



# Modeling of asphaltic sludge formation during acidizing process of oil well reservoir using machine learning methods

Sina Shakouri<sup>a,b</sup>, Maysam Mohammadzadeh-Shirazi<sup>a,b,\*</sup>

<sup>a</sup> Department of Petroleum Engineering, School of Chemical and Petroleum Engineering, Shiraz University, Shiraz, Iran

<sup>b</sup> Formation Damage and Well Treatment Research Group, IOR/EOR Research Institute, Shiraz University, Shiraz, Iran

## ARTICLE INFO

Handling Editor: Wojciech Stanek

### Keywords:

Acid stimulation

Asphaltic sludge

Machine learning

Multi-layer perceptron

Extreme gradient boosting

Categorical boosting

## ABSTRACT

Considering the global need for fossil fuels and its limited resources, maximum production from oil reservoirs is important. Acid treatment is a common method to stimulate oil reservoirs, but acid and oil interaction may form undesirable asphaltic sludge, and the prediction of this phenomenon by using machine learning models can be useful for field application. In this study, multi-layer perceptron (MLP), extreme gradient boosting (XGBoost), random forest (RF), and categorical boosting (CatBoost) as four machine learning models were employed to estimate the weight of asphaltic sludge formed. To this end, a data set containing 199 experimental data for seven different oil samples including a wide range of SARA fractions was used. The input parameters of the models included oil properties, acid properties, and the content of protective additives. The statistical analysis indicated that the MLP model has the highest accuracy with the coefficient of determination ( $R^2$ ) of 0.9517. In addition, the impact analysis of the input variables showed that the ferric ion concentration has the highest impact on asphaltic sludge formation with a relevance factor of 0.2755. Finally, using the leverage method, only 4 outlier data points were identified, which proved the validity of the model.

## 1. Introduction

In recent years, despite the increased attention to renewable energy, there is still demand on a global scale for fossil fuels. Considering the limited resources of fossil fuels and the high cost of drilling, it is essential that each production well reaches its maximum possible productivity. The acid stimulation process is the primary and regular way for well treatment, which aims to improve the fluid flow in the near wellbore region. Although, in some cases, it can cause new formation damages, which impedes the flow and thus reduces production. Acid-oil emulsion and sludge formation, aqueous phase trapping, clay swelling and dispersion, unfavorable wettability alteration, non-breaking gel-acid plugging, water and gas coning, insoluble precipitates raised from acid-rock side reactions, fines liberation and immigration are some of the various formation damages create during acid injection into the formation [1]. Among these, the formation and precipitation of asphaltic sludge has always been of interest [2]. Generally speaking, acid Sludge is an emulsion with a high viscosity, stabilized by organic particles rich in asphaltene [1]. Asphaltic sludge formation may have negative effects on near-wellbore permeability by blocking various pore spaces, altering rock wettability, and improving the stability of the emulsions [3]. Fig. 1

shows the sludge formed on a steel screen due to the incompatibility between acid and crude oil.

The formation of sludge is a complicated phenomenon, and its mechanism is not well known. It is affected by a wide range of parameters, including the type and strength of the acid [4], temperature [5], additives [6,7], iron concentration especially ferric [8], exposure time [9], mixing rate and acid mixture ratio (AMR) [1], and characteristics of crude oil [10].

On the other hand, using protective additives such as anti-sludge, anti-emulsion, and ferric ion reducer is the common method to prevent sludge formation [11]. The presence of suitable anti-sludge causes the formation of smaller size sludge, which may not be harmful to the oil reservoir's pores [1]. Also, when the acid and oil are emulsified in each other, more exposure time and surface area between the drops of acid and crude oil results in a significant amount of sludge and subsequent plugging of the pores, which is why anti-emulsion is utilized [12]. In addition, adding iron ion reducer causes the ferric ion to change to ferrous, subsequently decreasing the risk of sludging [6].

In order to prevent the formation damage and its costly consequence, it is essential to estimate the probable sludge formed before acidizing. Laboratory study as the compatibility test is a regular procedure for

\* Corresponding author. IOR/EOR Research Institute, School of Chemical and Petroleum Engineering, Shiraz University, Shiraz, Iran.

E-mail addresses: [mmohshirazi@gmail.com](mailto:mmohshirazi@gmail.com), [m.mohammadzadeh@shirazu.ac.ir](mailto:m.mohammadzadeh@shirazu.ac.ir) (M. Mohammadzadeh-Shirazi).



Fig. 1. The picture of the sludge formed on the steel screen.

checking the sludge formation tendency of crude oils. However, experimental testing is costly and time-consuming, and in some cases, the crude oil of the reservoir is not available, such as a newly drilled well that has not produced yet. Hence, the modeling and fast prediction of this phenomenon will be helpful.

Machine learning and data mining methods have been used in various petroleum engineering fields recently as reliable alternatives to expensive experimental tests as a result of advancements in computer science. In recent years, Wang et al. used machine learning models for relative permeability upscaling [13], Hui et al. used machine learning models to identify controlling factors of unconventional shale productivity [14], and also Kang et al. implemented deep learning models for prediction of drilling fluid lost-circulation zone [15]. Among these, some intelligent approaches have been developed within the field of formation damage. Kalam et al. [16] reported one of the first uses of the machine learning method to evaluate the formation damage caused by the invasion of drilling and completion fluids and additives. Their suggested artificial neural network (ANN) has been successful in predicting relative permeability and wettability and curves. Zuluaga et al. [17] have used fuzzy logic (FL) and artificial neural networks (ANNs) to estimate the effect of particle invasion on permeability decrease in unconsolidated rocks. Among implemented models, the ANN showed the best performance in forecasting permeability decrease utilizing flow-rate, initial porosity, particle concentration, and initial permeability. Rezaian et al. [18] have implemented ANNs in the formation caused by asphaltene deposition. The proposed ANN model predicted permeability reduction using initial permeability, asphaltene concentration, injection time, and velocity, the suggested model had an average absolute percent relative error of 8.3%. Foroutan and Moghadasi [19] have developed an ANN which predicted the relative permeability while mineral precipitation. This model was able to predict mineral precipitation with an average error of about 5%. Kamari et al. [20] have developed the coupled simulated annealing-least squares support vector machine (CSA-LSSVM), which predicted barium sulfate deposition at different NaCl concentrations and temperatures. The model had an average absolute relative deviation (AARD) of 0.0002%. Pourakabarian et al. [21] have developed an ANN model to predict sludge mass and volume using a data set including 120 compatibility test data, ignoring the effect of protective additives. For all data, the correlation coefficient of the

developed model was 0.9458. This study's main aim was to statistically analysis the results of the experimental tests. In a recent study, Larestani et al. [22] attempted to estimate the formation damage induced by mineral scaling. The best model developed in this research was gradient boosting decision tree (GBDT), with an average absolute percent relative error of 0.1465%. The significance of formation damage is undeniable in the oil and gas industry, as it directly influences the well productivity and ultimate recovery of reservoir. Formation damage control is critical for preserving the reservoir rock permeability to ensure stable production and cost reduction over the life of reservoirs. Machine learning models have shown their effectiveness in the application of controlling formation damage.

To the best of the authors knowledge, there has been an absence of a comprehensive model that involves the properties of crude oil, acid, and preventative additives in order to predict the asphaltic sludge formation. Furthermore, whereas the literature has concentrated on the formation damage caused by asphaltene deposition, the issue of the formation damage induced by asphaltic sludge has been less attended. The previous developed ANN model has not been taken into account the presence of preventative additives [21]. The inclusion of acid additives, together with more extensive dataset including various oil samples, and the comparative implementation of various machine learning models, can lead to the construction of an adequate predictive model for asphaltic sludge formation.

In the present study, machine learning models are developed to estimate the amount of asphaltic sludge formation, which is an unfavorable factor in the stimulation of oil reservoirs. These models can be useful and efficient since checking acid and crude oil compatibility experimentally is not always feasible due to operational or economic limitations. On the other hand, this precipitation phenomena is complex and the accurate estimation can be achieved through the machine learning models. In addition, the published experimental data with similar procedure is very limited and therefore, this objective has not been achieved yet. In this study, we used 199 published experimental data with known and similar procedure to model asphaltic sludge precipitation. The proposed model was well optimized and the importance of the independent parameters, notably operational additives, were founded. For this purpose, four intelligent models, namely multi-layer perceptron (MLP) neural network, extreme gradient boosting (XGBoost), random forest (RF), and categorical boosting (CatBoost) are used to estimate asphaltic sludge. To develop models, a data set consisting of crude oil properties, acid properties, the content of protective additives, and the formed sludge (g sludge/g oil) is assembled from the experimental tests. After assembling the data set, the data is divided into two parts, training and testing, and the models are trained with the training data and evaluated with the testing data. In order to accurately evaluate the models, statistical and graphical error analyses are performed, and the leverage method is used to prove the validity and applicability range of the model. In addition, the impact of input parameters on the formation of asphaltic sludge is evaluated and the most effective parameter is determined.

## 2. Asphaltic sludge and data collection

For reliable and successful modeling of a phenomenon, it is necessary to comprehend the fundamental mechanisms in order to interpret and select model inputs. For this purpose, in the first step, the mechanisms involved in the formation of asphaltic sludge and in the second step, the details of the collected data are described.

