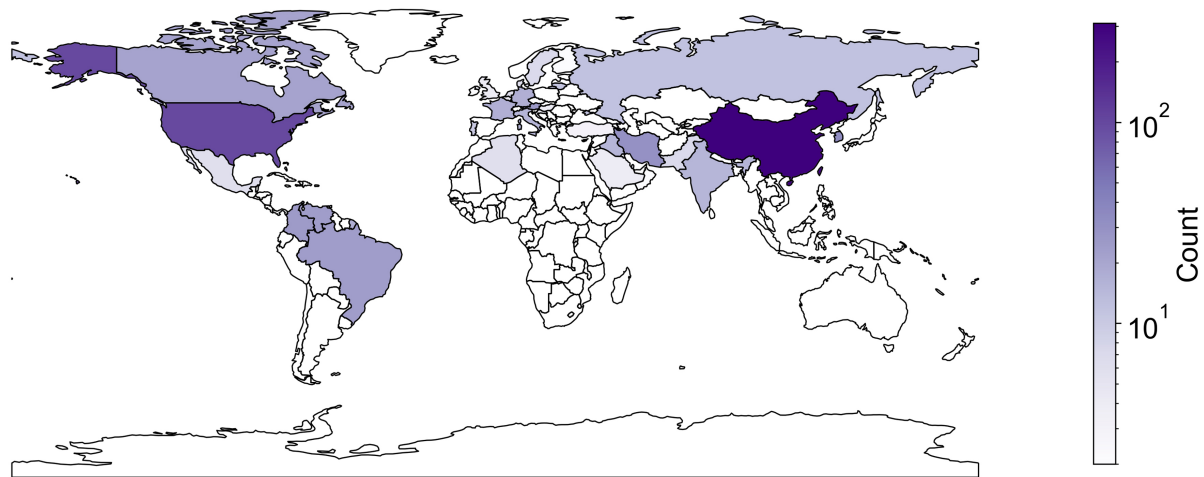


Figure 2. Distribution map of asphalt geographical origins.

highly dependent on the choice of algorithm. Consequently, even with a limited amount of data and numerous features, selecting an appropriate algorithm can enhance the model's predictive capability. This study therefore evaluates the best ML model by comparing different algorithms, with the MLR as the baseline, including extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), random forest (RF), and support vector machine (SVM).

MLR

MLR is the most basic algorithm used to model the relationship between two or more predictor variables and a continuous outcome variable. MLR is usually applied to scenarios where the relationship between features is clear and consistent with the assumptions. It has advantages with high computational efficiency and is easy to understand, but it cannot deal with problems containing complex nonlinear relationships. The model may be unstable when there are multiple covariates and is more sensitive to outliers.

XGBoost

XGBoost is an advanced implementation of the Gradient Boosting framework, designed to enhance performance and computational efficiency. It incorporates several innovations, such as a regularization technique to prevent overfitting, parallel tree boosting, and support for sparse data handling. XGBoost also uses a more efficient gradient-based optimization approach, making it faster and more scalable than traditional GB methods.

AdaBoost

AdaBoost is an ensemble learning algorithm that combines multiple weak classifiers to form a strong classifier. The algorithm works by iteratively adjusting the weights of incorrectly classified samples, focusing more on difficult cases in subsequent rounds. Each new classifier is added to minimize the overall error, and the final model is a weighted sum of all classifiers. AdaBoost is particularly effective for binary classification problems and can improve the performance of simple models, such as decision stumps while maintaining interpretability.

RF

RF is a powerful ensemble learning method widely used in ML for classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the mode of the classes

(classification) or mean prediction (regression) of the individual trees. The “random” in RF refers to two key aspects: random selection of data points for training each tree (bootstrap aggregation or bagging) and a random subset of features considered for each split in the tree. This randomness helps to decorrelate the trees, reducing overfitting and improving generalization performance. RF is robust to noisy data and works well with high-dimensional datasets. Additionally, it provides estimates of feature importance, aiding in interpretability. Overall, RF is a versatile and efficient algorithm suitable for various ML tasks.

SVM

SVM is a powerful supervised learning algorithm renowned for its effectiveness in both classification and regression tasks. Its key principle revolves around finding the optimal hyperplane that separates data points of different classes with the maximum margin, thereby maximizing the generalization ability of the model. SVM achieves this by transforming the input data into a higher-dimensional feature space, where it can construct an optimal separating hyperplane. SVM is particularly well-suited for situations with high-dimensional feature spaces and small to medium-sized datasets. Its ability to capture intricate decision boundaries and its robustness to overfitting make it a popular choice across various domains.

Hyperparameter optimization

In this study, Bayesian optimization was employed as a sophisticated approach for hyperparameter tuning in model training. This method uses probabilistic models to systematically explore the hyperparameter space. The process starts by sampling the objective function at a few randomly selected points, which serves as the initial data set for a surrogate model, commonly a Gaussian process. It acts as a probabilistic model of the objective function, providing estimates of performance across the hyperparameter space. To decide the next point to evaluate, an acquisition function is applied to balance the trade-off between exploring new regions and exploiting known areas of high performance. This point is chosen to maximize the potential gain in information or performance improvement. The selected hyperparameters are then evaluated, and the results are used to update the surrogate model. This cycle continues, with each iteration refining the model's predictions and directing computational efforts toward the most promising areas.

The 5-fold cross-validation repeated ten times was employed primarily for hyperparameter tuning to ensure robust selection of model parameters and avoid overfitting. This technique minimizes bias by randomly selecting samples for training and testing, enhancing the model's performance. In this study, 80% of the data was used for training, and the remaining 20% was used for testing. It is important to note that K-Fold cross-validation was applied only to the training dataset. In specific, the training data was divided into two groups using the K-Fold approach: one for training and the other for testing. The process was repeated over five rounds, where in each iteration, a different subset from the training data was used as the testing set while the model was trained on the remaining four subsets. To further strengthen the model's performance, this cross-validation procedure was repeated ten times; once the optimal hyperparameters were identified through this repeated cross-validation, the final model evaluation was performed on a separate hold-out test set to assess generalization performance and the model with the lowest error was selected.

Evaluating prediction results is also crucial across all ML applications. This involves comparing the predicted values with the actual measured values using statistical indices that reveal the differences between them. In this study, the coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE) were chosen as the error evaluation criteria, to determine the effectiveness and accuracy of all proposed models. R^2 , which ranges from 0 to 1, indicates the correlation between predicted and observed values, with a value closer to 1 representing a better model fit. MSE and MAE measure the dispersion between predicted and actual values, serving as indicators of error or loss. Lower values of MSE and MAE suggest a more precise and accurate model.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{i,o} - y_{i,p})^2}{\sum_{i=1}^n (y_{i,o} - \bar{y})^2} \quad (1)$$

$$MSE = \frac{\sum_{i=1}^n (y_{i,o} - y_{i,p})^2}{n} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^n |y_{i,o} - y_{i,p}|}{n} \quad (3)$$

In which $y_{i,o}$ and $y_{i,p}$ represent the observed value and predicted value of sample i , and $\bar{y}$ indicates the average value of all observed values; n signifies the sample size.

RESULTS AND DISCUSSION

Generic fractions data exploration

Asphalt can be fractionated into four generic fractions, namely SARA fractions, according to the decreasing polarity^[19]. Therefore, the proportions of SARA fractions are first explored in terms of five different types of asphalt, as shown in Figure 3.

