Petroleum Science 22 (2025) 894–908

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

Machine learning approaches for assessing stability in acid-crude oil emulsions: Application to mitigate formation damage

Sina Shakouri [a, b], Maysam Mohammadzadeh-Shirazi [a, b, *]

[a] Department of Petroleum Engineering, School of Chemical and Petroleum Engineering, Shiraz University, Shiraz, Iran
[b] Formation Damage and Well Treatment Research Group, IOR/EOR Research Institute, Shiraz University, Shiraz, Iran

ARTICLE INFO

Article history:
Received 28 March 2024
Received in revised form
13 July 2024
Accepted 18 September 2024
Available online 20 September 2024

Edited by Min Li

Keywords:
Acid-crude oil emulsion
Emulsion stability
Classification
Machine learning
Artificial neural network
Formation damage

ABSTRACT

The stability of acid-crude oil emulsion poses manifold issues in the oil industry. Experimentally evaluating this phenomenon may be costly and time-consuming. In contrast, machine learning models have proven effective in predicting and evaluating various phenomena. This research is the first of its kind to assess the stability of acid-crude oil emulsion, employing various classification machine learning models. For this purpose, a data set consisting of 249 experimental data points belonging to 11 different crude oil samples was collected. Three tree-based models, namely decision tree (DT), random forest (RF), and categorical boosting (CatBoost), as well as three artificial neural network models, namely radial basis function (RBF), multi-layer perceptron (MLP) and convolutional neural network (CNN), were developed based on the properties of crude oil, acid, and protective additive. The CatBoost model obtained the highest accuracy with 0.9687, followed closely by the CNN model with 0.9673. In addition, confusion matrix findings showed the superiority of the CatBoost model. Finally, by applying the SHapley Additive exPlanations (SHAP) method to analyze the impact of input parameters, it was found that the crude oil viscosity has the most significant effect on the model's output with the mean absolute SHAP value of 0.88.

1. Introduction

As the global population keeps growing, there will be an ever-increasing need for all forms of energy, such as fossil fuels. According to predictions published by Exxon Mobil, the global energy demand will increase by 25% between 2018 and 2040. Consequently, producing as much crude oil as possible from the existing reservoirs is crucial (Wojnar, 2018).

In the petroleum production industry, acidizing is known as a typical well stimulation technique due to its convenience, affordability, and favorable performance. The injected acid can dissolve plugged channels around the wellbore region and improves productivity, leading to an increase in crude oil production. In this process, the incompatibility between the acid solution (aqueous phase) and crude oil (organic phases) can cause stable acid-crude oil emulsion and acts as a serious formation damage. Formation and stability of acid-in-oil emulsion is one of the main causes of acidizing failure in oil wells and has negative consequences on the treatment efficacy (Mohammadzadeh Shirazi et al., 2019). Fig. 1 illustrates the happening of this phenomenon during the acidizing process.

During the acid stimulation process, the shear force created by the acid injection would act as an external factor for this emulsification. Some crude oil components such as asphaltene and resin, known as natural surfactants, tend to accumulate at the emulsion phase interfaces and inhibit the dispersed acid droplets to cohere by decreasing the interfacial energy (Alves et al., 2022). As a result, the interfacial area is not reduced, and a rigid film emulsion is created (Abbasi and Malayeri, 2022).

In light of the adverse effects that might result from the formation of acid-crude oil emulsion (Fredd and Fogler, 1998; Greene et al., 1974), evaluating and forecasting this before it happens is crucial. Determining the stability tendency of this undesirable emulsion helps prevent this by choosing a suitable acid type and concentration and using appropriate additives (Abbasi et al., 2024). For this purpose, researchers have employed experimental approaches, including centrifugation and the bottle test, to evaluate emulsion stability (da Silva et al., 2018; Hutin et al., 2016; Umar

* Corresponding author.
E-mail addresses: mmohshirazi@gmail.com, m.mohammadzadeh@shirazu.ac.ir (M. Mohammadzadeh-Shirazi).

https://doi.org/10.1016/j.petsci.2024.09.013

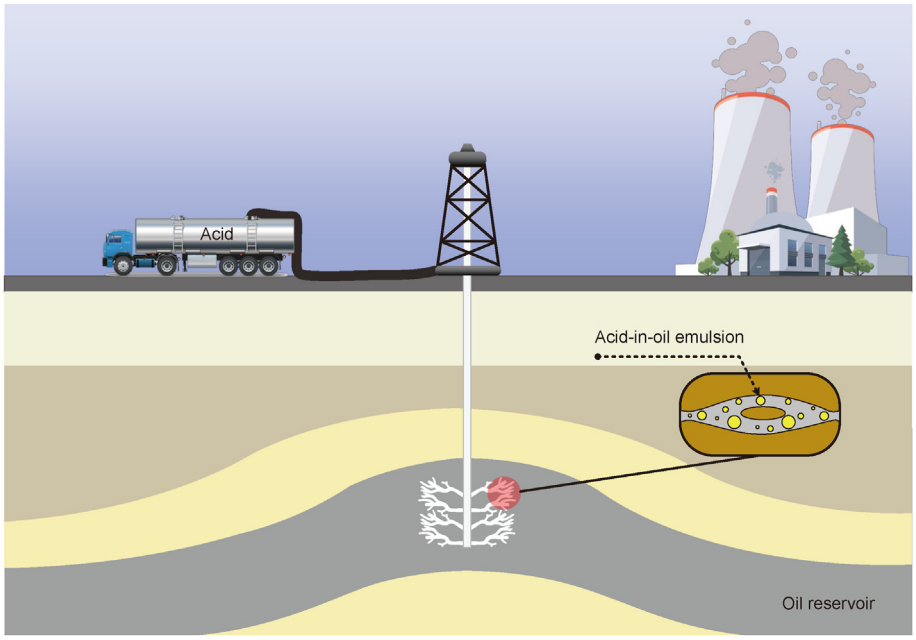


Fig. 1. An illustration of acid-in-oil emulsion formation during the acidizing process.

et al., 2018; Zhang et al., 2016). Additionally, the influence of several parameters, including crude oil properties like viscosity (Minakov et al., 2022) and asphaltene and resin content (Alves et al., 2022), iron concentration especially ferric (Pourakaberian et al., 2021), acid mixture ratio (AMR) (Mohammadzadeh Shirazi et al., 2019), and acid type and strength (Kharisov et al., 2012; Rietjens and Nieuwpoort, 2001), have been investigated on this emulsion stability. Employing protective additives like anti-emulsion, anti-sludge, and ferric ion reducer is the main strategy for avoiding this formation damage (Mohammadzadeh Shirazi et al., 2019).

Although experimental methods are common for evaluation, they are not always available due to financial or technical limitations and can be time-consuming, so developing advanced models such as machine learning models can be helpful and practical. In the last several decades, these novel techniques have been applied in various scientific fields as trustworthy replacements for expensive experimental investigations due to developments in artificial intelligence techniques and artificial intelligence (AI) has initiated the third wave of application (Liu et al., 2022). Regarding this matter, Deng et al. (2024) developed a hybrid machine learning optimization algorithm to estimate pore pressure, with the aim of enhancing drilling safety and efficiency; Pang et al. (2023) employed machine learning methods for predicting mud loss in carbonate formation drilling by utilizing seismic features and mud loss data; Bai et al. (2024) proposed an approach which utilizes data mining methods to find similar oil fields and estimate well production with machine learning models; Pei et al. (2024) used a fully connected neural network to accurately forecast the rate of penetration (ROP) with the goal of optimizing drilling operations; Shi et al. (2023) implemented a mixed-kernel machine learning approach to identify reservoir types in deep carbonates by using geophysical logging data; Zhang et al. (2023) proposed a hybrid neural network model based on the convolutional neural network (CNN) and gated recurrent unit (GRU) to predict bottom hole pressure (BHP) fluctuations.

A significant number of intelligent methods have been developed to predict reservoir formation damage. Zuluaga et al. (2002) have implemented an artificial neural network (ANN) and fuzzy logic to assess the impact of invasion by foreign particles on permeability decline in poorly consolidated rock. In order to predict permeability decline, the ANN performed best among applied models by using flow rate, initial porosity, initial permeability, and particle concentration. Rezaian et al. (2010) have evaluated the formation damage due to the deposition of asphaltene using ANNs. The ANN model estimated permeability decline utilizing asphaltene concentration, initial permeability, injection duration, and velocity with an average absolute percent relative error (AAPRE) of 0.83. In another study, Foroutan and Moghadasi (2013) have used an artificial neural network model to forecast relative permeability during precipitation of minerals to forecast relative permeability during mineral precipitation, achieving an average error of around 5%. Kamari et al. (2014) have implemented a least squares support vector machine model to predict the deposition of barium sulfate over a range of temperatures and concentrations of NaCl. The average absolute relative deviation of the proposed model was 0.0002%. In order to determine the volume and mass of formed acid sludge, Pourakaberian et al. (2021) have implemented an ANN using compatibility tests. The generated model had a correlation coefficient of 0.9458 for all data. In recent research on formation damage, Shakouri and Mohammadzadeh-Shirazi (2023) have developed four different machine learning models with the purpose of predicting asphaltic sludge formation. In this context, it was demonstrated that the MLP model with a correlation coefficient of 0.9517 had superior performance than other models in predicting asphaltic sludge formation.