### 2.1. Asphaltic sludge formation mechanism

The shear force generated by the acid injection forms an acid-in-oil emulsion during the acid stimulation operation. With the formation of an acid-in-oil emulsion, the conditions are created for the formation of asphaltic sludge. In general, Fig. 2 depicts the sequence of mechanisms

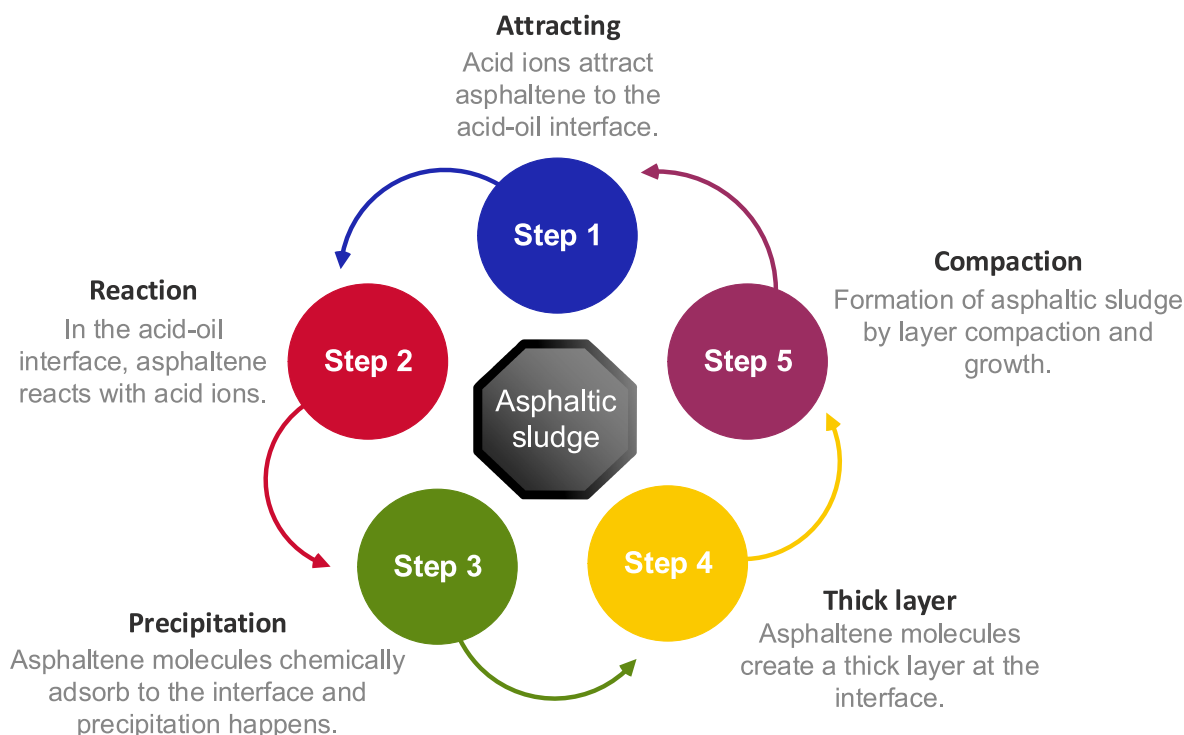


Fig. 2. The sequence of processes that would cause the formation of asphaltic sludge.

that cause the formation of asphaltic sludge. At a first step, the oil molecules and the hydrogen ions in the acid react together [23] as shown in Fig. 3. Hence, asphaltene would accumulate at the interface between the acid and crude oil phases [24], and then, acid-base reactions of the asphaltene molecules and their heteroatoms like S, N, and O would occur at the interface [7,25]. With regard to these interactions, asphaltenes can move from the crude oil to the interface and then accumulate there. As this happens, a protective layer forms around the acid drops, preventing them from contact [26]. After then, the layer that was formed at the interface will expand. As a result of this, the thickness of this layer of sludge rises, and it also gets more dense [7]. Therefore, it can be concluded that crude oil properties, acid properties, and protective additives play the main role in controlling the formation of asphaltic sludge.

## 2.2. Data collection

In this research, a credible data set was collected from our experimental tests to implement the models, and some of these data have been published in the literature [1]. The accurate bottle tests were performed

following the API Recommended Practice 42 [27] standard method with some modifications, which is reliable for the quantitative measurement of asphaltic sludge formation [1]. Hydrochloric acid was used in the experimental tests. Seven different crude oil samples were used with a wide range of SARA (Saturate, Aromatic, Resin, and Asphaltene) fractions to reliable the suggested model. Table 1 shows the chemical and physical properties of crude oil samples. The additives used in this study are shown in Table 2. The collected data set includes 199 data points. In the applied data set, each data point contains values for crude oil dynamic viscosity, saturate to aromatic ratio (Sa/Ar), asphaltene to resin ratio (As/Re), acid concentration (wt.%), acid to mixture ratio (AMR), mass concentration of ferric ion (mg/l), anti-sludge agent (wt.%), anti-emulsion agent (wt.%), ferric ion reducing agent (wt.%), and sludge mass (g sludge/g oil).

The statistical description of data is presented in Table 3. The first nine parameters were selected as the model's inputs, and the last was selected as the model's output. Before use, the data set was randomly divided into train and test data at 80 % and 20 %, respectively.

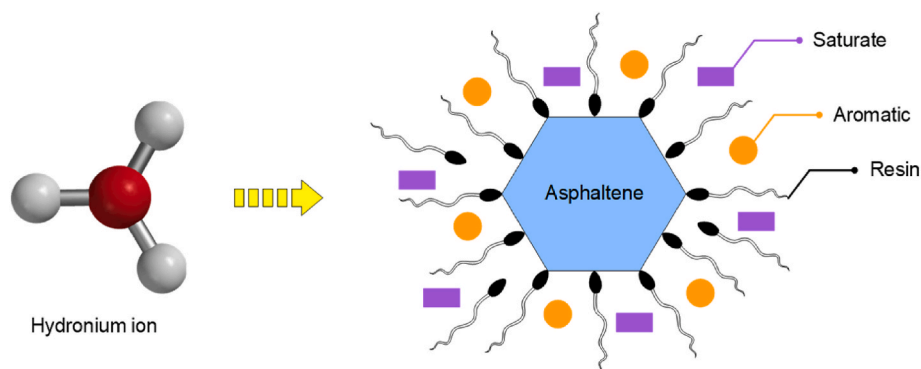
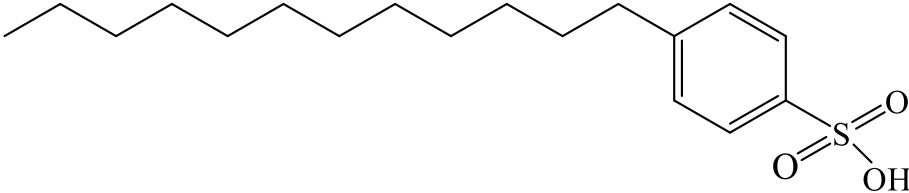
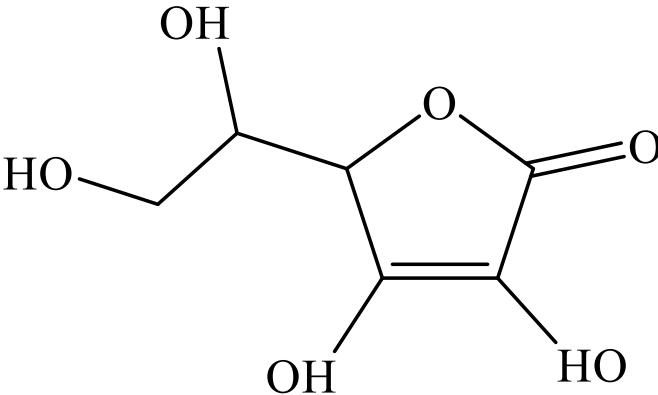
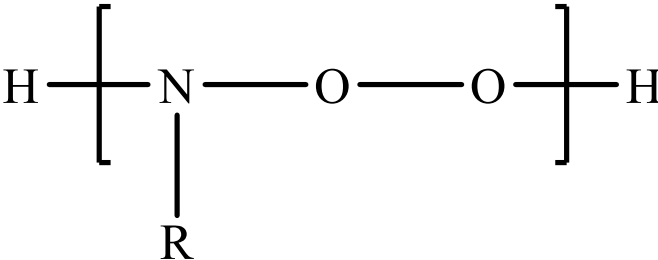


Fig. 3. Schematic illustration of a contact between hydronium ion and oil components.

**Table 1**  
Properties and characteristics of the crude oil samples.

Sample	Specific gravity (@ 25 °C)	Density (°API)	Viscosity (cp) (@ 25 °C)	As/Re	Sa/Ar	SARA Analysis (wt. %)			
						Sa	Ar	Re	As
A	0.9164	22.91	101.5	1.3226	1.1123	45.04	40.49	6.23	8.24
B	0.8731	30.57	25.20	0.1821	2.5250	63.96	25.33	9.06	1.65
C	0.8952	26.57	33.60	0.3773	1.2763	49.93	39.12	7.95	3.00
D	0.8711	30.94	18.20	0.0353	2.4639	66.97	27.18	5.65	0.20
E	0.9194	22.40	78.20	0.6846	1.1833	47.25	39.93	7.61	5.21
F	0.9330	19.00	2960	1.9350	1.4314	45.09	31.50	7.70	14.9
G	0.8903	27.40	80.30	0.8604	1.2721	48.75	38.32	6.95	5.98

**Table 2**  
Chemical structure and type of additives used.

Common name & chemical formula	Type	Structural formula
Dodecyl benzene sulfonic acid (C <sub>18</sub> H <sub>30</sub> O <sub>3</sub> S)	Anti-sludge	
Erythorbic acid (C <sub>6</sub> H <sub>8</sub> O <sub>6</sub> )	Ferric reducing	
N-alkylated polyhydroxyetheramines (NRO <sub>2</sub> H <sub>2</sub> )	Anti-emulsion	

### 3. Methodology

Four famous and cutting-edge models, namely random forest, extreme gradient boosting, multi-layer perceptron, and categorical boosting were used to achieve the results. All four methods have been previously documented as providing satisfying performance in previous

petroleum related studies. The models were developed according to the following methods:

#### 3.1. Random forest (RF)

The random forest technique is a kind of ensemble learning method

**Table 3**

The statistical analysis of the data used in this study.