Overall, the proportions of SARA fractions of VG are significantly different from both AG and RAP, but only the asphaltenes are significantly different compared to PMA. This can be attributed to polymer modifiers mainly present as asphaltenes in the SARA fractions. The proportions of saturates and aromatics in RAP are the lowest among the five types of asphalts, followed by AG. This is due to the fact that saturates evaporated at high temperatures, and aromatics gradually convert to resins and asphaltenes during aging^[20].

In addition, the aging status of the RAP is usually higher than it is from the indoor aging process. Therefore, more significant differences in SARA fractions can be found between VG and RAP. The data distribution range of SARA fractions of RJ is significantly smaller relative to the distribution ranges for other asphalts. This is possible because the asphalt regeneration process is usually based on the principle of component blending. As a result, the distribution of its SARA fractions can be actively controlled.

There is no significant difference between the light fractions of RJ and VG, since the rejuvenator can provide the light fraction lost due to volatilization and oxidation of the asphalt during aging and rebalance the asphalt fractions^[21]. It is worth noting that there are significant differences between the asphaltenes and resins of RJ and VG, indicating that the heavy fractions of AG cannot be fully rejuvenated to the level of VG.

Furthermore, two indices are derived from SARA fractions to characterize the colloidal structure of asphalts, including the colloidal instability index (CII) and asphaltenes index (AI)^[18]. Among them, CII refers to the ratio of (%asphaltenes + %saturates) and (%resins + %aromatics), and AI refers to the ratio of (%asphaltenes + %resins) and (%saturates + %aromatics). In general, CII is used to characterize the stability of asphaltenes in the asphalt binder. The lower the CII value, the higher the stability of asphaltenes in the asphalt binder. AI represents the asphaltenes content and is useful for assessing the colloidal stability^[22].

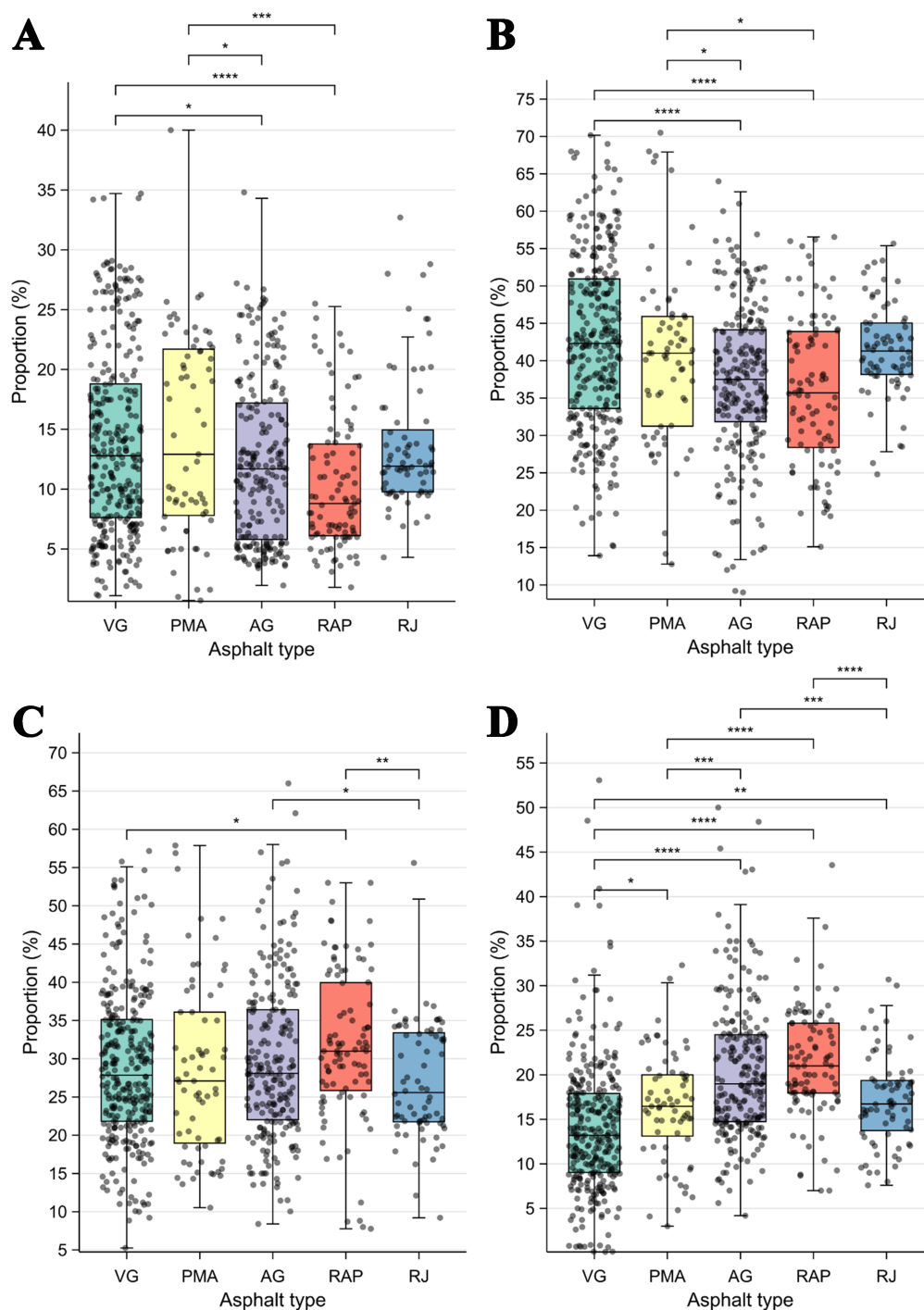


Figure 3. The box plot of SARA fractions for five asphalt binders: (A) Saturates; (B) Aromatics; (C) Resins; (D) Asphaltenes. **P* value < 0.05; ***P* value < 0.01; ****P* value < 0.001; *****P* value < 0.0001. SARA: Saturate, aromatic, resin, and asphaltene.

The distribution of these two SARA-derived indices for five types of asphalts is presented in Figure 4. It can be seen that the average CII values of the five asphalts increase in the order of VG, PMA, RJ, RAP, and AG. AI shows a similar changing pattern to that of CII. Overall, VG has the highest colloidal stability, and all other types of asphalts show varied increases in colloidal instability compared to VG.