Some attempts have been made to assess emulsions' stability in several fields using machine learning-based methods. de Souza et al. (2007) used an artificial neural network model to predict water-oil emulsion stability. The proposed model was able to predict the emulsion breaking height with a coefficient of determination of 0.899. Yetilmezsoy et al. (2011) developed a machine learning model called adaptive neuro-fuzzy inference system (ANFIS) to predict the stability of water-oil emulsion using oil properties with a coefficient of determination of 0.967. Kumar et al. (2011) implemented a model based on an ANN method to predict the stability of oil-water emulsion and identify the critical

concentrations of fatty alcohol with a coefficient of determination of 0.8920. Lee et al. (2022) used machine learning-based models to predict bilgewater emulsion stability. Several classification and regression models were employed, among which the random forest (RF) model with a F1 Score of 0.8244 and the decision tree (DT) model with mean absolute error (MAE) of 0.1611 performed the best, respectively. The proposed models indicate the ability of machine learning-based models in reliably predicting the stability of emulsions.

To the best of the authors' knowledge, no prior machine learning model has been developed for the prediction of acid-oil emulsion stability. A thorough literature review in the field of machine learning-based estimation of formation damage highlights a requirement to construct and propose a model for determining the stability class of acid-oil emulsion to prevent formation damage in the acidizing process. The present study aims to develop machine learning-based models to assess the stability of acid-crude oil emulsion. In order to accomplish this aim, 249 acid-crude oil emulsion stability data points are collected from bottle tests under the same experimental procedure at various crude oil properties, acid properties, and the amount of protective additives. The dataset is utilized to implement three tree-based machine learning models, namely decision tree (DT), random forest (RF), and categorical boosting (CatBoost), as well as three artificial neural network models, namely radial basis function (RBF), multi-layer perceptron (MLP) neural network, and convolutional neural network (CNN). Following the completion of model development, the models undergo validation using various statistical and graphical techniques. Ultimately, the SHapley Additive Explanations (SHAP) approach is applied to analyze the impact of input factors.

2. Acid-oil emulsion and data acquisition

Understanding the basic mechanisms of the phenomena is beneficial for selecting model inputs, which leads to more accurate and effective modeling. To achieve this objective, the mechanisms governing the formation of acid-crude oil emulsion are first described, followed by a description of the data set used.

2.1. Mechanism of acid-oil emulsion formation

The formation of stable acid-crude oil emulsion is undesirable because of its technical and financial consequences in the acid treatment process. Acid and crude oil, as two immiscible phases, form emulsions when enough shear force is exerted through mixing (Abbasi et al., 2023). The stability of this emulsion is from natural surfactants present in the crude oil called asphaltene and resin, which move towards the acid-oil emulsion interface and accumulate on it. Initially, this adsorption reduces the interfacial tension. With continuous accumulation of surfactants at the interface, a monolayer forms. This layer acts as a protective film, preventing direct contact and coalescence of droplets. Upon formation of this protective layer at the interface, further adsorption at this stage does not alter surface tension, and the emulsion is stabilized. (Abbasi et al., 2024). Also, ferric ion promotes this phenomenon due to its activity as a phase transfer catalyst for acid-base reactions (Kalhori et al., 2022). Fig. 2 illustrates the exact mechanism that causes this phenomenon to occur. Preventive additives including anti-sludge, anti-emulsion, and iron reducer hinder the stability of acid-crude oil emulsion (Abbasi et al., 2024; Mirvakili et al., 2012). According to the mechanism, this phenomenon could be affected by the viscosity of crude oil, saturate to aromatic ratio, asphaltene to resin ratio, acid concentration, acid to mixture ratio, amount of ferric ion, anti-sludge agent, anti-emulsion agent, and ferric reducing agent (Abbasi and Malayeri, 2022). The stability of an acid-crude oil emulsion can be estimated prior to the acid treatment process by using advanced models such as machine learning models and these effective input parameters.

2.2. Data acquisition

In this study, we conducted experiments to collect reliable data for the models. The compatibility bottle tests were done using some modifications of the standard technique established by API Recommended Practice 42 (American Petroleum Institute. Production Department, 1977), which is accurate for quantitatively evaluating acid-oil emulsion stability (Mohammadzadeh Shirazi et al., 2019). These experimental data have been published in the literature (Abbasi et al., 2023; Mohammadzadeh Shirazi et al., 2019; Pourakaberian et al., 2021). Hydrochloric acid was chosen as the desired acid, and the percentage of the acid that separated from the mixture was recorded. Fig. 3 illustrates the process of preparing emulsions and the method of classifying them. We have chosen time of 120 min for the stability measurement, as this is the most suitable time to determine whether the emulsion remains stable or not. Moreover, significant changes do not occur beyond 120 min is due to the formation of a monolayer, which stabilizes the emulsions relatively. In the early minutes, the stability of the emulsion can be uncertain because the system is still changing and the protective layer around the acid droplets may not yet be fully formed.

In order to confirm the accuracy and applicability of the model, 11 different samples of crude oil containing a broad range of SARA fractions were employed. The chemical and physical characteristics of the crude oil samples are detailed in Table 1. It is worth mentioning that the crude oil samples used in our study were selected based on their rheological properties, and they exhibited Newtonian behavior.

Basically, the acid-oil compatibility test is not highly repeatable, and the results of studies have shown some variation in the repetition of this test (Abbasi et al., 2023; Mohammadzadeh Shirazi et al., 2019); hence, it is more appropriate to provide a phase separation range to make up for this limitation. Moreover, in most cases, the emulsion is either completely stable or unstable, and in a few cases, it is placed outside of these two classes (Pourakaberian et al., 2021). Therefore, the most appropriate way to model this phenomenon is to classify it. For this purpose, the data was classified into four classes. As seen in Fig. 4, the data classification method is shown in Fig. 4(a), and the percentage of data in each class is shown in Fig. 4(b). Also, to expand the capabilities of the developed models, some experiments were done using the preventative additives and applied in the models. Table 2 describes the characteristics of the used additives.

The collected dataset contains 249 experimental data points. Each row of data includes values for crude oil viscosity, saturate to aromatic ratio, asphaltene to resin ratio, acid concentration, acid to mixture ratio, mass concentration of ferric ion, anti-sludge agent, anti-emulsion agent, and ferric ion reducing agent. Table 3 shows a statistical evaluation of the input data. In the present study, the synthetic minority oversampling (SMOTE) strategy was used to deal with the issue of imbalanced data. The SMOTE is an efficient resampling approach that has proven useful in a wide range of applications (Zhang et al., 2020). The SMOTE technique extends the original dataset by creating synthetic data points based on feature space (Chawla et al., 2002).

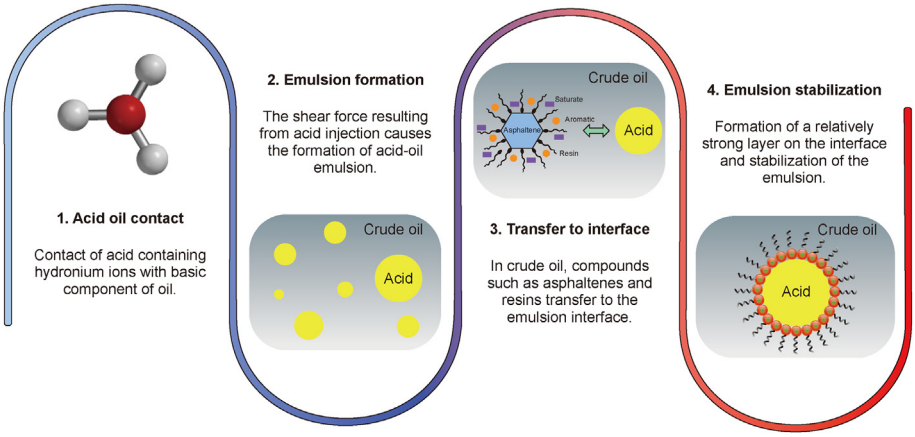


Fig. 2. The sequence of events leading to the formation of a stable acid-crude oil emulsion.

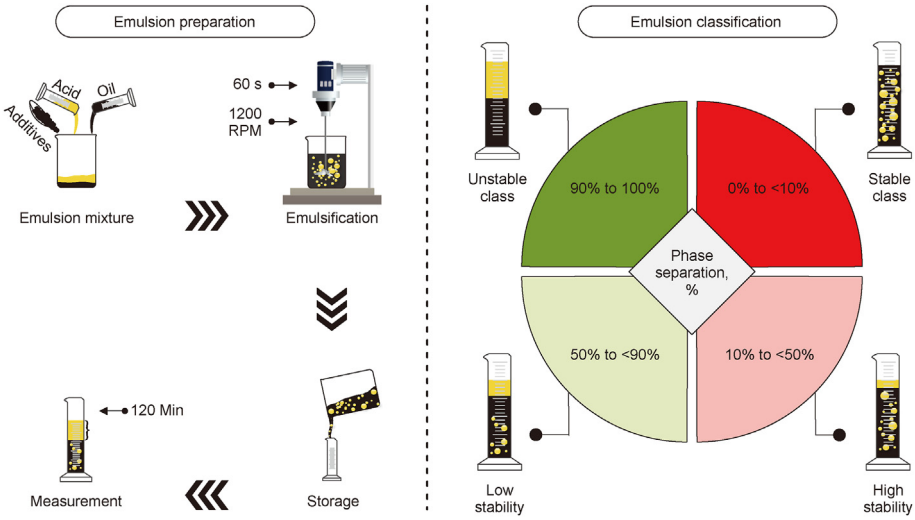


Fig. 3. The process of preparing emulsions and the classification of them based on phase separation.