Parameter	Minimum	Maximum	Average	Standard deviation
Viscosity of crude oil (cP)	18.2	2960	548.912	1102.401
Saturate to Aromatic ratio	1.112	2.525	1.719	0.596
Asphaltene to Resin ratio	0.035	1.935	0.68	0.648
Acid concentration (wt. %)	10.5	32.5	18.077	5.073
Acid to mixture ratio (cc/cc)	0.16	0.84	0.484	0.121
Ferric ion (mg/l)	0	3000	1587.204	1181.738
Anti-sludge agent (wt. %)	0	1	0.123	0.266
Anti-emulsion agent (wt.%)	0	1	0.137	0.272
Ferric reducing agent (wt.%)	0	0.5	0.064	0.168
Sludge mass (g sludge/ g oil)	0.002	0.177	0.055	0.043

where each tree is trained parallel to generate an ensemble of Decision Trees. In random forest, the greedy strategy determines the significance of each tree at each step [28]. In addition, RF can evaluate the importance of each input feature and conserve the most informative features [29]. The RF method includes a technique known as bagging or bootstrap aggregation to optimize the variable selection and variety of the trees. The model will determine how to divide the input data into multiple sub-datasets based on the population of the trees. Bagging, a form of random sampling technique, assigns one-third of the data for the training stage of a subtree development procedure, while the remains are referred to as out-of-bag (OOB) samples. Furthermore, when employing the random forest algorithm, the cross-validation approach is

not required since multiple bagging during the training process avoids over-fitting [30], which are the advantages of using RF applied in this study. The random forest structure is shown in Fig. 4.

Assume  $D$  to be the training data set with  $n$  observations,  $D = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$ , and if  $D_t$  denotes the training data set for the tree  $h_t$ , then the predicted output related to the OOB data set of sample  $x$  can be defined as follows [31]:

$$H^{oob}(x) = \operatorname{argmax} \sum_{t=1}^T I(h_t(x) = y) \tag{1}$$

The error of the OOB data set is described as follows:

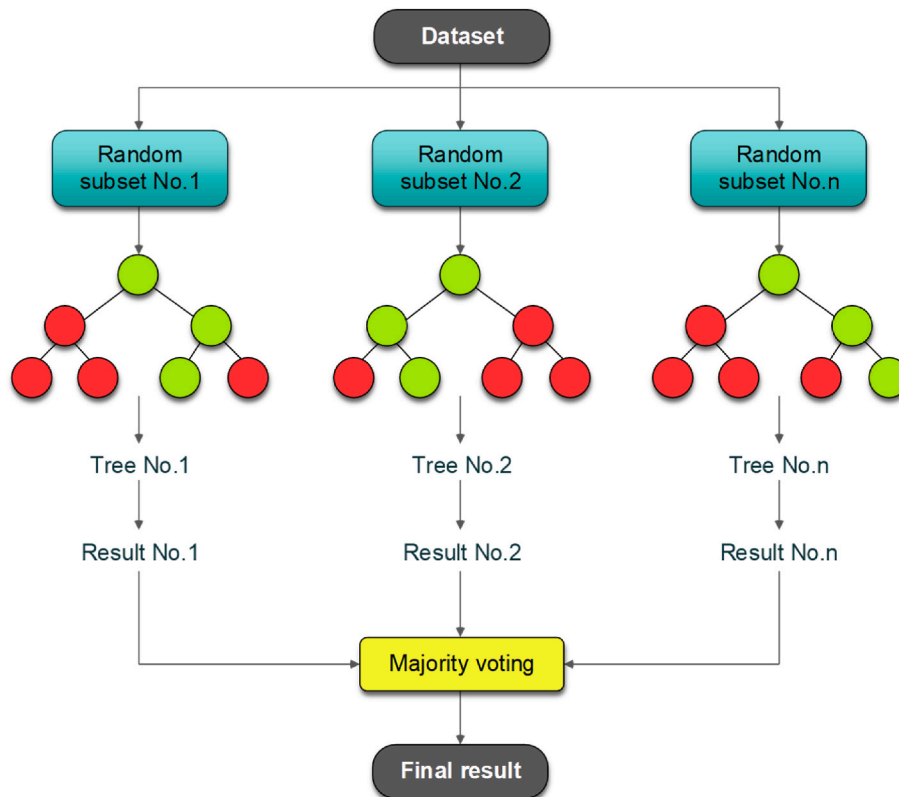
$$\varepsilon^{oob}(x) = \frac{1}{|D|} \sum_{(x,y) \in D} I(H^{oob}(x) \neq y) \tag{2}$$

The RF should operate randomly, and this feature is managed by the parameter  $K$ , which is formulated as  $K = \log_2 d$  [30].

### 3.2. Extreme gradient boosting (XGBoost)

The XGBoost method is a subset of the gradient boosting decision tree (GBDT) group of tree-based models. Many data science issues are accurately solved using a parallel tree-boosting method supplied by XGBoost. XGBoost was developed to be highly effective and flexible. Tree-based ensemble approaches use a set of regression trees (CARTs) to determine the most optimal fit for a given set of training data, using a regularized objective function. As shown in Fig. 5, each CART has a root node together with internal and leaf nodes. Based on the binary decision approach, the root node, which contains all data, is classed as an internal node, while the leaf nodes indicate the final categories. Gradient boosting progressively generates a strong group from a set of basic CARTs. Moreover, the weight of each CART must be synced throughout the training procedure [32].

A group of  $n$  trees should be trained in accordance with the following



**Fig. 4.** Schematic illustration of the random forest method.



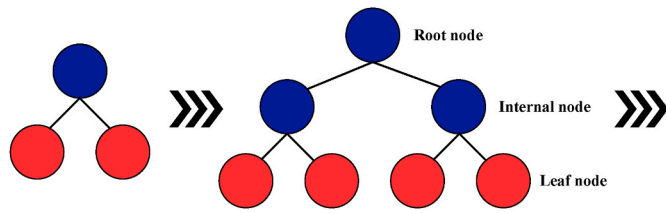


Fig. 5. Schematic illustration of the CART model (Level-wise tree growth).

equation for a given data set in order to model  $y$  as the output:

$$\hat{y} = \sum_{k=1}^N f_k(X_i), f_k \in f \text{ With } f = \{f(X) = \omega_{q(x)}\}, (q : \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T) \quad (3)$$

Also,  $n$  and  $m$  relate to instances and dimension features, respectively. Additionally, the sample is converted into a binary leaf index by the decision rule  $q(x)$ . Regression tree space is denoted by  $f$ ,  $\omega$  denotes the leaf weight,  $f_k$  is the  $k^{th}$  independent tree, and  $T$  is the number of leaves.

Using the following equation, the regularized objective function is decreased to determine the group of trees:

$$L = \sum_i^n l(\hat{y}_i, y_i) + \sum_k^N \Omega(f_k) \text{ With } \Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (4)$$

where  $l$  represents the differentiable convex loss function, and  $\Omega$  is the regularization term that reduces the complexity of the model and assists in preventing overfitting.  $\gamma$  indicates the minimal loss reduction needed to split a new leaf, while  $\lambda$  denotes the regulation coefficient. It is important to notice that the parameters  $\gamma$  and  $\lambda$  in these equations help to decrease model variance and overfitting [33].

The learning factor rate is applied to the newly added weights after each boosting stage in XGBoost. This reduces the risk of overfitting by limiting the impact of future new trees on existing trees [34].

### 3.3. Multi-layer perceptron (MLP) neural network

In the 1980s, the MLP, the most well-known artificial neural network (ANN) with several layers, was introduced [35]. MLP is a powerful tool

for solving complex nonlinear problems, it can process large amounts of input data without a problem, and it is possible to have the same amount of accuracy with a smaller sample size. MLP is a type of feedforward ANNs consisting of several layers. The input layer is the first layer that is relevant to the input data, the output layer is the last layer that corresponds to the output of the model, and the layers in the middle that process the data are hidden layers [36]. Each neuron in the hidden layers will connect to each neuron in the next and previous layers. The amount of each neuron multiplying in its corresponding weight in the previous layer is added together, and a bias factor is added to these values. The resultant value is then transmitted to an activation function [37]. To achieve an accurate and efficient ANN, it is necessary to optimize the number of hidden layers and their neurons [38]. Typically, the number of neurons is estimated by trial and error [39]. The MLP model's efficiency depends on the optimization techniques used to train the model [40]. In this study, the MLP model based on the Adam optimization technique [41] is developed. Fig. 6 shows a schematic of the MLP neural network developed in this study.