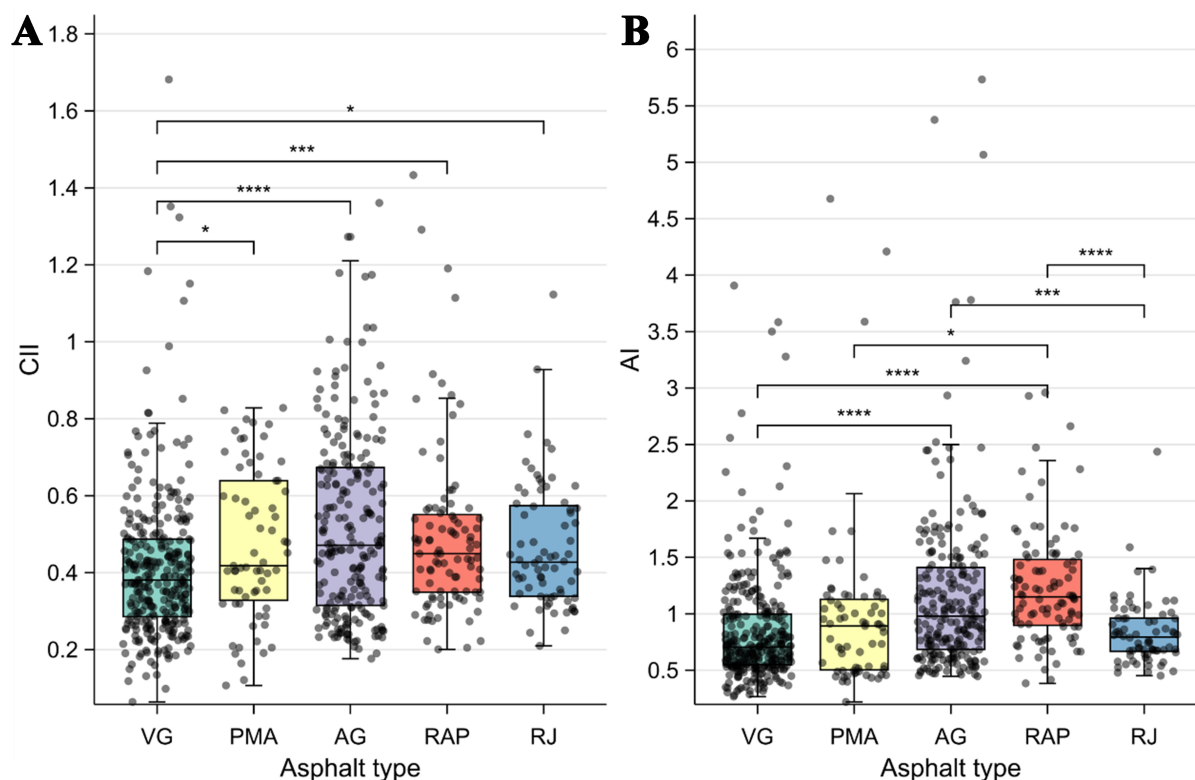


Figure 4. The box plot of SARA-derived indices for five asphalt binders: (A) CII; (B) AI. *Significance level P -value ≤ 0.05 ; ***significance level P -value ≤ 0.001 ; ****significance level P -value ≤ 0.0001 . SARA: Saturate, aromatic, resin, and asphaltene; CII: colloidal instability index; AI: asphaltenes index.

Furthermore, the significant difference in the distribution of CII and AI values between AG, RAP, and VG is much larger than it is between PMA, RJ, and VG. It is shown that the colloidal stability of asphalts is affected by aging significantly. In addition, the distribution range of CII and AI values of RAP is much greater than that of the other types of asphalts. This may be due to different aging conditions in different literature, which further leads to greater variation. Therefore, the changes in SARA-derived indices are also supposed to be reflected in the change in asphalt type.

In addition, the significance matrix of SARA fractions for five asphalt binders for different asphalt origins is presented in Figure 5. Only the three regions with the largest amount of data are shown. It can be seen that the significance levels of fractions among different asphalt types vary depending on their origins, thereby confirming the substantial influence of source on fractions. Furthermore, East and Central Asian asphalts exhibit significantly greater differences in fractions compared to Middle Eastern asphalts, while North American asphalts show relatively less variation in fractions across different asphalt types. Notably, asphaltene emerges as the fraction most significantly affected by asphalt type across all asphalt origins, demonstrating the highest sensitivity to material classification differences.

Fraction-property linkage modeling

Based on the exploration of generic fractions data above, further attempts are made to understand the linkage between asphalt fractions and properties. The SARA fractions and the SARA-derived indices as well as asphalt geographical origin and asphalt type are employed as input variables in the ML modeling, which included both continuous variables and categorical variables.

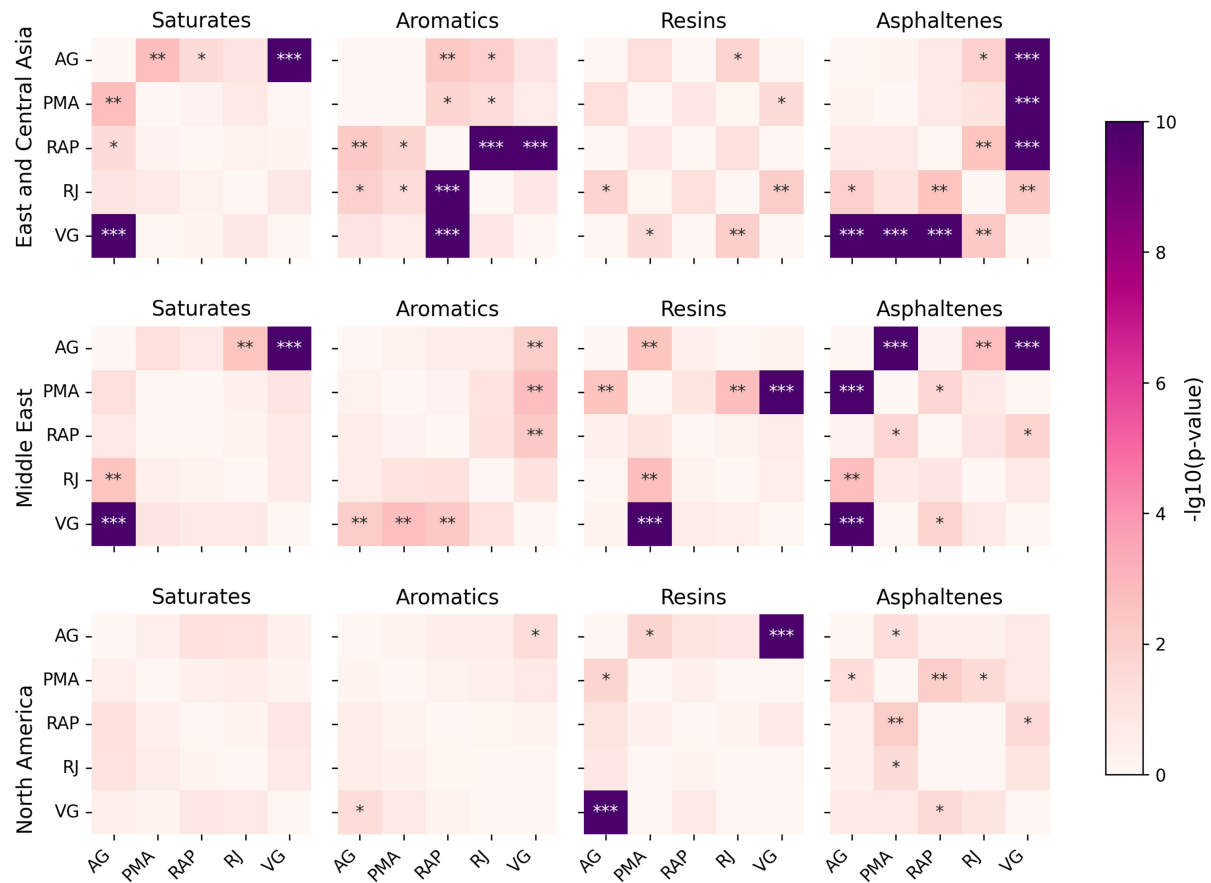


Figure 5. The significance matrix of SARA fractions for five asphalt binders across different asphalt origins. *Significance level P -value ≤ 0.05 ; **significance level P -value ≤ 0.01 ; ***significance level P -value ≤ 0.001 . SARA: Saturate, aromatic, resin, and asphaltene.

Figure 6 shows the pair plot between SARA fractions and SARA-derived indices. Firstly, there is no correlation between the four SARA fractions. In addition, CII is found to show a similar pattern to asphaltenes, while AI is related to aromatics to some extent ($R^2 = 0.72$). For different types of asphalt, the correlation between each pair does not present significant differences.