Table 1
Characteristics of the crude oil samples in the dataset.

Sample	SARA Analysis				Sa/Ar	As/Re	Specific gravity (@ 25 °C)	Viscosity, cp (@ 25 °C)	Density (°API)
	Sa	Ar	Re	As					
A	45.04	40.49	6.23	8.24	1.1123	1.3226	0.9164	101.5	22.91
B	63.96	25.33	9.06	1.65	2.5250	0.1821	0.8731	25.20	30.57
C	49.93	39.12	7.95	3.00	1.2763	0.3773	0.8952	33.60	26.57
D	66.97	27.18	5.65	0.20	2.4639	0.0353	0.8711	18.20	30.94
E	47.25	39.93	7.61	5.21	1.1833	0.6846	0.9194	78.20	22.40
F	45.09	31.50	7.70	14.9	1.4314	1.9350	0.9330	2960	19.00
G	48.75	38.32	6.95	5.98	1.2721	0.8604	0.8903	80.30	27.40
H	43.4	35.6	12.9	8.1	1.2191	0.6279	0.9432	466.9	18.52
I	54.7	25.9	14.7	4.7	2.1119	0.3197	0.9178	131.2	22.67
J	52.8	31.7	9.7	5.8	1.6656	0.5979	0.9459	96.54	18.09
K	47.6	31.7	17.0	3.7	1.5015	0.2176	0.9159	121.8	22.99

3. Methodology

In the field of machine learning, tree-based models are a popular and widely used approach owing to the many benefits they provide, including their simplicity and interpretability. In contrast, although artificial neural network-based models are challenging to comprehend and interpret, they have a remarkable ability to learn and model nonlinear and complex relationships. In this regard, three tree-based models, including decision tree (DT), random forest (RF), and categorical boosting (CatBoost), as well as three artificial neural network models, including radial basis function (RBF), multi-layer perceptron (MLP) neural network, and convolutional neural network (CNN), were applied. In this part, the models are presented in detail, and the process of developing and

(a)

Phase separation, %	Class
0% to <10%	Stable
10% to <50%	High stability
50% to <90%	Low stability
90% to 100%	Unstable

(b)

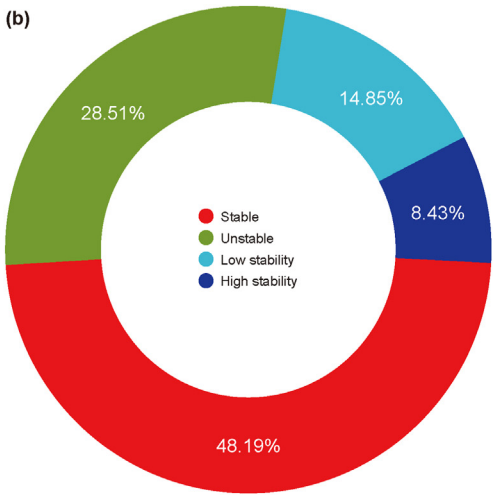


Fig. 4. Data classification method **(a)** and percentage of data in each class **(b)**.

Table 2
The acid-oil emulsion protective additives used in the dataset with their chemical structure.

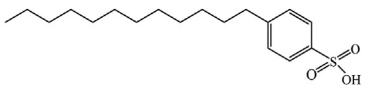
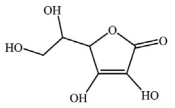
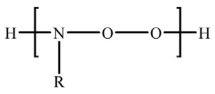
Name	Schematic molecular structure
Anti-sludge (Dodecyl benzene sulfonic acid ($C_{18}H_{30}O_3S$))	
Ferric reducing agent (Erythorbic acid ($C_6H_8O_6$))	
Anti-emulsion (N-alkylated polyhydroxyetheramines (NRO_2H_2))	

Table 3
The statistical evaluation of the data set employed in the present research.

Parameter	Maximum	Minimum	Standard deviation	Average
Asphaltene/Resin	1.935	0.0353	0.5529	0.6128
Saturate/Aromatic	2.525	1.112	0.5308	1.6704
Viscosity of crude oil, cP	2960	18.2	924.5207	430.6635
AMR, cc/cc	0.84	0.16	0.1487	0.5
Acid concentration, wt.%	32.5	10.5	4.9186	17.6626
Ferric ion, mg/L	3000	0	1066.248	1602.3453
Ferric reducing, wt.%	0.5	0	0.1502	0.0502
Anti-emulsion, wt.%	1	0	0.2424	0.1024
Anti-sludge, wt.%	1	0	0.2360	0.923

evaluating them is described. In order to estimate the stability of acid-oil emulsion, after collecting and pre-processing the data, statistical and graphical analysis is performed to assess the developed models. In addition, to determine the impact of the input parameters, the relevance factor is calculated to clarify the relationship between the input parameters and model output. The chain of activities performed in this study is shown in Fig. 5.

3.1. Machine learning

Machine learning is a scientific field that concentrates on creating models with the capacity to learn patterns based on previous data without being specifically programmed (Samuel, 1959). In conventional programming, the user provides the computer with data and rules; the computer then uses these to determine an intended result. In contrast, in the field of machine learning, the model receives data and answers, and the rules are the outcome. Thus, this model predicts the answer for new data using the learned rules.

A machine learning model can learn in different ways depending on the type of data set; as seen in Fig. 6, supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the main methods of machine learning models. The most common method for machine learning is supervised learning, which involves providing the model with data in the form of labeled samples. The model learns the pattern of the relations

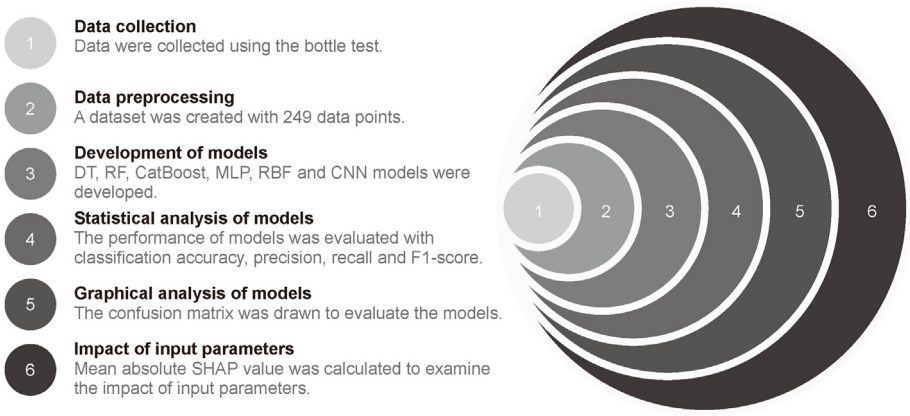


Fig. 5. The chain of activities performed to develop and evaluate the models in this study.

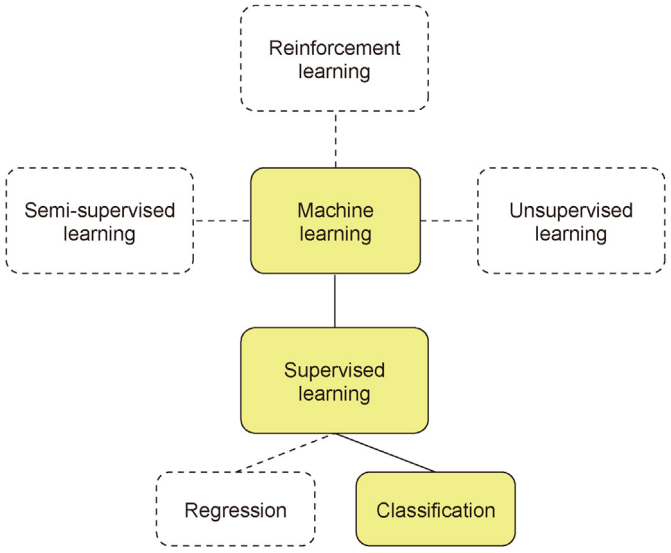


Fig. 6. Types of learning methods for a machine learning model. As highlighted, the supervised classification method was used in this study.

between input data and sample labels during training. Two types of supervised learning are classification and regression (Qiao et al., 2020). In predictive classification modeling, after using the data to train the model, the new data can be classified (Alpaydin, 2020). In this study, supervised classification models were used, and the developed models are described as follows.