### 3.4. Categorical boosting (CatBoost)

CatBoost is a kind of the categorical gradient boosting technique, that uses binary decision trees as its primary predictor as shown in Fig. 7 [42,43]. This method functions with minimal information loss on categorical features. The most important notion for comprehending CatBoost's approach is based on the differentiation between training and testing datasets [44]. In addition, the indicator function 1 is another important notion for understanding how CatBoost encodes the categorical features, as seen below [44]:

$$\text{Indicator function } 1_{k,t} = \begin{cases} 1, & \text{if } k = t \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The aforementioned function is an important part of the formula used by CatBoost to convert the categorical data into quantitative values. Furthermore, the boosting mechanism that is used by the CatBoost technique takes use of categorical columns, and processing techniques are used in these columns. The most critical are target-based statistics and One-Hot-Max-Size (OHMS). The main phases of the CatBoost technique are creating a random subgroup of variable records, changing labels to quantitative values, and transforming feature to

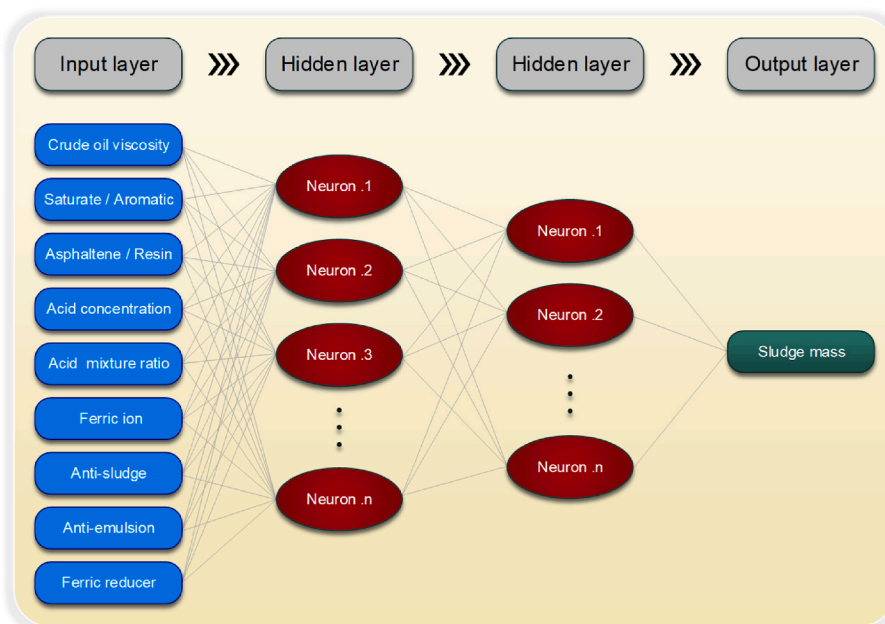


Fig. 6. Schematic illustration of the MLP model developed in this study.

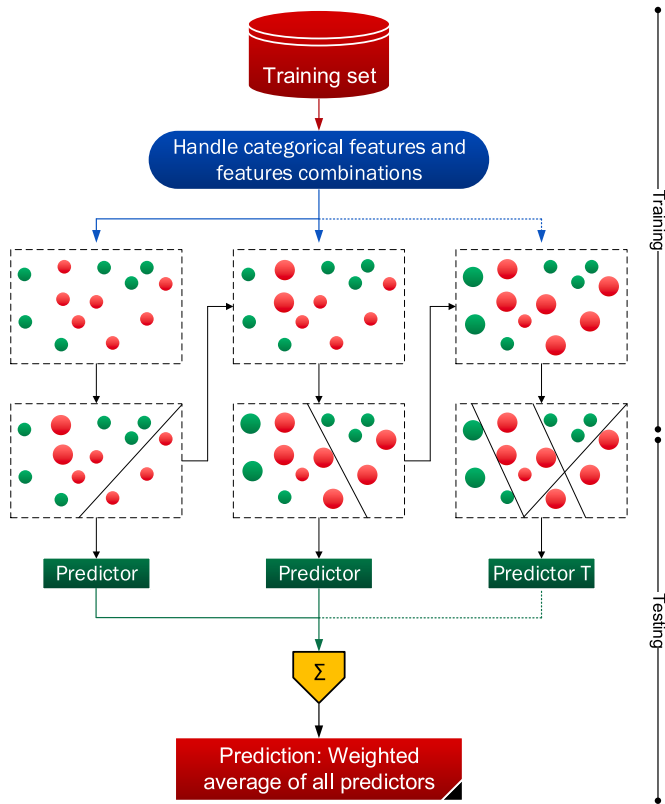


Fig. 7. Schematic illustration of the CatBoost model structure.

numerical format [45]. To avoid over-fitting, the CatBoost algorithm performs random permutations to predict leaf values while choosing the tree construction, which provides a significant advantage. In this method, the predicted value is obtained as below [42]:

$$T = H(x_i) = \sum_{n=1}^n c_n 1_{\{x \in R_n\}} \quad (6)$$

where,  $H$  denotes the DT function,  $x_i$  represents the explanatory variable, and  $R_n$  defines the disjoint section related to the tree's leaves.

### 3.5. Performance evaluation of models

#### 3.5.1. Statistical evaluation of models

To statistically evaluate the performance and dependability of the models, various statistical parameters, including coefficient of determination ( $R^2$ ), mean absolute error (MAE), standard deviation (SD), and root mean square error (RMSE) are served as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^N (y_{i,exp} - \bar{y}_{exp})^2} \quad (7)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_{i,exp} - y_{i,pred}| \quad (8)$$

$$SD = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \left( \frac{y_{i,exp} - y_{i,pred}}{y_{i,exp}} \right)^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{i,exp} - y_{i,pred})^2} \quad (10)$$

#### 3.5.2. Graphical evaluation of models

Graphical error assessments are used to illustrate and compare the models' results. Cross plots, error distribution graphs, and cumulative frequency graphs were used to visually evaluate the performance of the suggested models. An overview of these graphical techniques is provided below:

The cross plot for the target parameter shows the predicted value compared to the actual value. Increased data point concentration around the unit-slope line in such graphs indicates a model with more accuracy. Graph of error distribution illustrates the percent relative error versus independent variables. Finally, the cumulative frequency graph displays the absolute relative error versus the cumulative error frequency.

#### 3.6. Development of models

In order to develop the models and prevent overfitting, the hyperparameters were optimized using the KerasTuner method for the MLP model and the gridsearch method for the RF, XGBoost, and CatBoost models. Table 4 provides the optimum hyperparameter values for the models.

## 4. Results and discussion

In this study, four effective algorithms that have already obtained acceptable results for modeling, were employed to predict the amount of asphaltic sludge. After selecting the optimal hyperparameter values, the models were developed using training data, and then testing data was served to evaluate the models. Fig. 8 shows the process of data preparation and model development to predict sludge formation.

#### 4.1. Basic data analysis

In the current research, the Pearson correlation matrix of the variables in the data set was determined for the purposes of basic data analysis. Fig. 9 illustrates the Pearson correlation matrix, which provides insight into relationships between different variables. Each cell displays the Pearson coefficient between two variables. The Pearson coefficient, which ranges from  $-1$  to  $+1$ , is employed to measure the linear relationship between two attributes. A negative value indicates a negatively linear relationship between two attributes, whereas a positive value indicates a positively linear relationship. A value of 0 indicates the absence of a linear relationship.

At first glimpse, it appears that there is a significant positive relationship between the parameters of acid concentration, AMR, and

Table 4  
The optimum values of hyperparameters for the models.

Hyperparameter	Search range	Model			
		RF	XGBoost	CatBoost	MLP
max_depth	5–50	17	6	–	–
max_leaf_nodes	10–140	70	–	–	–
n_estimators	20–2500	152	–	–	–
max_features	'Sqrt', 'log2', 1,	1	–	–	–
	2, 3				
eta	0.01–0.5	–	0.1	–	–
subsample	0.1–1	–	0.3	–	–
colsample_bytree	0.01–1	–	0.94	–	–
iterations	20–2000	–	–	1000	–
Depth	2–20	–	–	7	–
l2_leaf_reg	2–20	–	–	4	–
learning_rate	0.01–1.5	–	–	1.258	0.002
Hidden layers	1–256	–	–	–	[32*72]
Epoches	100–500	–	–	–	400
Batch size	2–32	–	–	–	8
Activation function	Sigmoid–ReLU	–	–	–	ReLU

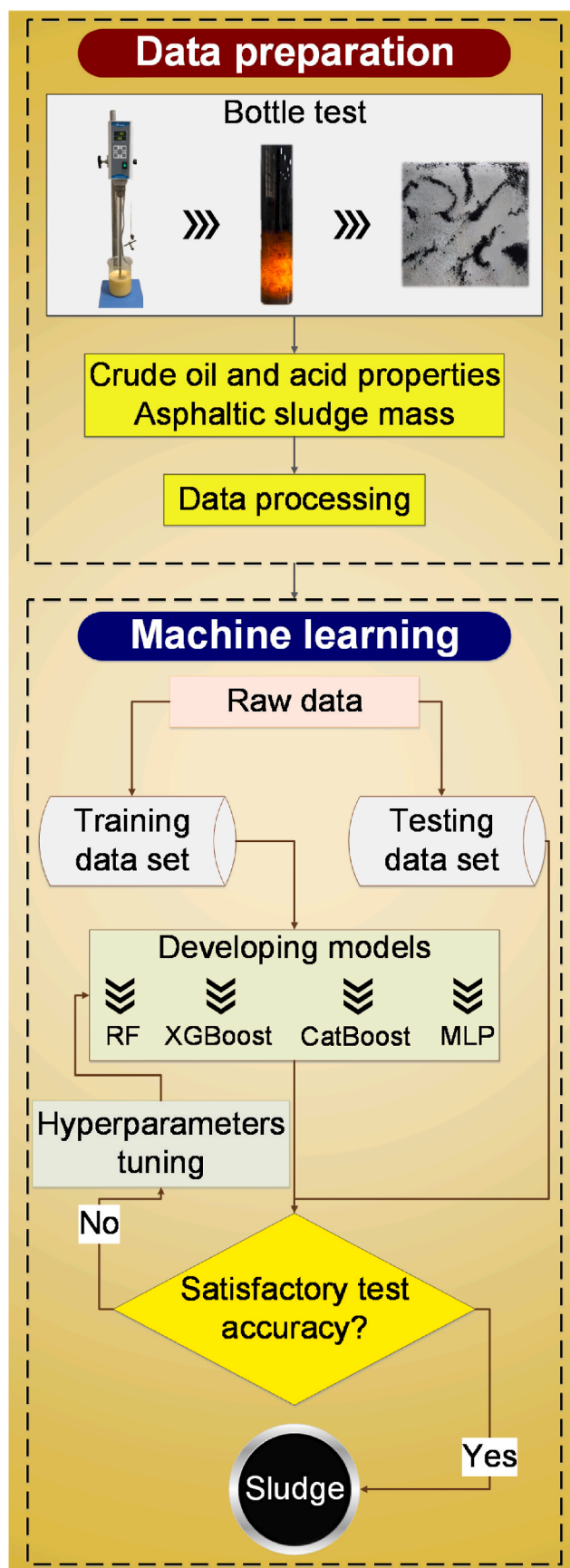


Fig. 8. The process of preparing data and developing models to predict sludge formation.