To analyze the relationship between SARA fractions and bulk properties, several common asphalt properties are selected, including penetration, softening point, rutting factor (64 °C), and rotational viscosity (135 °C).

The distribution of the collected data for four properties is plotted in **Figure 7**. The distribution of penetration approximates a normal distribution, centered between 60 and 80 (0.1 mm). The distribution of softening point exhibits a slightly skewed distribution, peaking around 50 °C. It has a longer right tail extending beyond 70 °C, suggesting higher softening point values. The distribution of rutting factor is notably right skewed distribution, with most values between 0 and 4 kPa, and fewer instances exceeding 6 kPa. The distribution of rotational viscosity exhibits similar characteristics. The distributions indicate that the rutting factor and the rotational viscosity are characterized by sparsity. In other words, most samples are in the low value range and fewer are in the high value range.

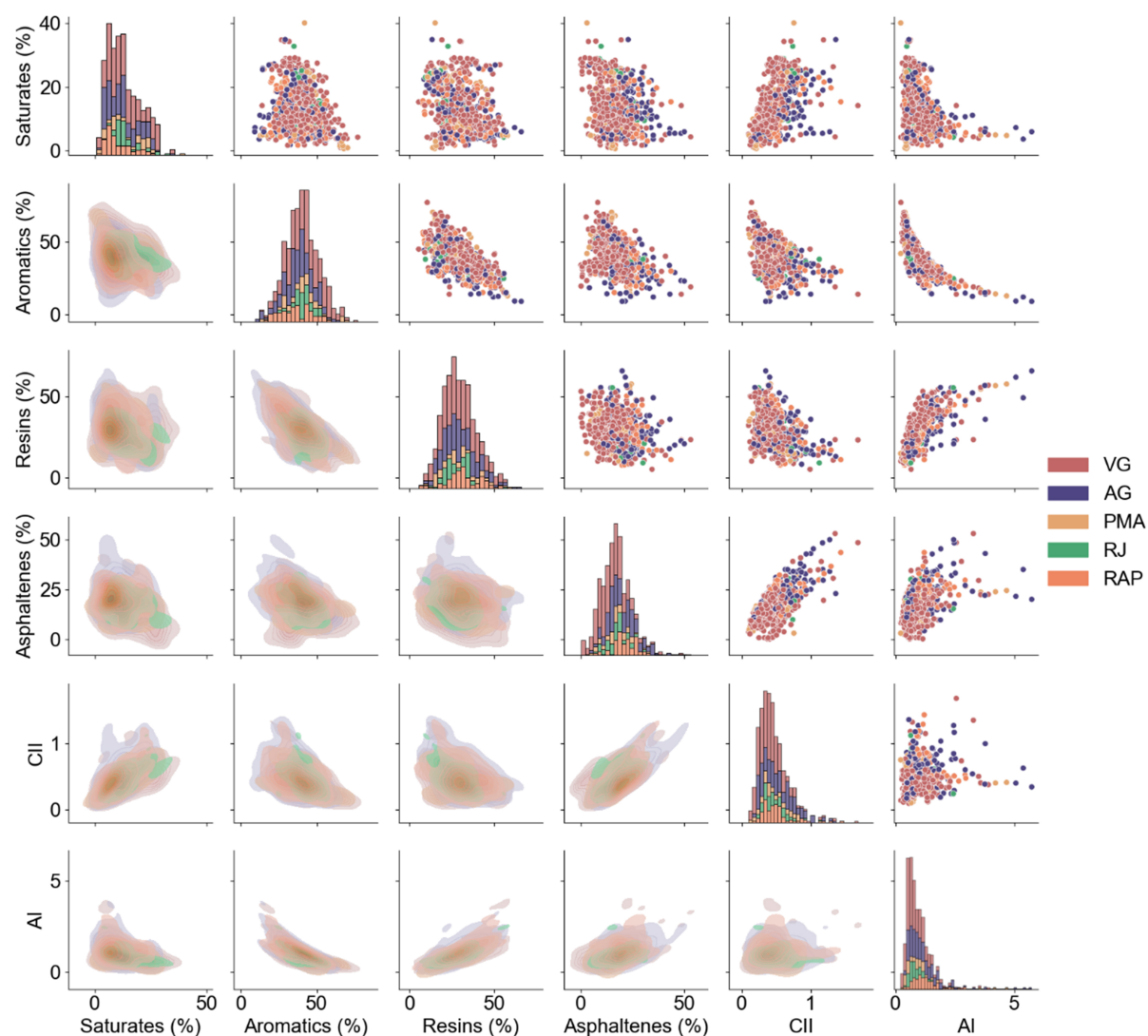


Figure 6. The pair plot between SARA fractions and SARA-derived indices. SARA: Saturate, aromatic, resin, and asphaltene.

Table 1 summarizes the three performance evaluation metrics of five models with SARA fractions, SARA-derived indices, the origin and type of asphalt as input variables and four asphalt properties as output variables, respectively. In general, the XGBoost and the RF perform well on the training set with a R^2 greater than 0.9. However, it shows a significant drop in prediction performance on the testing set, indicating probable overfitting. For the other three models, the average R^2 values of the training set for all the performance prediction models are around 0.6, while the R^2 value of the testing sets decreased to varying degrees.

The average difference between the R^2 value of the training set and testing set is the smallest for penetration, which is only 0.05. However, it is 0.09 for the softening point and it is 0.13 for the rotational viscosity, presenting a gradual decrease in the generalization ability of the model. The previous analysis of the input parameter characteristics shows that the rutting factor and rotational viscosity distributions are significantly right skewed, with only a small number of high value samples. This data imbalance is the main reason for the reduced generalization ability of the model. In addition, data imbalance has been identified as the

Table 1. Performance evaluation and comparison of different ML models for asphalt properties prediction

Asphalt property	ML model	Training set			Testing set		
		MSE	MAE	R ²	MSE	MAE	R ²
Penetration	MLR	253.59	12.68	0.56	392.17	15.64	0.51
	XGBoost	0.86	0.54	0.99	226.26	11.37	0.72
	AdaBoost	177.41	11.44	0.69	304.48	14.36	0.62
	RF	37.30	4.84	0.94	238.64	12.15	0.70
	SVM	207.07	10.59	0.64	312.82	13.68	0.61
Softening point	MLR	26.85	3.53	0.56	55.2	5.09	0.51
	XGBoost	4.65	1.44	0.92	37.14	4.19	0.67
	AdaBoost	17.41	3.43	0.72	46.53	4.96	0.58
	RF	4.51	1.50	0.93	40.11	4.30	0.64
	SVM	27.95	3.34	0.55	58.98	5.18	0.47
Rutting factor	MLR	1.41	0.81	0.57	1.94	0.98	0.39
	XGBoost	0.14	0.26	0.96	2.32	0.85	0.27
	AdaBoost	0.59	0.58	0.82	2.74	1.02	0.14
	RF	0.21	0.30	0.93	2.24	0.87	0.30
	SVM	0.90	0.57	0.73	2.10	0.85	0.34
Rotational viscosity	MLR	0.07	0.18	0.46	0.06	0.19	0.52
	XGBoost	0.03	0.10	0.79	0.06	0.18	0.53
	AdaBoost	0.04	0.16	0.69	0.07	0.19	0.48
	RF	0.01	0.07	0.90	0.05	0.17	0.61
	SVM	0.05	0.12	0.61	0.06	0.18	0.51

ML: Machine learning; MSE: mean squared error; MAE: mean absolute error; R²: the coefficient of determination; MLR: multivariate linear regression; XGBoost: extreme gradient boosting; AdaBoost: adaptive boosting; RF: random forest; SVM: support vector machine.

primary cause of overfitting in XGBoost and RF algorithms, which is corroborated by previous research. It is recommended that subsequent studies employ mitigation strategies such as over-sampling, under-sampling, cost-sensitive and boosting algorithms to resolve data imbalance^[23,24].