3.2. Decision tree (DT)

The DT model is a supervised tree-based algorithm that is used for both regression and classification purposes (Geurts et al., 2009). The DT method has many notable benefits, including minimal data preparation requirements and optimal performance on huge data sets (Yang and Fong, 2013). There are four parts to this algorithm: First, the root node, which is the starting point of the tree and contains the input data. In the end, the leaf nodes form the flowchart's endpoint and indicate the system's output. In addition, the internal nodes are located between root and leaf nodes, and the branches that connect the nodes. A schematic illustration of the DT model's framework is shown in Fig. 7.

In the DT method, splitting, stopping, and pruning are the three main techniques for creating a tree (Song and Lu, 2015). Beginning with the root node, the input data is split into decision nodes and branches. The splitting operation continues until a stopping requirement is satisfied. In addition, the pruning procedure involves eliminating the low-value branches (Patel and Upadhyay, 2012).

3.3. Random forest (RF)

As has been explained, despite the many advantages of DT, it also has some disadvantages. One main disadvantage is the possibility of overfitting. DTs frequently exhibit high variance and low prediction bias, a condition known as over-fitting, which allows the model to detect small perturbations and random noise in the training sample. In addition, the overall DT may not be optimal because this model ignores the global optimum (Liu et al., 2023b). The problems mentioned above can be solved using ensemble techniques, combining the outcomes from various trees into one conclusion (Brieuc et al., 2018). RF is a DT ensemble learning system in which every tree is trained simultaneously (Yan et al., 2023). In RF, the greedy method defines the significance of each tree at every stage (Wu and Misra, 2019). Additionally, RF can assess the significance of a feature and keep the most useful input information of a feature (Shaikhina et al., 2019). Fig. 8 depicts the basic structure of the RF model.

The RF approach uses a procedure called bagging that enhances the variety of the trees and variable selection. The algorithm will decide how to split the data it receives to several datasets according to the number of the provided trees.

3.4. Categorical boosting (CatBoost)

The categorical gradient boosting method known as CatBoost uses binary decision trees as its main predictors (Prokhorenkova et al., 2018). This approach functions with minimal loss of information for categorical features. The difference between training and testing datasets is the most important concept for the CatBoost approach (Hancock and Khoshgoftaar, 2020). Furthermore, the indicator function 1 is a key concept to comprehend how CatBoost categorical features are encoded, which is defined as follows (Hancock and Khoshgoftaar, 2020):

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

The function mentioned above is a critical component of the formula employed by CatBoost for converting the category

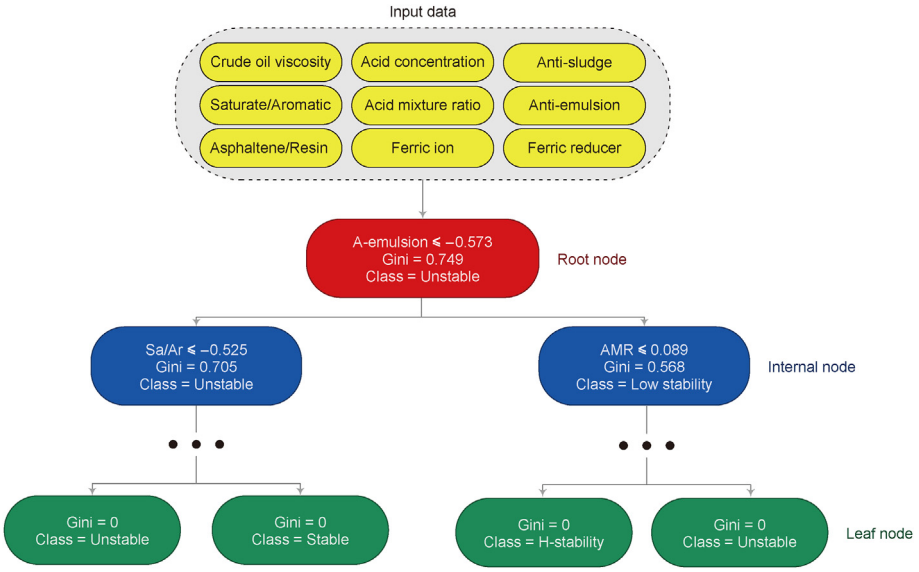


Fig. 7. Schematic illustration of a DT model.

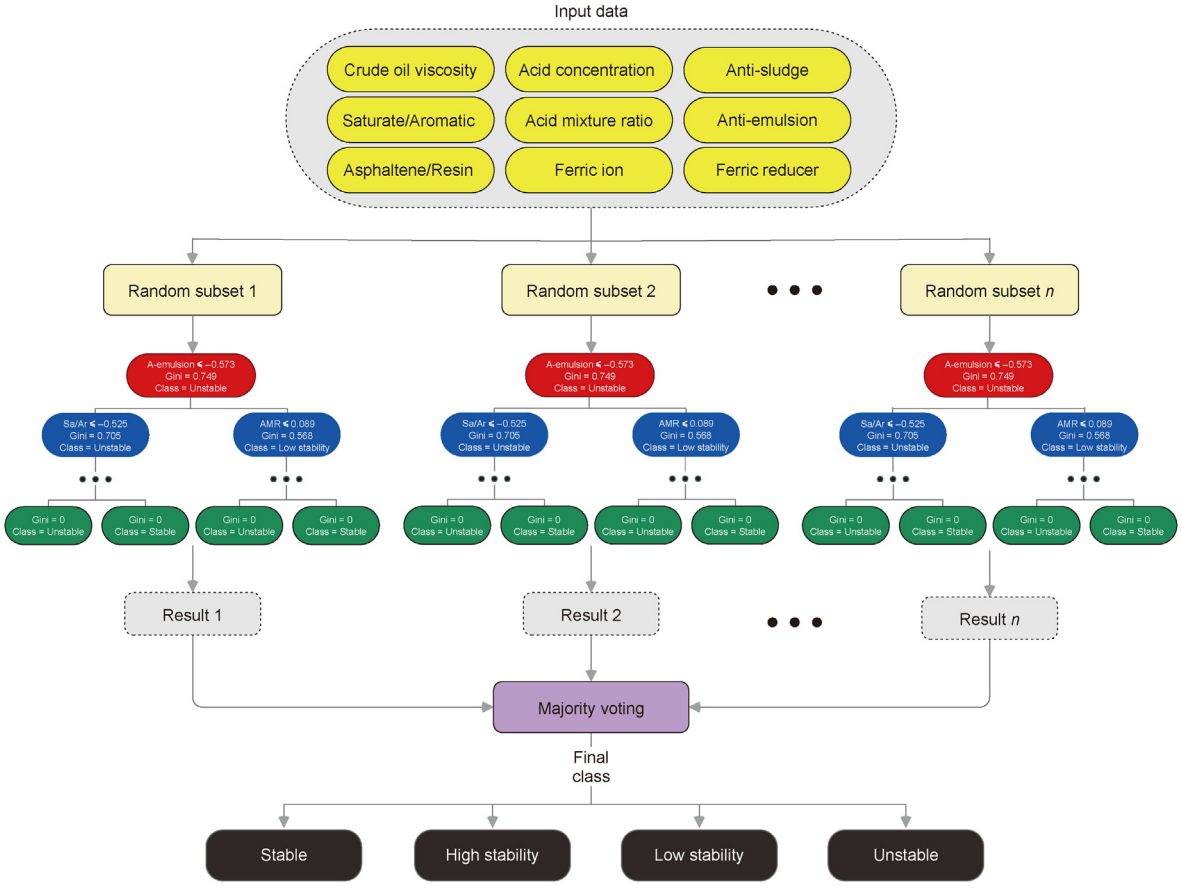


Fig. 8. The basic structure of the RF model.

variables into number values. Also, the CatBoost algorithm uses categorized columns to boost performance. The two most important are One-Hot-Max-Size and target-based statistics, which are applied in this work. The basic phases in the CatBoost method include forming a random set of variables, converting labels to numbers, and transforming features into numerical values (Abdi et al., 2021). To prevent overfitting, the CatBoost methode employs random permutations for predicting leaf values during selecting the tree structure, which gives it an important advantage. The basic framework of the CatBoost model is shown in Fig. 9.

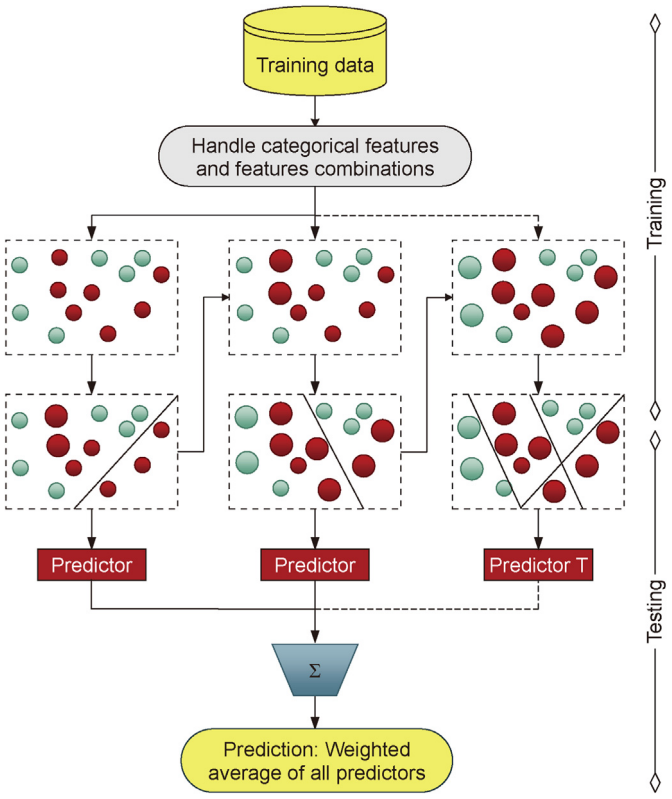


Fig. 9. The basic framework of the CatBoost model.