especially ferric ion with asphaltic sludge. Furthermore, it is noteworthy to mention that three additives, namely anti-sludge, anti-emulsion, and ferric reducing, exhibit a negative relationship with asphaltic sludge, aligning with expectations. With regard to the relationship between input features, a strong relationship is seen between viscosity and asphaltene-to-resin ratio. It is important to point out that experimental results have shown that the viscosity of oil samples increases with increasing asphaltene concentration under constant temperature conditions [46].

#### 4.2. Statistical evaluation of the developed models

Various statistical parameters, including  $R^2$ ,  $MAE$ ,  $SD$ , and  $RMSE$ , were used to evaluate the efficiency of the models. The calculated values are shown in Table 5. As seen in the table, the MLP model has provided the most accurate prediction, with an  $RMSE$  of 0.0115. XGBoost, RF, and CatBoost models have predicted with  $RMSE$  values of 0.0121, 0.0160, and 0.0132 respectively. The high accuracy of a model for prediction can sometimes be a result of overtraining, which is another critical issue that should be considered. In order to prevent this, the outcomes of training and test results will be compared. If there is a significant gap between the metrics of the training and test results, the model can be overfitted. According to Table 5, there is no significant gap between the metrics of the training and test results; therefore, the models were not overfitted.

#### 4.3. Graphical evaluation of developed models

Several graphical error analyses were conducted to demonstrate the reliability of the developed methods. The cross plots for the developed models are shown in Fig. 10. The greater concentration of data close to the unit-slope line shows that predicted values are closer to the actual ones, and the model is more reliable, consequently. The superior performance of the developed MLP model is evident compared to other applied models.

The error distribution graph for the proposed models is shown in Fig. 11. This graph illustrates experimental data versus relative error. The forecasted data error decreases as the points condense and place close to the zero line. As illustrated in Fig. 11, the points in the MLP model are closer to the zero line, which shows better accuracy and a lower relative error than the other models.

In addition, Fig. 12 shows the cumulative frequency graph, which shows the cumulative frequency of all data versus the absolute relative errors. Higher curvature indicates a model with better accuracy. This indicates that a model has more predicted data with a lower absolute relative error compared to other models. Based on Fig. 12, it can be concluded that the MLP model is more accurate than other models.

#### 4.4. Trend analysis

Trend analysis was performed to assess the validity of the MLP model to predict asphaltic sludge formation. Hence, the experimental data of asphaltic sludge for two new distinct crude oil samples was obtained at different AMR values, and they were compared by the values predicted by MLP model. The characteristics of these crude oils are outlined in Table 6. As illustrated in Fig. 13, there is a noticeable rise in the asphaltic sludge of oil sample H, when the AMR increases. On the other hand, for oil sample I, the weight of asphaltic sludge has increased slightly. As can be seen, The MLP model has predicted the trend for both oil samples with reasonable accuracy.

#### 4.5. Analyzing the quality of experimental data and the model's applicability range

The leverage approach was applied for the MLP model (as the optimal model) to identify suspicious data as well as the model's application range. The Leverage approach, with a clear visual



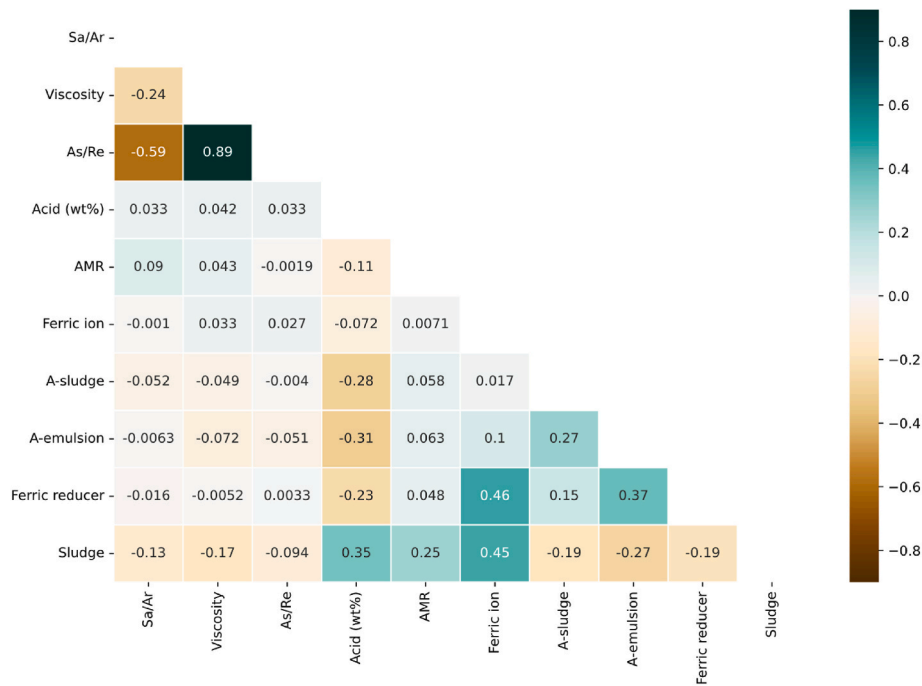


Fig. 9. The Pearson correlation matrix of the variables in the data set.

Table 5

Calculated statistical parameters for the suggested models.

Model		$R^2$	MAE	SD	RMSE	No. of data
MLP	Train	0.9690	0.0037	0.0424	0.0069	159
	Test	0.9517	0.0073	0.0485	0.0115	40
	Total	0.9645	0.0045	0.0436	0.0081	199
RF	Train	0.9477	0.0067	0.0379	0.0096	159
	Test	0.8947	0.0116	0.0428	0.0160	40
	Total	0.9338	0.0077	0.0388	0.0112	199
XGBoost	Train	0.9304	0.0077	0.0386	0.0111	159
	Test	0.9395	0.0096	0.0446	0.0121	40
	Total	0.9328	0.0081	0.0398	0.0113	199
CatBoost	Train	0.9342	0.0078	0.0399	0.0107	159
	Test	0.9284	0.0103	0.0431	0.0132	40
	Total	0.9327	0.0083	0.0404	0.0113	199

representation in William’s plot, is one of the crucial methods for identifying outliers [47–49]. Fig. 14 shows William’s plot for the proposed MLP model. Differences between the model’s predictions and experimental data are represented by the standardized residual and the hat matrix, which is described as follows [50]:

$$H = X(X^T X)^{-1} X^T \tag{11}$$

$$H^* = \frac{3(n+1)}{m} \tag{12}$$

In this equation,  $X^T$  denotes the transpose of the matrix  $X$ . Additionally,  $H^*$  is determined as the value for the Leverage limit, where the parameters  $m$  and  $n$  indicate the total amount of data in the data set and the number of input parameters, respectively. If the majority of the data points fall between the boundaries  $-3 \leq R \leq 3$  and  $0 \leq H \leq H^*$ , the developed model is regarded as reliable, and its evaluations are done within the applicability range.

As seen in Fig. 14, most data fall between  $-3 \leq R \leq 3$  and  $0 \leq H \leq 0.1507$ , and just four data points were identified as suspicious. Based on these findings, it can be concluded that the experimental data are of great quality and the MLP model is reliable.

#### 4.6. Impact analysis of input variables

The impact analysis of the parameter is used to determine the influence of input parameter on the output of model. This analysis is conducted using the following relevancy factor [51]:

$$r(Inp_k, out) = \frac{\sum_{i=1}^n (Inp_{k,i} - Inp_{avg,k})(Out_i - Out_{avg})}{\sqrt{\sum_{i=1}^n (Inp_{k,i} - Inp_{avg,k})^2 - \sum_{i=1}^n (Out_i - Out_{avg})^2}} \tag{13}$$

Hence,  $Out_i$  means the  $i^{th}$  value of the predicted output, while  $Out_{avg}$  defines the average of the output data.  $Inp_{avg,k}$  and  $Inp_{k,i}$  denote the average value and the  $i^{th}$  value of the  $k^{th}$  input, respectively. This factor, which ranges from  $-1$  to  $1$ , shows the impact of inputs data on the model’s output in three manners as follows [52]:

- 1 If the relevancy factor is less than 0, by increasing the input variable, the value of the output variable decreases.
- 2 If the relevancy factor is equal to 0, it can be concluded that there is no relation between the input variable and the output variable, or the relation is not monotonic.
- 3 If the relevancy factor is more than 0, the input variable has an increasing impact on the output variable.