Overall, the AdaBoost model has the best training results and outperforms the other prediction models for the testing set. The MLR model and SVM model, on the other hand, perform more conservatively on the training set, but the results of the testing set are not much different from the training set, indicating their greater generalization ability. Finally, it should be noted that the purpose of our study was not to provide an accurate prediction model of asphalt properties but to explore and gain more insights into the implicit linkage between generic fractions and bulk properties of asphalt with ML.

Feature importance analysis

The properties of asphalt are not only dependent on or controlled by the proportion of each fraction but also by other factors, such as the structural characteristics of fractions. This can be reflected in the types and origins of asphalts to some extent. Therefore, the impact of such features, including asphalt types and origins, on the prediction accuracy of the model is analyzed and the importance of different factors in developing the predictive model based on SARA fractions is explored.

For the penetration prediction model, three additional models are developed: One excluding the asphalt origin from the input variables, another excluding the asphalt type, and the third excluding both asphalt origin and type. The model performance evaluation results are presented in [Table 2](#).

Table 2. Performance evaluation of the prediction model for penetration

ML model	Input variables	Training			Testing		
		MSE	MAE	R ²	MSE	MAE	R ²
MLR	SARA & Origin & Type	253.59	12.68	0.56	392.17	15.64	0.51
	SARA & Origin	351.23	15.22	0.39	557.33	19.76	0.30
	SARA & Type	290.19	13.71	0.50	413.43	16.16	0.48
	SARA	434.58	16.91	0.24	632.81	21.08	0.21
XGBoost	SARA & Origin & Type	0.86	0.54	0.99	226.26	11.37	0.72
	SARA & Origin	1.56	0.90	0.99	326.99	14.23	0.59
	SARA & Type	3.09	1.31	0.99	260.18	12.69	0.67
	SARA	71.90	6.98	0.87	443.42	17.53	0.44
AdaBoost	SARA & Origin & Type	177.41	11.44	0.69	304.48	14.36	0.62
	SARA & Origin	264.45	14.00	0.54	464.08	19.04	0.42
	SARA & Type	201.81	11.83	0.65	345.92	15.00	0.57
	SARA	359.66	15.48	0.37	497.01	19.28	0.38
RF	SARA & Origin & Type	37.30	4.84	0.94	238.64	12.15	0.70
	SARA & Origin	42.12	4.99	0.93	327.91	14.69	0.59
	SARA & Type	31.71	4.40	0.94	279.21	13.30	0.65
	SARA	90.22	7.20	0.84	391.82	16.32	0.51
SVM	SARA & Origin & Type	207.07	10.59	0.64	312.82	13.68	0.61
	SARA & Origin	249.35	10.96	0.57	449.39	16.06	0.44
	SARA & Type	236.48	11.39	0.59	326.35	13.58	0.59
	SARA	400.71	16.75	0.23	591.61	20.52	0.26

ML: Machine learning; MSE: mean squared error; MAE: mean absolute error; R²: the coefficient of determination; MLR: multivariate linear regression; SARA: saturate, aromatic, resin, and asphaltene; XGBoost: extreme gradient boosting; AdaBoost: adaptive boosting; RF: random forest; SVM: support vector machine.

Figure 8 shows the R² for the different models and the change of the R² for the models with distinct input variables compared to the model of *Both*. Note that the model named *Both* indicates that the feature of both origin and type is considered as additional input variables. The model named *Origin* indicates that only the feature of asphalt origin is considered as additional input variables. The model named *Type* indicates that only the feature of asphalt type is considered as additional input variables. The model named *SARA* indicates that the feature of both origin and type is not considered as additional input variables. Overall, all the models with some input variables removed shows lower R², indicating that the accuracy of the predictions decrease. Compared to the model with SARA-related variables only, the R² of the training set increases by 0.1 to 0.36, and the R² of the testing set increases by 0.14 to 0.33 when the asphalt type is added to the input variables. The R² of the training set increases by 0.09 to 0.34, and the R² of the testing set increases by 0.04 to 0.18 when the asphalt origin is added. Besides, the asphalt type contributes more significantly to improving prediction accuracy than the asphalt origin. On the other hand, the AdaBoost model has the best accuracy of prediction when the input variables contain both asphalt origin and type or only SARA fractions. The SVM model performed better when the input variables lack either the asphalt type or origin.

Due to the black-box property of ML models in their engineering application, SHAP technique is further used to quantify the contribution of each SARA fraction and SARA-derived indices to the asphalt properties. The SHAP technique is a widely adopted method of explaining the predictions of the model and understanding the effects of various variables on the model. In a ML model, the contribution of each input variable to the model is expressed as a SHAP value. The SHAP values are calculated by evaluating model output differences due to the removal of specific input variables^[25,26].

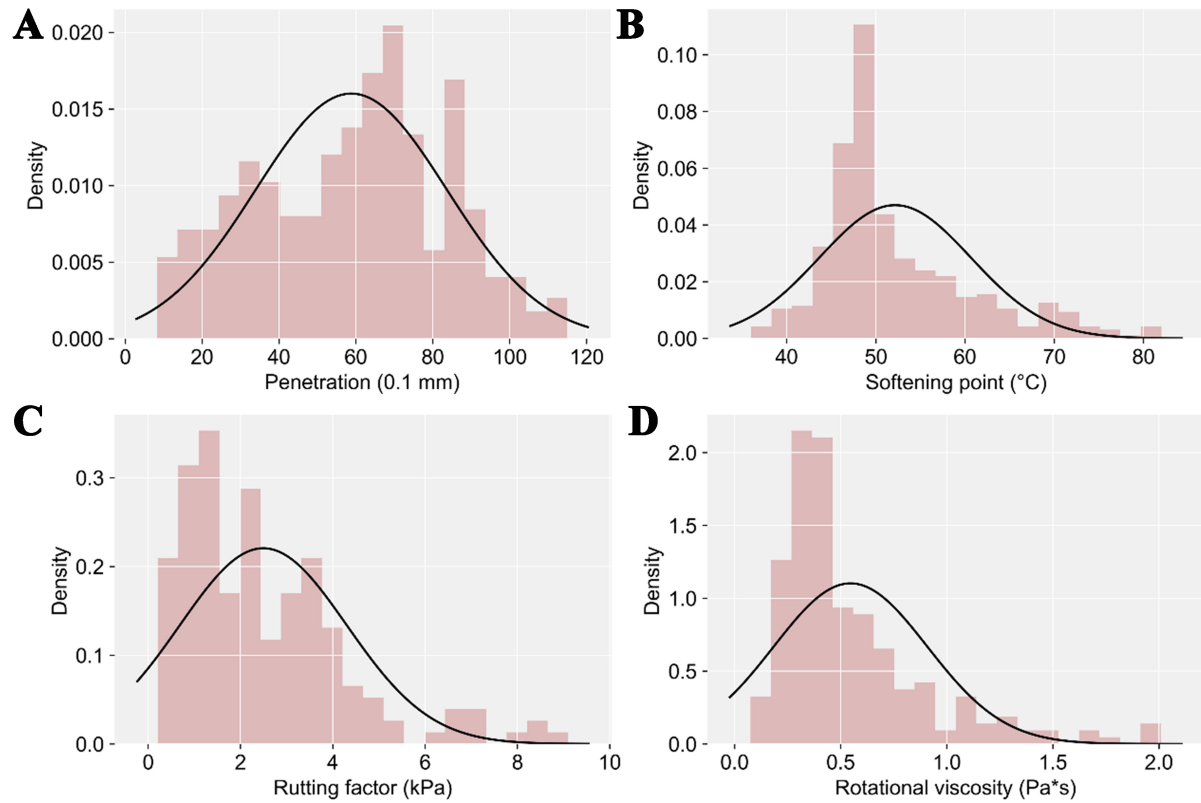


Figure 7. Probability density of the collected data for asphalt properties: (A) Penetration; (B) Softening point; (C) Rutting factor; (D) Rotational viscosity.