3.5. Radial basis function (RBF) neural network

The radial basis function neural network is a popular model of neural networks used in classification and regression issues. In fact, the concept of RBF neural networks originates from the mathematical theory of function approximation (Liu et al., 2023a). By transforming the data into a space with more dimensions, RBF can process and provide accurate solutions for every issue involving scattered and multivariate data (Broomhead and Lowe, 1988). RBF neural networks typically have a three-layer architecture with just one hidden layer between the input (first) and output (third) layers (Hemmati-Sarapardeh et al., 2018). Fig. 10 depicts a graphical illustration of the RBF structure implemented in this study. In RBF, radial basis functions $\phi_i(x)$ are used for computing regression on a function $f(x)$. This regression is obtained using linear superposition of basis functions. $f(x)$ can be calculated as follows (Broomhead and Lowe, 1988):

$$f(x_i) = w^{\mathrm{T}}\phi(x_i) + b \tag{2}$$

where w^{T} is the output layer weight, $\phi(x_i)$ denotes the transfer function, and b defines the bias. During the RBF model development process, various radial basis transfer functions can be implemented; however, the Gaussian function is the most commonly employed radial basis function in the RBF neural network (Xia et al., 2023). In this study, the Gaussian function was used, which is defined as follows (Hemmati-Sarapardeh et al., 2019):

$$\phi(r) = \exp\left(\frac{r^2}{2\sigma^2}\right) \quad \text{with} \quad \sigma > 0 \tag{3}$$

where σ denotes the spread coefficient.

3.6. Multi-layer perceptron (MLP) neural network

One of the most well-known kinds of artificial neural networks is the multi-layer perceptron (MLP), which is a sort of feedforward artificial neural network with multiple layers (Wasserman and Schwartz, 1988). The initial layer is the input layer, and it is responsible for receiving input data, the output layer is the last layer of the model which represents the model's output, and the hidden layers are the layers in between, which are used for data processing (Kiannejad Amiri et al., 2023; Lashkarbolooki et al., 2012). To determine each neuron's value in the hidden layers or output, the value of each neuron in the previous layer is multiplied by its related specific weight and added together, and a bias is added to these values (Gao et al., 2022). Finally, an activation function is applied to the obtained value (Mohammadi et al., 2021).

Each input layer in a MLP model has the same number of neurons as the number of input parameters, and the output layer has the same number of neurons as the existing classes. Hence, in this study, due to the presence of 4 different classes, the last layer includes 4 neurons. The number of hidden layers and their neurons must be tuned in order to create an accurate and robust MLP model (Khamehchi et al., 2020). The efficacy of neural network models is significantly affected by optimization strategies (Hagan and Menhaj, 1994). In order to achieve this goal, an optimization technique called Adam's, which is one of the most well-known optimization algorithms of artificial neural networks (Kingma and Ba, 2014), was used in the present study. The schematic of the MLP model constructed for the present research is shown in Fig. 11.

3.7. Convolutional neural network (CNN)

In recent years, convolutional neural networks (CNNs) have attracted a lot of attention due to their impressive performance in fields including classification, image processing, and pattern recognition (Albawi et al., 2017). Compared to other artificial neural networks, CNNs reduce parameters, which is a significant advantage (Lv et al., 2023). The schematic of the CNN algorithm is shown in Fig. 12. Two of the most significant concepts behind the CNN network are the convolutional phase and the fully-connected phase. In the convolution phase, features are calculated in multi-dimensional sequences, and values are computed by convolution of conveyed data obtained from the previous step and filter window (kernel). The conveyed data is processed from left to right and top to bottom by the filter screen. A set of features selected from the dataset forms a feature map (Xue et al., 2024).

3.8. Procedure of model development

The first step for developing machine learning models is to select training and testing subsets. The data in this research were randomly divided to avoid bias and intention in the process of selecting subsets. In order to establish training and testing subsets, 80% of the data was assigned for training, and 20% was assigned for testing. Then, the hyperparameter values were chosen for model development. Model performance is significantly affected by the selection of hyperparameter values. The main aim of setting hyperparameters is to identify suitable hyperparameter values to obtain the highest performance of any machine learning model (Castro-Amoedo et al., 2024). Employing the Randomized Search technique is one of the most common ways to achieve this aim (Pedregosa et al., 2011), which was employed in this research. In brief, in the first stage, the specific hyperparameters associated with the desired algorithm are identified, and then the range of search values is determined. In the last stage, the final model is constructed based on the optimal set of hyperparameters. In

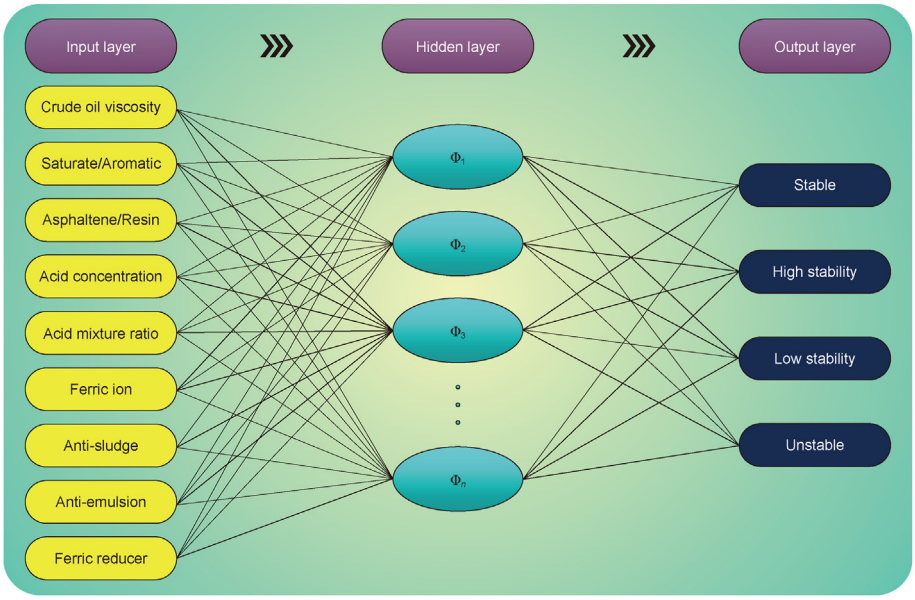


Fig. 10. A schematic illustration of the developed RBF model.

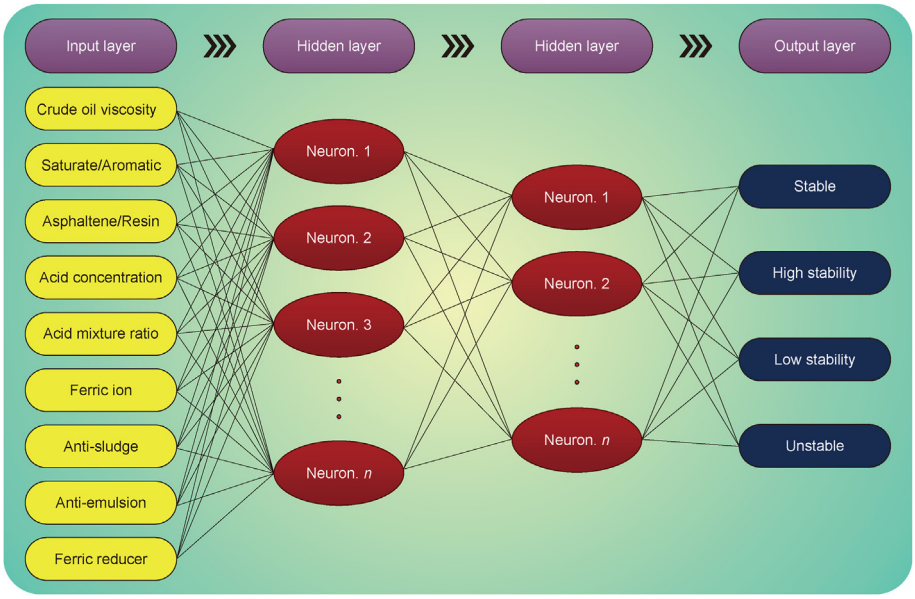


Fig. 11. A schematic illustration of the developed MLP model.