Fig. 15 indicates the relevance factor values for the input variables of the MLP model. This chart shows that ferric ion, acid concentration, and acid to mixture ratio (AMR) have incremental impacts on sludge formation. Contrarily, ferric ion reducing agent, anti-emulsion agent, anti-sludge agent, asphaltene to resin ratio, viscosity of crude oil, and saturate to aromatic ratio have diminishing impacts on this. Among the parameters with the highest relevancy factor, ferric ion has the highest positive relevancy factor with a relevancy factor of 0.2755, which means that with the increase of ferric ion, the amount of sludge also increases; the observed impact can be related to the significant influence of ferric ions and their complexes on the acid-oil interface, leading to increased emulsion stability and ultimately increasing asphaltic sludge formation [53]. On the other hand, anti-emulsion agent has the highest negative relevancy factor with a relevancy factor of  $-0.2735$ , which means that

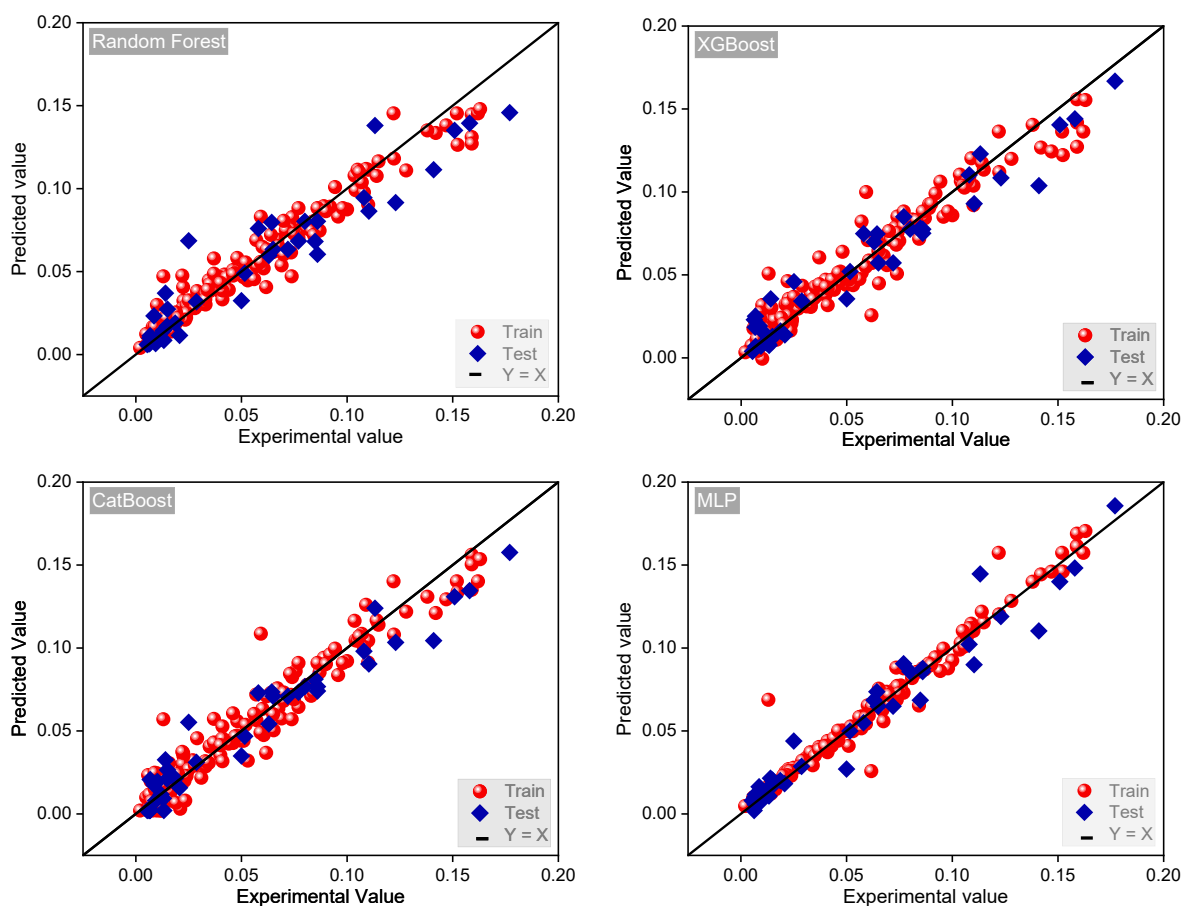


Fig. 10. Cross plots of the developed models.

with the increase of anti-emulsion agent, the amount of sludge decreases. In fact, the sludge is a viscous emulsion stabilized by organic components; therefore, utilizing of the anti-emulsion additive inhibits the stability of the acid-oil emulsion, which decreases the formation of asphaltic sludge [1].

In terms of the impact of additives, anti-emulsion agent, ferric ion reducing agent, and anti-sludge agent are placed with the relevance factor of  $-0.2735$ ,  $-0.2032$ , and  $-0.1899$ , respectively. This finding proved the importance of ferric ion reducing agent to suppress sludging besides other regular protective additives. Parameters related to the crude oil properties, such as viscosity, asphaltene to resin ratio, and saturate to aromatic ratio, are placed with the relevance factor of  $-0.1681$ ,  $-0.0796$ , and  $-0.0131$ , respectively; this indicates that the viscosity of crude oil has a more impact on the formation of asphaltic sludge than other crude oil properties. Finally, from the view of acid properties, ferric ion, acid concentration, and acid to mixture ratio are placed with the relevance factor of  $0.2755$ ,  $0.1181$ , and  $0.0912$ , respectively.

As a result, among four superior machine learning models, the MLP as the best model provides an accurate prediction of asphaltic sludge formation; this helps the sludge formation prediction prior to acidizing operations in situations where compatibility testing is not possible or economical. Also, the impact of various parameters, which is important due to the complexity of the relationships governed by this phenomenon, was clarified using relevancy factor. Therefore, the values of the effective parameters can be more accurately designed before the acid stimulation operation to control the formation of asphaltic sludge more effectively.

Although the current study offers findings and insights about the use of machine learning models for asphaltic sludge modeling, it is crucial to

recognize the inherent limitations of this research. One significant limitation arises from the fact that machine learning models are often constructed using a particular dataset, which might potentially restrict their effectiveness when applied to other domains. Consequently, these models typically need retraining in order to be applicable to new domains. Another limitation of this study is that although the MLP model exhibited superior performance, it is commonly perceived as a black box model owing to certain attributes that creates challenges in comprehending the rationales behind a model's decision-making process. Therefore, in situations where comprehending the decision-making process is essential, a dearth of transparency may undermine trust in the predictions made by the model.

## 5. Summary and conclusions

It is essential that every producing well achieve its maximum feasible production due to the limited resources of fossil fuels and the high cost of drilling. Therefore, in order to prevent the formation of asphalt sludge, it is necessary to predict this phenomenon for maximum production from oil reservoirs. For this purpose, in the present study, a data set containing 199 compatibility experimental data of asphaltic sludge formation for seven different crude oil samples covering a wide range of SARA analysis values was collected. Nine input parameters related to crude oil properties, acid properties, and the amount of additives, were used to predict the formed sludge mass. To this end, four machine learning models, namely multi-layer perceptron (MLP) neural network, random forest (RF), extreme gradient boosting (XGBoost), and Categorical boosting (CatBoost) were developed, and the following results were obtained:

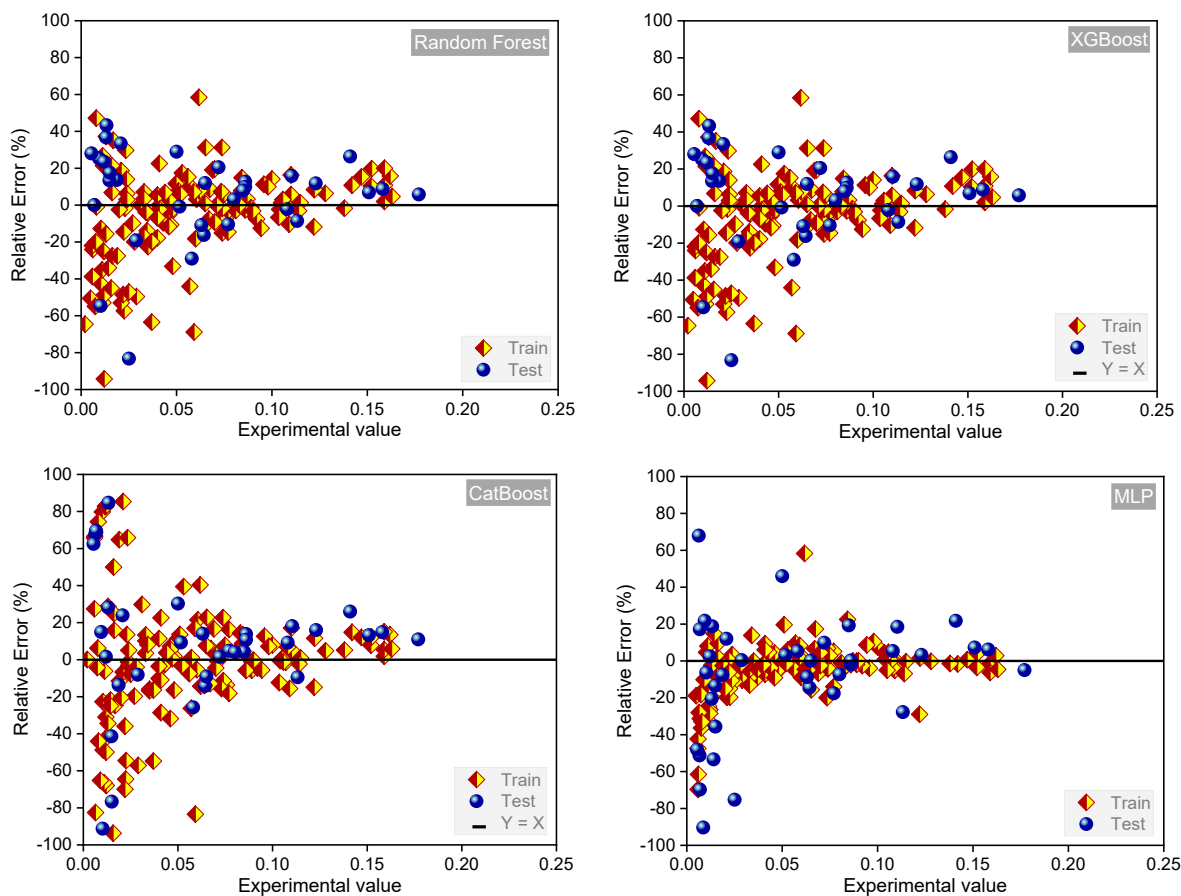


Fig. 11. Error distribution graphs of the developed models.

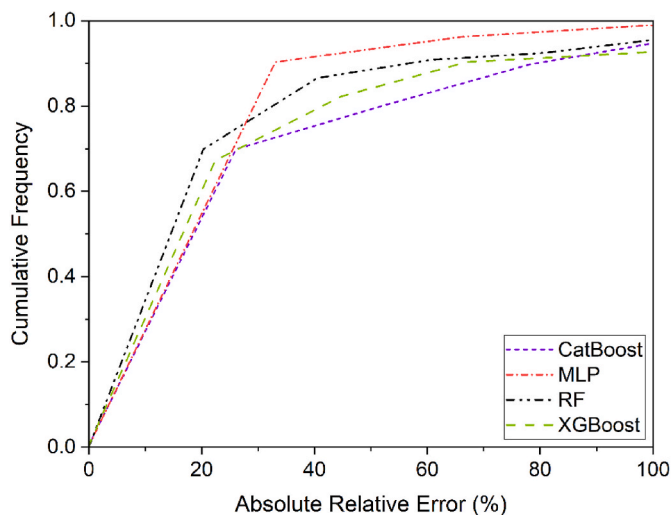


Fig. 12. The cumulative frequency plot of the proposed models.

1. Among the proposed models, the MLP model showed the best performance with *RMSE* value of 0.0115. The *RMSE* values for XGBoost, CatBoost, and RF models were obtained as 0.0121, 0.0132, and

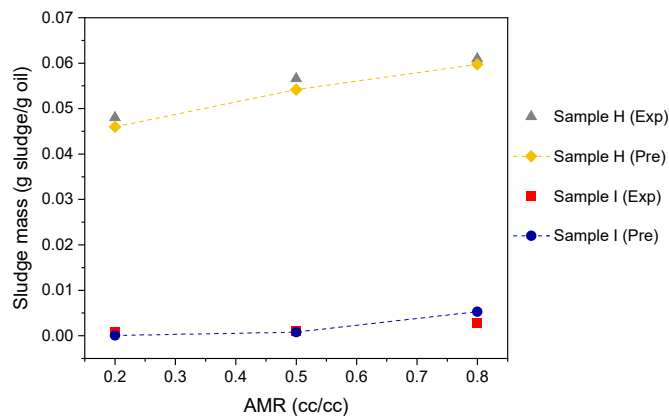


Fig. 13. Predictions of MLP model with experimental asphaltic sludge data for two different oil samples.

**Table 6**  
Properties and characteristics of the crude oil samples served for validation analysis.

Sample	Density ( $^{\circ}$ API)	Viscosity (cp) (@ 25 $^{\circ}$ C)	As/Re	Sa/Ar	SARA Analysis (wt. %)			
					Sa	Ar	Re	As
H	20.30	140	1.3076	1.2837	47.5	37	6.5	8.5
I	27.86	56	1.0714	1.2368	47	38	7	7.5

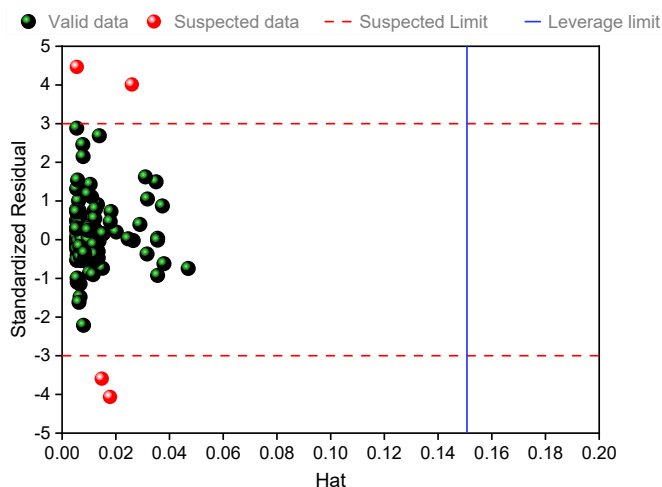


Fig. 14. William's plot to identify suspected data and validate the range of MLP's applicability.

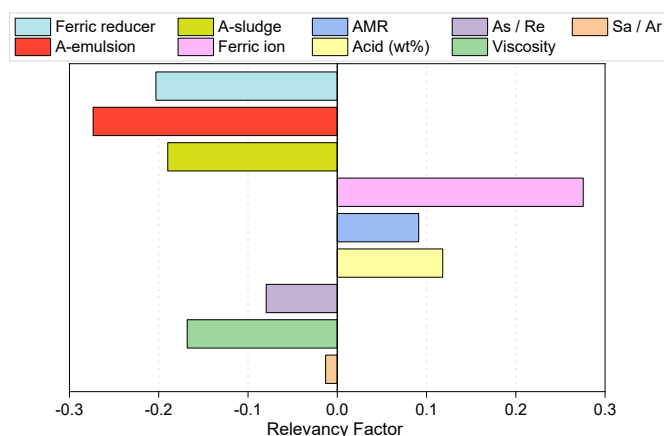


Fig. 15. Evaluation of the impact of input variables on the formation of sludge using relevancy factor.

0.0160, respectively; this superiority can be attributed to the MLP model's capacity to interpret complex nonlinear relationships and intricate data patterns.

- The MLP model and the data were proven to be valid and reliable using the leverage method, which indicated that just 4 data points were suspected.
- The impact analysis of the input variables showed that the impact of the input variables was as follows: Ferric ion > Anti-emulsion agent > Ferric ion reducing agent > Anti-sludge agent > Viscosity > Acid concentration > Acid to mixture ratio > Asphaltene to Resin ratio > Saturate to Aromatic ratio. Except for ferric ion, acid concentration, and acid to mixture ratio, the other factors have a decreasing impact on the formation of asphaltic sludge.
- Anti-emulsion agent with a relevance factor of  $-0.2735$  was the most effective additive, followed by ferric ion reducing agent and anti-sludge agent with relevance factors of  $-0.2032$  and  $-0.1899$ , respectively.
- The results obtained from the trend analysis of the MLP model in two distinct oil samples, including different AMR ranges, demonstrated reliable performance for this model.
- In some operational cases, it is not possible or economical to do compatibility tests before the acid stimulation operation, so it is necessary to estimate the formation of asphaltic sludge before it to minimize formation damage consequences. To achieve this, an

accurate estimation can be made using the proposed machine learning model.

#### Credit author statement

**Sina Shakouri:** Investigation, Conceptualization, Data Curation, Methodology, Validation, Visualization, Software, Formal analysis, Writing - Original Draft. **Maysam Mohammadzadeh Shirazi:** Resources, Supervision, Conceptualization, Data Curation, Methodology, Validation, Writing - Review & Editing.

#### Declaration of competing interest

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

#### Data availability

Data will be made available on request.