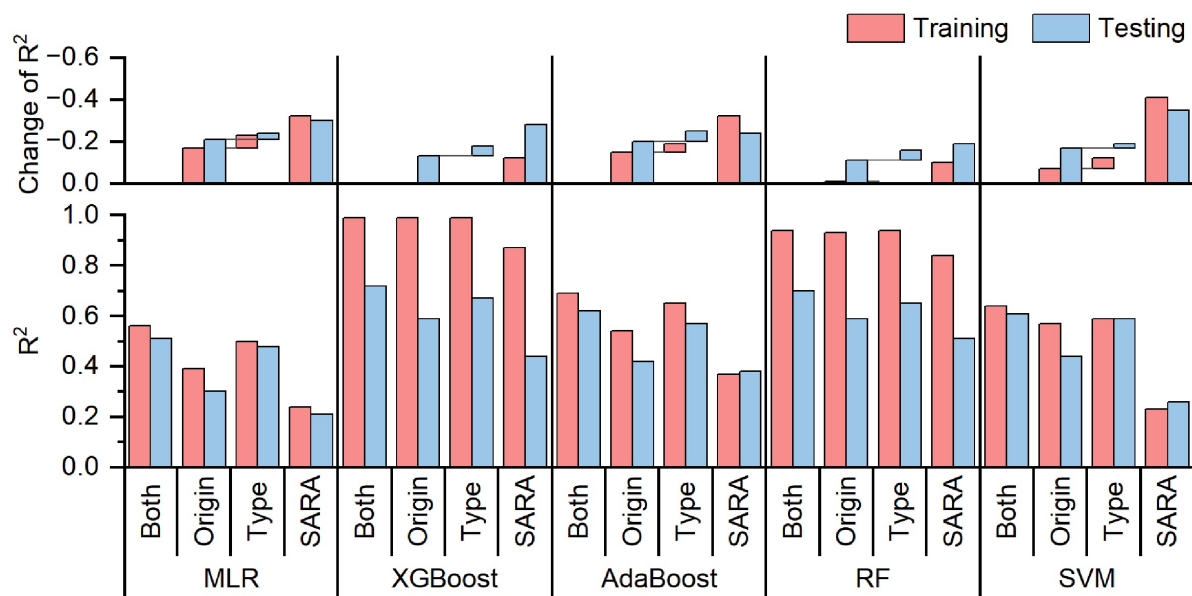


Figure 8. Performance evaluation and comparison of the prediction models with different input variables.

Figure 9 shows the mean SHAP value for different input variables in the prediction model of penetration. In both RF model and XGBoost model, the degree of influence of each variable on asphalt penetration is ranked from largest to smallest, as AI, asphaltenes, saturates, resins, CII, and aromatics, with asphaltenes being the most significant influencing factor. AI contributes more to asphalt penetration compared to CII. It suggests that the heavy fractions in asphalts are more important factors in determining the penetration of asphalts. But saturates show a higher contribution than resins. In other words, the penetration of asphalt is mainly determined by the proportion of asphaltenes and saturates.

Figure 10 shows the contribution of different input variables to softening point. Asphaltenes are also the largest factor for softening point in both RF model and XGBoost model. It is worth noting that the contribution of AI and CII is higher than the other three fractions. Therefore, the proportion of asphaltenes in asphalt is the most important factor in determining the softening point of asphalt.

There is a difference in the RF and XGBoost model, in terms of these variables that make a major contribution to the rutting factor according to Figure 11. However, the three variables that contributed the most in both models are the same: resins, asphaltenes, and CII. It suggests that the proportion of asphaltenes and resins has the greatest influence on the rutting factor of asphalts. In fact, resins play an important role in promoting the dispersion of asphaltenes and preventing their aggregation and segregation in asphalt. Therefore, the rutting factor of asphalt is primarily determined by the stability of asphaltenes.

As shown in Figure 12, the separate contribution of the variables to the rotational viscosity is not the same in the different models. Asphaltenes are the largest contributor to rotational viscosity in both models. Thus, the absolute percentage of asphaltenes may be an important determinant of rotational viscosity. It is worth noting that the four most contributing factors in both models are also aromatics, resins, asphaltenes, and CII, which are the same determinants of the rutting factor. Therefore, the rotational viscosity of asphalt is also determined by the stability of asphaltenes. In the XGBoost model, the mean SHAP value of the four variables is essentially the same. This may indicate that the proportion and stabilization degree of the asphaltene in the asphalt are equally important for rotational viscosity.

Overall, the importance of asphaltenes fraction and its related variables is relatively higher than others. In other words, the asphaltenes are the most important factors to all asphalt properties considered in this study. Existing studies have demonstrated that asphaltenes exist in asphalt as fine particles (1-10 nm) and serve as the core component of the asphalt colloidal system, exhibiting strong self-associating tendencies^[27,28]. With increasing asphaltene concentration, asphaltene molecules initially form nano-sized aggregates, which further develop into larger clusters upon continued concentration enhancement. The stability of these asphaltene clusters is strongly influenced by temperature and concentration variations. Consequently, the asphaltene fraction in asphalt significantly governs its rheological properties, particularly in terms of viscoelastic behavior and temperature-dependent performances^[29,30]. Higher asphaltene fraction promotes stronger intermolecular interactions, leading to increased viscosity and enhanced resistance to deformation, which aligns with the SHAP values indicating their dominant influence. Conversely, saturates, being non-polar and lighter, contribute to the micellar network's stability.

CONCLUSION

In this study, approximately 800 datasets of SARA fractions from petroleum asphalts of different global origins were collected to investigate the relationship between asphalt fractions and properties using ML algorithms. The main findings and conclusions are summarized below:

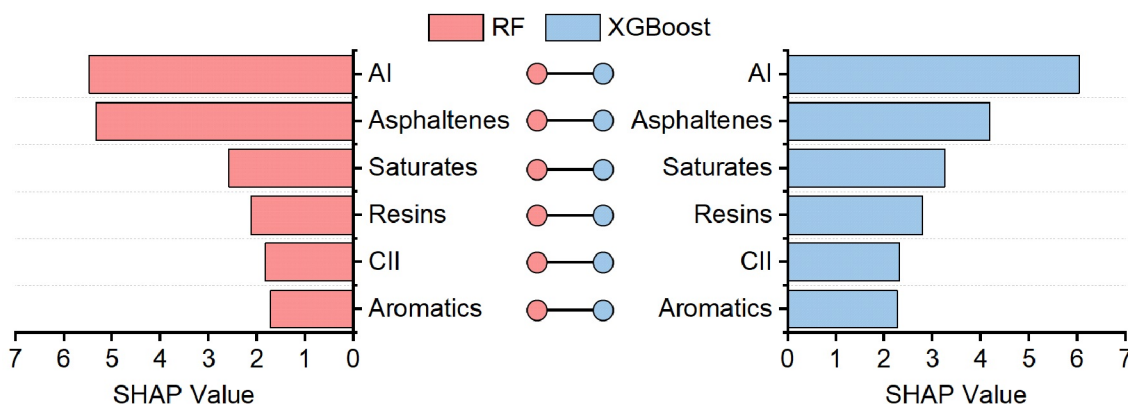


Figure 9. The mean SHAP values of different input variables in the prediction model of penetration. SHAP: The SHapley Additive exPlanations.