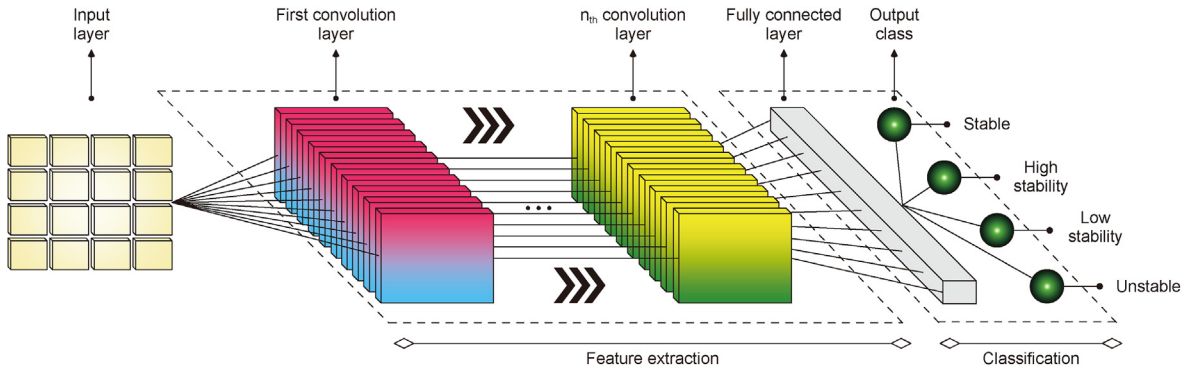


Fig. 12. A schematic illustration of the CNN algorithm.

addition, the cross-validation strategy was applied, where data sampling methods such as cross-validation are often used for unbiased evaluation of models during the model development process (Berrar, 2019).

3.9. Performance evaluation of models

In this study, to confirm the validity of each model, the models were evaluated using several criteria. The evaluation criteria are divided into two groups: statistical and visual evaluation, which are described below.

3.9.1. Statistical evaluation

In order to evaluate the developed models, *Accuracy*, *Precision*, *Recall*, and *F*1 *Score* metrics were used, which are defined as follows (Garud et al., 2018; Liu et al., 2023c):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Where *TP* stands for true positive, *FP* for false positive, *FN* for false negative, and *TN* for true negative, the definition of *Accuracy* is the ratio of samples that were predicted correctly to the whole number of samples.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Precision is measured by the ratio of correctly predicted positive samples to all predicted positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Recall of a classification system is measured by the ratio of the number of correctly predicted positive samples to the total number of samples relating to that class.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

The weighted average of *Precision* and *Recall* gives *F*1 *Score*. All of the aforementioned criteria range from 0 to 1, with the closer to 1, the better the model's performance.

3.9.2. Visual evaluation

The confusion matrix is a table that describes the performance of a classification machine learning model. Confusion matrices are useful for visualizing the results of a classification model (Dobos et al., 2023; Singh et al., 2021). With n output classes in the model, the confusion matrix will be a $n \times n$ matrix. The diagonal elements in the confusion matrix indicate how many samples of each class are correctly classified in the test data. On the other hand, non-diagonal elements indicate incorrect predictions. Therefore, the most accurate classifier will have a confusion matrix consisting completely of diagonal elements and zero for non-diagonal (Arjaria et al., 2021). In addition, the superior model was interpreted using the SHAP method. SHAP is a method based on coalitional game theory that is used for explaining the machine learning models (Lundberg and Lee, 2017). This method was applied to evaluate the impact of input parameters.

4. Results and discussion

The main focus of this research is the analysis of the accuracy and validity of the developed models, as well as comparing the performance of tree-based and artificial neural network-based models. This section is divided into three parts: 1- results for selecting the most optimal hyperparameters of each proposed model, 2- evaluation and comparison of the models using various graphical and statistical metrics, 3- analyzing the most accurate model based on SHApley Additive Explanations (SHAP).

4.1. Development of the models

The DT, RF, CatBoost, RBF, MLP, and CNN models were used to predict the stability of acid-oil emulsion. The grid search method was applied to select the most optimal values of hyperparameters (Dong et al., 2023). The type of hyperparameter, the search range, and the optimum value obtained are presented in Tables 4 and 5. Also, the 10-fold cross-validation technique was served in the development and evaluation of the models.

4.2. Evaluation of the models

Various criteria were used to analyze and evaluate the model's reliability, which are presented in two sections: statistical evaluation and visual evaluation.

4.2.1. Statistical evaluation

In order to properly evaluate the performance, four statistical criteria, including *Accuracy*, *Precision*, *Recall*, and *F*1 *Score* were used for training, testing, and total data sets. The results are reported in detail in Table 6. The analysis of statistical criteria results for tree-based and artificial neural network models is described as follows:

4.2.1.1. Tree-based models. According to Table 6, the most successful tree-based algorithm was CatBoost, which predicted the emulsion stability classes with an *Accuracy* of 0.9687, followed by RF and DT models with an *Accuracy* of 0.9583 and 0.9270. The other statistical criteria, including *Precision*, *Recall*, and *F*1 *Score* values, were 0.9691, 0.9687, and 0.9686 for the CatBoost model, respectively, which shows its superiority. In addition, the slight difference between *Accuracy*, *Precision*, *Recall*, and *F*1 *Score* values proves the trustworthiness of the CatBoost model. Also, the small gap between training and testing accuracy indicates that the models have not been over-fitted and are reliable.

4.2.1.2. Artificial neural network models. Table 6 shows that the CNN model provided the highest performance with an *Accuracy* of 0.9673 among the neural network models. The MLP and RBF models provided lower performance with the *Accuracy* of 0.9583 and 0.9479, respectively. The obtained values of 0.9688 for *Precision*, 0.9573 for *Recall*, and 0.9675 for *F*1 *Score* all confirm the superiority of the CNN model. In addition, the insignificance of the difference between the training and testing *Accuracy* of neural

Table 4
The optimal hyperparameter values for the tree-based models.

Hyperparameter	Search range	Model		
		DT	RF	CatBoost
max_depth	5–50	11	16	–
max_leaf_nodes	10–140	82	70	–
n_estimators	20–2500	–	560	–
max_features	'Sqrt', 'log 2', 1, 2, 3	–	1	–
iterations	20–2000	–	–	1000
learning_rate	0.01–1.5	–	–	1.258
depth	2–20	–	–	7
l2_leaf_reg	2–20	–	–	4

Table 5
The optimal hyperparameter values for the artificial neural network models.

Hyperparameter	Search range	Model		
		MLP	RBF	CNN
learning_rate	0.001–0.9	0.012	0.008	0.0011
loss function	Categorical cross-entropy	Categorical cross-entropy	Categorical cross-entropy	Categorical cross-entropy
Hidden layers activation function	[sigmoid-ReLU- Gaussian-LeakyReLU]	ReLU	Gaussian	LeakyReLU
Output layer activation function	Softmax	Softmax	Softmax	Softmax
1st Hidden layer size	1–120	20	95	16
2nd Hidden layer size	1–120	45	–	48
1st Convolutional layer filters	1–256	–	–	32
2nd Convolutional layer filters	1–256	–	–	64
3rd Convolutional layer filters	1–256	–	–	128
Epoch	100–1000	400	300	400
Batch size	4–64	8	8	8
Kernel size	1–6	–	–	3

Table 6
Calculated the statistical criteria for the implemented models.

Criteria	Data set	Tree-Based model			Artificial neural network model		
		DT	RF	CatBoost	MLP	RBF	CNN
Accuracy	Train	0.9479	0.9635	0.9713	0.9609	0.9635	0.9716
	Test	0.9270	0.9583	0.9687	0.9583	0.9479	0.9673
	Total	0.9437	0.9625	0.9708	0.9604	0.9604	0.9708
Precision	Train	0.9500	0.9641	0.9721	0.9618	0.9663	0.9730
	Test	0.9347	0.9587	0.9691	0.9609	0.9516	0.9688
	Total	0.9464	0.9627	0.9713	0.9615	0.9632	0.9716
Recall	Train	0.9479	0.9635	0.9713	0.9609	0.9635	0.9716
	Test	0.9270	0.9583	0.9687	0.9583	0.9479	0.9673
	Total	0.9437	0.9625	0.9708	0.9604	0.9604	0.9708
F1 Score	Train	0.9483	0.9635	0.9714	0.9610	0.9636	0.9717
	Test	0.9287	0.9576	0.9686	0.9582	0.9487	0.9675
	Total	0.9443	0.9624	0.9709	0.9605	0.9606	0.9709

network models indicates that the models were not overfitted and are reliable. Fig. 13 demonstrates the *Accuracy* of the models comparatively. As can be seen, the CatBoost model has the highest *Accuracy* with 0.9687, followed by the CNN model with an *Accuracy* of 0.9673 with a slight difference.