#### Abbreviations

- AMR = Acid to mixture ratio
- As = Asphaltene
- Ar = Aromatic
- CatBoost = Categorical boosting
- cP = Centipoise
- MAE = Mean absolute error
- MLP = Multi-layer perceptron
- Sa = Saturate
- SD = Standard deviation
- Re = Resin
- RF = Random Forest
- RMSE = Root mean square error
- R2 = Coefficient of determination
- wt.% = Weight percent
- XGBoost = Extreme gradient boosting

#### References

- Mohammadzadeh Shirazi M, Ayatollahi S, Ghotbi C. Damage evaluation of acid-oil emulsion and asphaltic sludge formation caused by acidizing of asphaltic oil reservoir. *J Petrol Sci Eng* 2019;174:880–90.
- Mirkhoshshah SM, Mahani H, Ayatollahi S, Mohammadzadeh Shirazi M. Pore-scale insights into sludge formation damage during acid stimulation and its underlying mechanisms. *J Petrol Sci Eng* 2021;196:107679.
- Alrashidi H, Farid Ibrahim A, Nasr-El-Din H. Bio-oil Dispersants effectiveness on AsphalteneSludge during carbonate acidizing treatment. In: SPE Trinidad and Tobago section energy resources conference. OnePetro; 2018.
- Abdollahi R, Shadizadeh SR, Zargar G. Experimental investigation of acid-induced sludge precipitation: using acid additives in Iran. *Energy Sources, Part A Recovery, Util Environ Eff* 2014;36(16):1793–9.
- Kharisov RY, Folomeev AE, Sharifullin AR, Bulgakova GT, Telin AG. Integrated approach to acid treatment optimization in carbonate reservoirs. *Energy Fuels* 2012;26(5):2621–30.
- Jacobs I, Thorne M. Asphaltene precipitation during acid stimulation treatments. In: SPE formation damage control symposium. OnePetro; 1986.
- O'Neil B, Maley D, Lalchan CA. Prevention of acid-induced asphaltene precipitation: a comparison of Anionic Vs. Cationic surfactants. *J Can Petrol Technol* 2015;54:49–62.
- Delorey JA, Taylor RS. Recent studies into iron/surfactant/sludge interactions in acidizing. In: Annual technical meeting; 1985. All Days.
- Rietjens M. Sense and non-sense about acid-induced sludge. In: SPE European formation damage conference. OnePetro; 1997.
- Suzuki F. Precipitation of asphaltic sludge during acid stimulation treatment: cause, effect, and prevention. In: SPE western regional meeting. OnePetro; 1993.
- Mirvakili A, Rahimpour MR, Jahanmiri A. Effect of a cationic surfactant as a chemical Destabilization of crude oil based emulsions and asphaltene stabilized. *J Chem Eng Data* 2012;57(6):1689–99.
- Abbasi A, Malayeri MR. Stability of acid in crude oil emulsion based on interaction energies during well stimulation using HCl acid. *J Petrol Sci Eng* 2022;212:110317.

- [13] Wang Y, Li H, Xu J, Liu S, Wang X. Machine learning assisted relative permeability upscaling for uncertainty quantification. *Energy* 2022;245:123284.
- [14] Hui G, Chen Z, Wang Y, Zhang D, Gu F. An integrated machine learning-based approach to identifying controlling factors of unconventional shale productivity. *Energy* 2023;266:126512.
- [15] Kang Y, Ma C, Xu C, You L, You Z. Prediction of drilling fluid lost-circulation zone based on deep learning. *Energy* 2023:127495.
- [16] Kalam M, Al-Alawi S, Al-Mukheini M. Assessment of formation damage using artificial neural networks. In: International symposium on formation damage control: Lafayette LA; 1996. p. 301–9. 14–15 February 1996.
- [17] Zuluaga E, Alvarez H, Alvarez J. Prediction of permeability reduction by external particle invasion using artificial neural networks and fuzzy models. *J Can Petrol Technol* 2002;41(6).
- [18] Rezaian A, Kordestany A, Haghghat Sefat M. An artificial neural network approach to formation damage prediction due to Asphaltene deposition. In: Nigeria annual international conference and exhibition. OnePetro; 2010.
- [19] Foroutan S, Moghadasi J. A neural network approach to predict formation damage due to calcium sulphate precipitation. In: SPE European formation damage conference & exhibition. OnePetro; 2013.
- [20] Kamari A, Gharagheizi F, Bahadori A, Mohammadi AH. Rigorous modeling for prediction of barium sulfate (barite) deposition in oilfield brines. *Fluid Phase Equil* 2014;366:117–26.
- [21] Pourakaberian A, Ayatollahi S, Shirazi MM, Ghotbi C, Sisakhti H. A systematic study of asphaltic sludge and emulsion formation damage during acidizing process: experimental and modeling approach. *J Petrol Sci Eng* 2021;207:109073.
- [22] Larestani A, Mousavi SP, Hadavimoghaddam F, Hemmati-Sarapardeh A. Predicting formation damage of oil fields due to mineral scaling during water-flooding operations: gradient boosting decision tree and cascade-forward back-propagation network. *J Petrol Sci Eng* 2022;208:109315.
- [23] Hinojosa RA. Asphaltene damage in matrix acidizing. Texas A&M University; 1996.
- [24] Houchin L, Dunlap D, Arnold B, Domke K. The occurrence and control of acid-induced asphaltene sludge. In: SPE formation damage control symposium. OnePetro; 1990.
- [25] Rietjens M, Nieuwpoort M. Acid-sludge: how small particles can make a big impact. In: SPE European formation damage conference. OnePetro; 1999.
- [26] Wong T, Hwang R, Beaty D, Dolan J, McCarty R, Franzen A. Acid-Sludge characterization and remediation improve well productivity and save costs in the Permian basin. *SPE Prod Facil* 1997;12(1):51–8.
- [27] RP42 A. Recommended practices for laboratory testing of surface active agents for well stimulation. Dallas: API; Jan; 1977.
- [28] Wu Y, Misra S. Intelligent image segmentation for organic-rich shales using random forest, wavelet transform, and hessian matrix. *Geosci Rem Sens Lett IEEE* 2019;17(7):1144–7.
- [29] Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed Signal Process Control* 2019;52:456–62.
- [30] Breiman L. Random forests. *Mach Learn* 2001;45(1):5–32.
- [31] Chen J, Li K, Tang Z, Bilal K, Yu S, Weng C, et al. A parallel random forest algorithm for big data in a spark cloud computing environment. *IEEE Trans Parallel Distr Syst* 2016;28(4):919–33.
- [32] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–94.
- [33] Zhang J, Sun Y, Shang L, Feng Q, Gong L, Wu K. A unified intelligent model for estimating the (gas + n-alkane) interfacial tension based on the eXtreme gradient boosting (XGBoost) trees. *Fuel* 2020;282:118783.
- [34] Dev VA, Eden MR. Gradient boosted decision trees for Lithology classification. In: Muñoz SG, Laird CD, Realf MJ, editors. Computer aided chemical engineering. Elsevier; 2019. p. 113–8.
- [35] Wasserman PD, Schwartz T. Neural networks. II. What are they and why is everybody so interested in them now? *IEEE expert* 1988;3(1):10–5.
- [36] Lashkarbolooki M, Hezave AZ, Ayatollahi S. Artificial neural network as an applicable tool to predict the binary heat capacity of mixtures containing ionic liquids. *Fluid Phase Equil* 2012;324:102–7.
- [37] Mohammadi M-R, Hemmati-Sarapardeh A, Schaffie M, Husein MM, Ranjbar M. Application of cascade forward neural network and group method of data handling to modeling crude oil pyrolysis during thermal enhanced oil recovery. *J Petrol Sci Eng* 2021;205:108836.
- [38] Khamehchi E, Mahdiani MR, Amooie MA, Hemmati-Sarapardeh A. Modeling viscosity of light and intermediate dead oil systems using advanced computational frameworks and artificial neural networks. *J Petrol Sci Eng* 2020;193:107388.
- [39] Sarapardeh AH, Larestani A, Menad NA, Hajrezaie S. Applications of artificial intelligence techniques in the petroleum industry. Gulf Professional Publishing; 2020.
- [40] Mohammadi M-R, Hadavimoghaddam F, Pourmahdi M, Atashrouz S, Munir MT, Hemmati-Sarapardeh A, et al. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Sci Rep* 2021;11(1):17911.
- [41] Kingma DP, Adam Ba J. A Method For Stochastic Optimization. In: International Conference on Learning Representations; 2015. p. 1–13.
- [42] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulina A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Process Syst* 2018:31.
- [43] Lv Q, Zheng R, Guo X, Larestani A, Hadavimoghaddam F, Riazi M, et al. Modelling minimum miscibility pressure of CO<sub>2</sub>-crude oil systems using deep learning, tree-based, and thermodynamic models: application to CO<sub>2</sub> sequestration and enhanced oil recovery. *Separ Purif Technol* 2023;310:123086.
- [44] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *Journal of Big Data* 2020;7(1):94.
- [45] Abdi J, Hadavimoghaddam F, Hadipoor M, Hemmati-Sarapardeh A. Modeling of CO<sub>2</sub> adsorption capacity by porous metal organic frameworks using advanced decision tree-based models. *Sci Rep* 2021;11(1):24468.
- [46] Ghanavati M, Shojaei M-J, S AAR. Effects of asphaltene content and temperature on viscosity of Iranian Heavy crude oil: experimental and modeling study. *Energy Fuels* 2013;27(12):7217–32.
- [47] Goodall CR. 13 Computation using the QR decomposition. 1993.
- [48] Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 2007;26(5):694–701.
- [49] Leroy A, Rousseeuw P. Robust regression and outlier detection. 1987. rrod.
- [50] Hemmati-Sarapardeh A, Hatami S, Taghvaei H, Naseri A, Band SS, Chau K-w. Designing a committee of machines for modeling viscosity of water-based nanofluids. *Engineering Applications of Computational Fluid Mechanics* 2021;15(1):1967–87.
- [51] Hosseinzadeh M, Hemmati-Sarapardeh A. Toward a predictive model for estimating viscosity of ternary mixtures containing ionic liquids. *J Mol Liq* 2014; 200:340–8.
- [52] Nakhaei-Kohani R, Taslimi-Renani E, Hadavimoghaddam F, Mohammadi M-R, Hemmati-Sarapardeh A. Modeling solubility of CO<sub>2</sub>-N<sub>2</sub> gas mixtures in aqueous electrolyte systems using artificial intelligence techniques and equations of state. *Sci Rep* 2022;12(1):3625.
- [53] Ganeeva YM, Yusupova TN, Barskaya EE, Valiullova AK, Okhotnikova ES, Morozov VI, et al. The composition of acid/oil interface in acid oil emulsions. *Petrol Sci* 2020;17(5):1345–55.