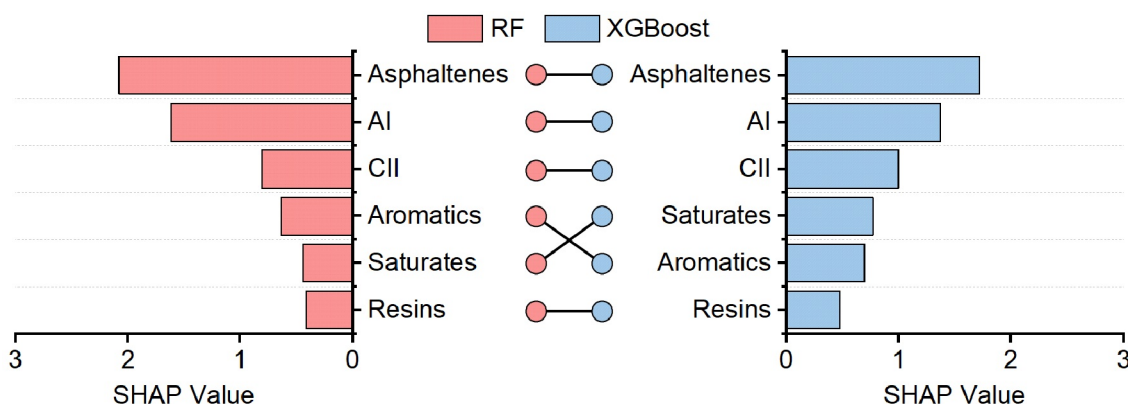


Figure 10. The mean SHAP values of different input variables in the prediction model of softening point. SHAP: The SHapley Additive exPlanations.

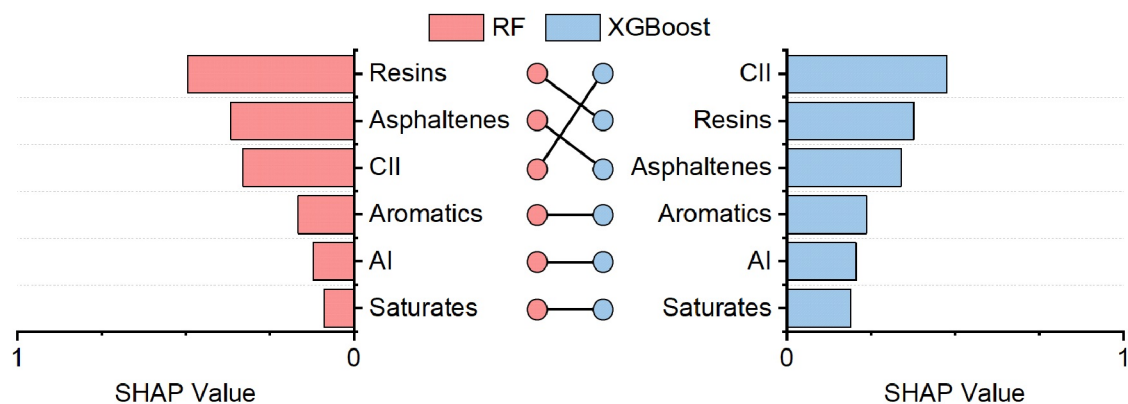


Figure 11. The mean SHAP values of different input variables in the prediction model of rutting factor. SHAP: The SHapley Additive exPlanations.

• The proportions of SARA fractions in VG differ significantly from those in AG and RAP, but only asphaltene content shows a significant difference compared to PMA. There are notable differences in asphaltenes and resins contents between RJ and VG, whereas light fractions are similar. The differences in

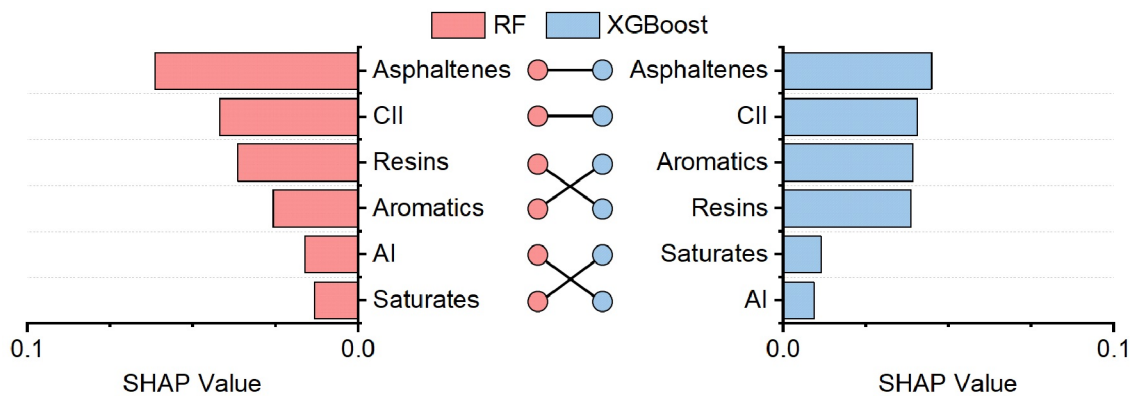


Figure 12. The mean SHAP values of different input variables in the prediction model of rotational viscosity. SHAP: The SHapley Additive exPlanations.

CII and AI distributions between AG, RAP, and VG are much larger than those between PMA, RJ, and VG. RJ has the narrowest distribution range for SARA fractions, while RAP has the widest.

- The XGBoost and RF models perform well on the training set but experienced significant performance drops on the testing set due to overfitting. In contrast, the AdaBoost model achieves optimal training results and outperforms other models on the testing set. The SVM model shows greater robustness and better generalization across both training and testing sets.

- Models that include asphalt origin and type as input variables achieve higher prediction accuracy compared to those using only SARA fractions. Asphalt type has a more substantial impact on improving accuracy than asphalt origin. The AdaBoost model achieves the highest accuracy when both asphalt origin and type are included as input variables, while the SVM model performs better when either variable is excluded.

- Asphalt penetration is primarily influenced by the proportions of asphaltenes and saturates. The proportion of asphaltenes is the most critical factor affecting the softening point. The rutting factor is mainly determined by the stability of asphaltenes. For rotational viscosity, both the proportion of asphaltenes and the colloidal stability of asphalt are equally important.

DECLARATIONS

Acknowledgments

We would like to thank all relevant researchers for reporting and presenting the asphalt fractions and properties data in their papers, which were crucial for the progress of this research.

Authors' contributions

Performed the research, analyzed data, wrote the programmers, and drafted the manuscript: Li, J.; Wu, J.; He, W.

Designed the study, performed the research, analyzed data, revised, and finalized the manuscript: Ma, J. M.; Hu, M.