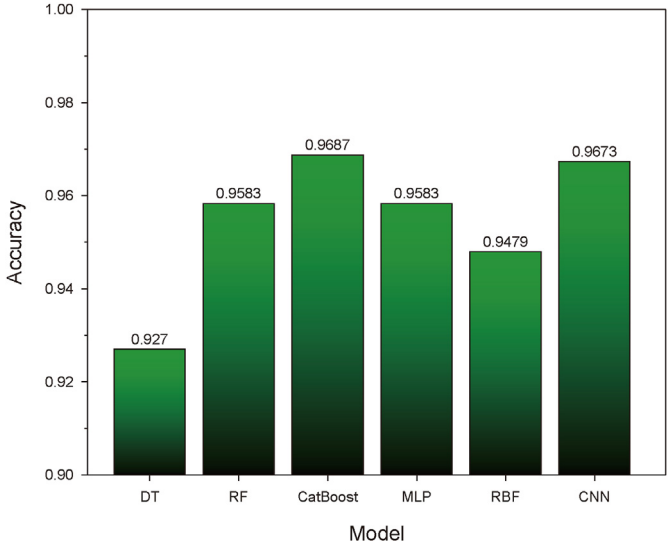


Fig. 13. Comparison of the developed models' *Accuracy*.

4.2.2. Visual evaluation

In order to evaluate the developed models more deeply, visual evaluation was performed using the confusion matrix. The confusion matrix is a two-dimensional plot that shows actual and predicted labels on two axes. This graph shows the percentage of correctly and incorrectly labeled samples for stable, high stability, low stability, and unstable classes. Fig. 14 shows the confusion matrices for all the tree-based and artificial neural network models. In the following, the results of confusion matrices are examined.

4.2.2.1. Tree-based models. Fig. 14 shows that the correct classification of tree-based models is higher than 92% for each of the four classes. Especially for stable and unstable classes, which are the most important ones, the correct classification of 95% and higher was achieved. In addition, the confusion matrix proves the superiority of the CatBoost model, which has the highest accuracy in three classes. It can also be seen that the DT model has classified 7.5% of the data in the unstable class instead of the high stability class, which shows the highest error in the classification of this class.

4.2.2.2. Artificial neural network models. According to Fig. 14, it can be seen that all artificial neural network models have a correct classification of 90% and higher for all four classes. These models have an accuracy of 91.66% and higher for stable and unstable classes. According to the confusion matrix results, it has been proved that the CNN model has the best performance among all the neural network models. In addition, the RBF model classified 10% of the data in the unstable class instead of the high stability class, which shows the highest error in the classification of this class.

Generally, the confusion matrices revealed specific misclassifications that highlight areas for improvement. For the Decision Tree (DT) model, a notable misclassification occurred where 7.5% of unstable samples were incorrectly categorized as having high stability. This can be attributed to the model's tendency to overfit to patterns in the training data, which may not generalize well to new data. Similarly, the Radial Basis Function (RBF) model showed a 10% error rate for unstable samples, incorrectly labeling them as high stability, indicating its struggle with capturing nonlinear relationships between features and labels. Misclassifications between high stability and unstable classes indicate a need for expanding the training dataset to enhance the models' ability to distinguish between different classes.

In order to visualize the performance of the developed models, a quadrilateral diagram was drawn based on the accuracy of each class. As shown in Fig. 15, CatBoost and CNN models cover almost the same area, and their better performance than other models is proven.

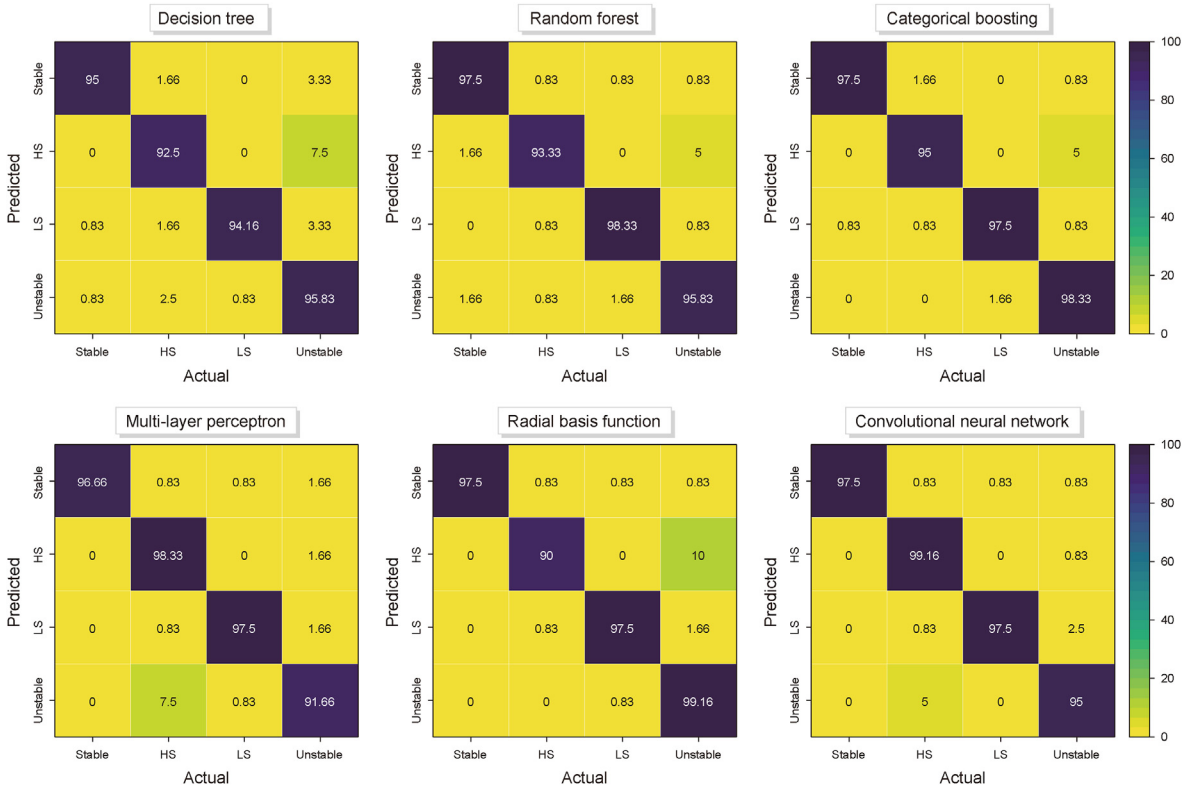


Fig. 14. Confusion matrices of the developed models.

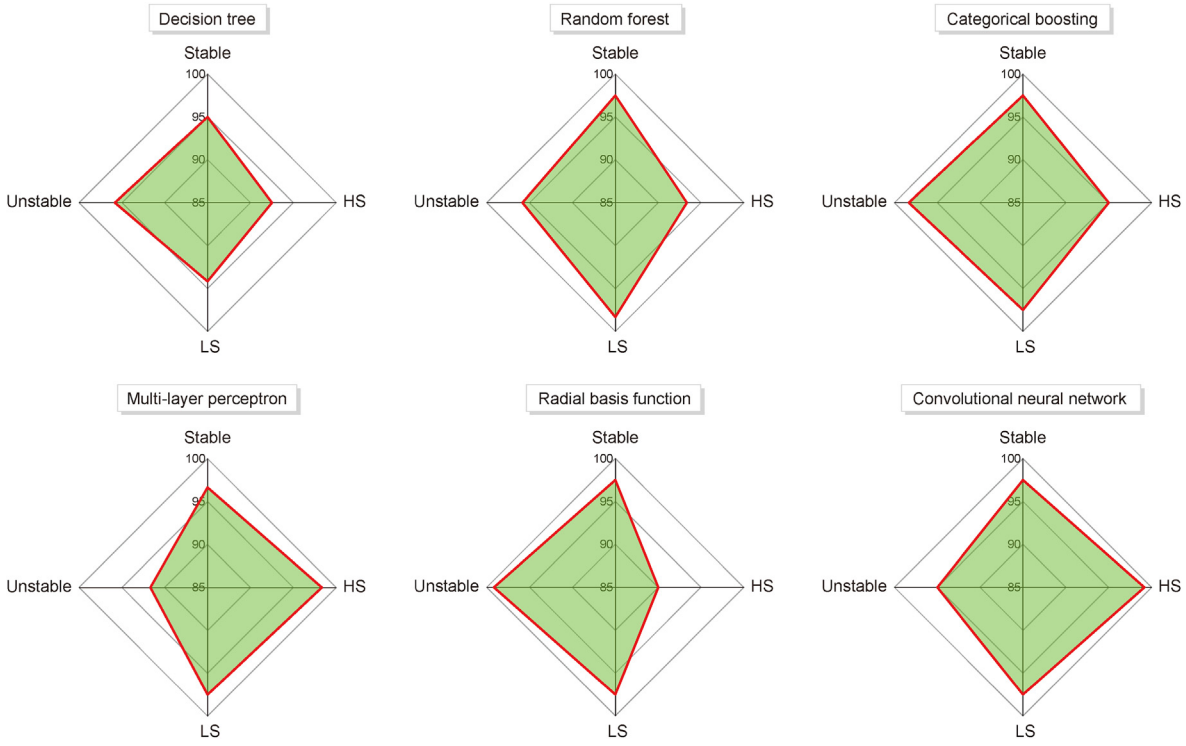


Fig. 15. Quadrilateral diagram of the developed models based on the accuracy of each class.

4.2.3. Computational efficiency: training time comparison

In this section, the training times of different models used in the study are compared. The system configuration for this evaluation is as follows: Intel(R) Core(TM) i5-8350U CPU@1.70GHz, 1.90 GHz, 16.0 GB RAM, and Python version 3.13. The total training times for each model are presented in Table 7.