Discussed the results: Li, J.; Wu, J.; He, W.; Ma, J.; Xiang, Q.; Ma, J. M.; Hu, M.

Availability of data and materials

All experimental data collected in the study are contained in [Supplementary Materials](#).

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Hu, M.; Ji, S.; Li, M.; Liu, L.; Cheng, H. Revealing the aging-induced chemical and microstructure evolution of asphalt via AFM-IR and quantum chemistry simulation. *Fuel* **2025**, *395*, 135218. [DOI](#)
2. Abdul Jameel, A. G.; Van Oudenhoven, V.; Emwas, A.; Sarathy, S. M. Predicting octane number using nuclear magnetic resonance spectroscopy and artificial neural networks. *Energy. Fuels*. **2018**, *32*, 6309-29. [DOI](#)
3. Alvarez, E.; Marroquín, G.; Trejo, F.; Centeno, G.; Ancheyta, J.; Díaz, J. A. Pyrolysis kinetics of atmospheric residue and its SARA fractions. *Fuel* **2011**, *90*, 3602-7. [DOI](#)
4. Chakravarthy, R.; Naik, G. N.; Savalia, A.; et al. Determination of naphthenic acid number in petroleum crude oils and their fractions by mid-fourier transform infrared spectroscopy. *Energy. Fuels*. **2016**, *30*, 8579-86. [DOI](#)
5. Corbett, L. W. Composition of asphalt based on generic fractionation, using solvent deasphalting, elution-adsorption chromatography, and densimetric characterization. *Anal. Chem.* **1969**, *41*, 576-9. [DOI](#)
6. Elbaz, A. M.; Gani, A.; Hourani, N.; Emwas, A.; Sarathy, S. M.; Roberts, W. L. TG/DTG, FT-ICR mass spectrometry, and NMR spectroscopy study of heavy fuel oil. *Energy. Fuels*. **2015**, *29*, 7825-35. [DOI](#)
7. Lesueur, D. The colloidal structure of bitumen: consequences on the rheology and on the mechanisms of bitumen modification. *Adv. Colloid. Interface. Sci.* **2009**, *145*, 42-82. [DOI](#) [PubMed](#)
8. Loeber, L.; Muller, G.; Morel, J.; Sutton, O. Bitumen in colloid science: a chemical, structural and rheological approach. *Fuel* **1998**, *77*, 1443-50. [DOI](#)
9. Pfeiffer, J. P.; Saal, R. N. J. Asphaltic bitumen as colloid system. *J. Phys. Chem.* **1940**, *44*, 139-49. [DOI](#)
10. Hu, M.; Lyu, L.; Pahlavan, F.; Han, P.; Sun, D.; Fini, E. H. Toward sustainable non-emitting asphalts: understanding diffusion-adsorption mechanisms of hazardous organic compounds. *Adv. Sustain. Syst.* **2025**, *9*, 2400868. [DOI](#)
11. Roja, K. L.; Masad, E. Influence of chemical constituents of asphalt binders on their rheological properties. *Transp. Res. Rec. J. Transp. Res. Board.* **2019**, *2673*, 458-66. [DOI](#)
12. Wang, J.; Zhang, R.; Wang, R.; et al. Prediction of the fundamental viscoelasticity of asphalt mixtures using ML algorithms. *Constr. Build. Mater.* **2024**, *442*, 137573. [DOI](#)
13. Hofko, B.; Eberhardsteiner, L.; Füssl, J.; et al. Impact of maltene and asphaltene fraction on mechanical behavior and microstructure of bitumen. *Mater. Struct.* **2016**, *49*, 829-41. [DOI](#)
14. Sakib, N.; Hajj, R.; Hure, R.; Alomari, A.; Bhasin, A. Examining the relationship between bitumen polar fractions, rheological performance benchmarks, and tensile strength. *J. Mater. Civ. Eng.* **2020**, *32*, 04020143. [DOI](#)
15. Xu, Y.; Zhang, E.; Shan, L. Effect of SARA on rheological properties of asphalt binders. *J. Mater. Civ. Eng.* **2019**, *31*, 04019086. [DOI](#)
16. Weigel, S.; Stephan, D. Relationships between the chemistry and the physical properties of bitumen. *Road. Mater. Pavement. Des.* **2018**, *19*, 1636-50. [DOI](#)
17. Wang, T.; Wang, J.; Hou, X.; Xiao, F. Effects of SARA fractions on low temperature properties of asphalt binders. *Road. Mater. Pavement. Des.* **2021**, *22*, 539-56. [DOI](#)
18. Wang, J.; Wang, T.; Hou, X.; Xiao, F. Modelling of rheological and chemical properties of asphalt binder considering SARA fraction. *Fuel* **2019**, *238*, 320-30. [DOI](#)
19. Qu, X.; Fan, Z.; Li, T.; et al. Understanding of asphalt chemistry based on the six-fraction method. *Constr. Build. Mater.* **2021**, *311*,

125241. DOI
20. Li, J.; Xing, X.; Hou, X.; Wang, T.; Wang, J.; Xiao, F. Determination of SARA fractions in asphalts by mid-infrared spectroscopy and multivariate calibration. *Measurement* **2022**, *198*, 111361. DOI
 21. Gong, M.; Yang, J.; Zhang, J.; Zhu, H.; Tong, T. Physical–chemical properties of aged asphalt rejuvenated by bio-oil derived from biodiesel residue. *Constr. Build. Mater.* **2016**, *105*, 35–45. DOI
 22. Xiao, X.; Wang, J.; Wang, T.; Amirkhanian, S. N.; Xiao, F. Linear visco-elasticity of asphalt in view of proportion and polarity of SARA fractions. *Fuel* **2024**, *363*, 130955. DOI
 23. Kim, M.; Kang, D.; Kim, H. B. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert. Syst. Appl.* **2015**, *42*, 1074–82. DOI
 24. Li, Z.; Kamnitsas, K.; Glocker, B. Overfitting of neural nets under class imbalance: analysis and improvements for segmentation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P, Khan A, editors. Medical Image Computing and Computer Assisted Intervention - MICCAI 2019. Cham: Springer International Publishing; 2019. pp. 402–10. DOI
 25. Koushik, A.; Manoj, M.; Nezamuddin, N. SHapley Additive exPlanations for explaining artificial neural network based mode choice models. *Transp. Dev. Econ.* **2024**, *10*, 200. DOI
 26. Yan, T.; Xing, X.; Xia, T.; Wang, D. Relation between fault characteristic frequencies and local interpretability shapley additive explanations for continuous machine health monitoring. *Eng. Appl. Artif. Intell.* **2024**, *136*, 109046. DOI
 27. Shan, L.; Xie, R.; Wagner, N. J.; He, H.; Liu, Y. Microstructure of neat and SBS modified asphalt binder by small-angle neutron scattering. *Fuel* **2019**, *253*, 1589–96. DOI
 28. Yen, T. F. The colloidal aspect of a macrostructure of petroleum asphalt. *Fuel. Sci. Technol. Int.* **1992**, *10*, 723–33. DOI
 29. Barré, L.; Jestin, J.; Morisset, A.; Palermo, T.; Simon, S. Relation between nanoscale structure of asphaltene aggregates and their macroscopic solution properties. *Oil. Gas. Sci. Technol. Rev. IFP.* **2009**, *64*, 617–28. DOI
 30. Tan, Y.; Li, G.; Dan, L.; Lyu, H.; Meng, A. Research progress of bitumen microstructures and components. *J. Traffic. Transp. Eng.* **2020**, *20*, 1–17. DOI