Table 7
The total training time of the models developed in this research.

Model	Total training time, s
Decision Tree	0.1340
Random Forest	10.2199
Categorical Boosting	80.7828
Radial Basis Function	437.5897
Multi-Layer Perceptron	525.6515
Convolutional Neural Network	2227.7067

As shown in Table 7, there is a clear distinction between the total training time of tree-based models and Artificial neural network models. Tree-based models such as DT and RF exhibited significantly lower training times due to their simpler structures. CatBoost model, although more complex, still maintains a relatively low training time of 80.7828 s. Conversely, ANNs like MLP, RBF, and CNN require substantially longer training times due to their complex architectures. Despite the lengthy training time, CNN achieved an accuracy comparable to CatBoost. However, CNN's training time was approximately 27.6 times longer than CatBoost's. This highlights the increased computational requirements associated with more complex models like CNNs. In larger datasets, the resource demands of such models become even more pronounced, affecting their efficiency.

4.3. Impact of input parameters

It is beneficial to identify which dataset parameters play a significant role in the formation of acid-oil emulsion. According to the results of the statistical and graphical analysis, the CatBoost model was the superior model, and the prediction results of this model are used to analyze the input data. SHAP value provides a method to rank the importance of input parameters on the ML model's output (Meng et al., 2023). The mean absolute SHAP value of each input parameter is used to calculate the importance index, which represents the average influence of input parameters. In this method, a higher mean absolute SHAP value indicates greater importance (Yao et al., 2023). Fig. 16 represents the importance of the feature for the acid-oil emulsion probability class. It can be seen that crude oil viscosity, ferric ion, anti-emulsion additive, and acid concentration with the mean absolute SHAP values of 0.88, 0.83, 0.81, and 0.75 are the most important input parameters, respectively. Previous studies have revealed that crude oil viscosity, ferric ion concentration, and acid concentration play a crucial role in the formation and stabilization of acid-oil emulsion (Abbasi et al., 2023; Mohammadzadeh Shirazi et al., 2019; Taylor et al., 1999). On the other hand, AMR and saturate/aromatic parameters have relatively lower mean absolute SHAP values of 0.62 and 0.64, respectively, while they are still important due to their high values.

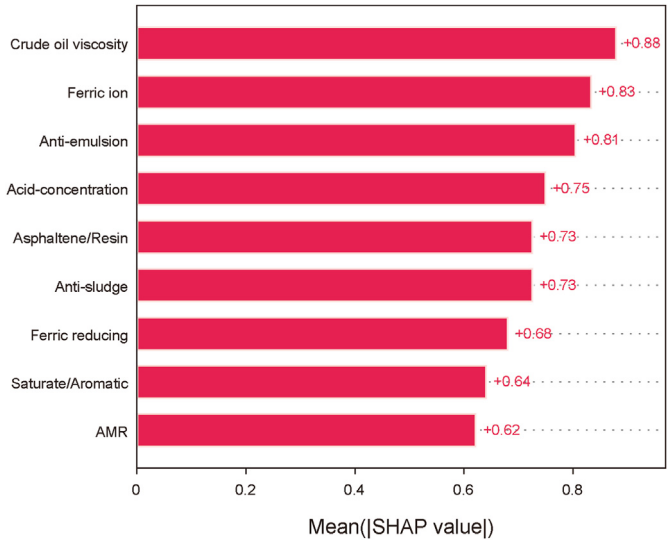


Fig. 16. Ranking the importance of input parameters based on the mean absolute SHAP value.

Finally, implementing machine learning-based models provides two significant advantages: One is giving a tool for rapid decision-making, which is especially helpful when there is a lack of time to undertake compatibility tests. Through machine learning models, the stability class of acid-oil emulsion can be accurately predicted; furthermore, the suitable amount of additives and the acid concentration can be found. Two, the optimal values of effective parameters to control the acid-oil emulsion will be determined without bias. In other words, based on previous experiences, the most suitable values are chosen to control the stability of the emulsion; consequently, errors made by experts or biased decisions cannot influence the results.

5. Limitations and future work

While this study offers findings and insights into the application of machine learning models for predicting acid-oil emulsion stability, it is essential to acknowledge certain limitations. One significant limitation is the size of our dataset. Although 249 data points were sufficient to develop and validate our models, expanding the dataset could enhance model accuracy and reduce errors as the models would learn more patterns. Future research should consider expanding the dataset to include a broader range of crude oil and acid properties.

Additionally, gathering data that measures the impact of temperature or other influential features could further refine our models and improve their precision. Future work could also explore alternative hyperparameter optimization methods beyond Grid Search, such as Genetic Algorithms, to efficiently navigate the hyperparameter space and potentially uncover more optimal configurations. This approach could lead to enhanced performance and robustness of the models.

6. Conclusion

In this study, the application of machine learning methods to assess the acid-oil emulsion stability was investigated for the first time. For this purpose, three tree-based algorithms and three artificial neural network algorithms were developed using a data set consisting of 249 experimental data points obtained through a particular experimental methodology. Several statistical metrics and visual evaluations were applied to evaluate the performance of the models. The following findings are obtained.

1. In the category of tree-based models, the CatBoost model predicted stability with an *Accuracy* of 0.9687, as the most accurate model, followed by the RF and DT models, with an *Accuracy* of 0.9583 and 0.9270, respectively.
2. In the category of neural network-based models, the CNN model predicted stability with an *Accuracy* of 0.9673 as the most accurate model, followed by the MLP and RBF models, with an *Accuracy* of 0.9583 and 0.9479, respectively.
3. Comparing the implemented models' results showed that the CatBoost model was the most accurate. Nevertheless, it is worth noting that the *Accuracy* of the CNN model was very close to this and can be advised. Generally, tree-based models are more

understandable, quicker, and have lower computational costs than neural network-based models.

4. According to the confusion matrix, both CNN and CatBoost models showed similar results; both models correctly classified the total data for all four classes by more than 95%. In addition, both models incorrectly classified only 0.83% of the total data into unstable class instead of stable.
5. The implementation of the SHAP method revealed that all the features influenced the stability of acid-crude oil emulsion. Additionally, all features were identified as significant, and none of them were found to be redundant. Nonetheless, it was discovered that crude oil viscosity, ferric ion, anti-emulsion additive, and acid concentration with the mean absolute SHAP value of 0.88, 0.83, 0.81, and 0.75, respectively, were the most important features.

The findings have proved that implementing machine learning-based models provides a quick instrument for operational decision-making and finding optimal values. In order to improve reliability, future research can be concentrated on expanding the dataset and evaluating the performance of various machine learning algorithms.

CRediT authorship contribution statement

Sina Shakouri: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maysam Mohammadzadeh-Shirazi:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Data curation, Conceptualization.

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.

References

Abbasi, A., Malayeri, M.R., 2022. Stability of acid in crude oil emulsion based on interaction energies during well stimulation using HCl acid. J. Pet. Sci. Eng. 212, 110317. https://doi.org/10.1016/j.petrol.2022.110317.

Abbasi, A., Malayeri, M.R., Mohammadzadeh Shirazi, M., 2023. Stability of spent HCl acid-crude oil emulsion. J. Mol. Liq. 383, 122116. https://doi.org/10.1016/j.molliq.2023.122116.

Abbasi, A., Mohammadzadeh-Shirazi, M., Malayeri, M.R., 2024. Functionality of chemical additives and experimental conditions during formation of acid-induced emulsion and sludge. J. Mol. Liq. 398, 124257. https://doi.org/10.1016/j.molliq.2024.124257.

Abdi, J., Hadavimoghaddam, F., Hadipoor, M., et al., 2021. Modeling of CO_2 adsorption capacity by porous metal organic frameworks using advanced decision tree-based models. Sci. Rep. 11 (1), 24468. https://doi.org/10.1038/s41598-021-04168-w.

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET).

Alpaydin, E., 2020. Introduction to Machine Learning. MIT press, Cambridge, USA.

Alves, C.A., Romero Yanes, J.F., Feitosa, F.X., et al., 2022. Influence of asphaltenes and resins on water/model oil interfacial tension and emulsion behavior: comparison of extracted fractions from crude oils with different asphaltene stability. J. Pet. Sci. Eng. 208, 109268. https://doi.org/10.1016/j.petrol.2021.109268.

American Petroleum Institute. Production Department, 1977. Recommended practices for laboratory testing of surface active agents for well stimulation. In: API, Dallas, USA.

Arjaria, S.K., Rathore, A.S., Cherian, J.S., 2021. Chapter 13 - kidney disease prediction using a machine learning approach: a comparative and comprehensive analysis. In: Kautish, P.N.S., Peng, S.L. (Eds.), Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics. Academic Press, pp. 307–333. https://doi.org/10.1016/B978-0-12-821633-0.00006-4.

Bai, W.P., Cheng, S.Q., Guo, X.Y., et al., 2024. Oilfield analogy and productivity prediction based on machine learning: field cases in PL Oilfield, China. Pet. Sci. https://doi.org/10.1016/j.petsci.2024.02.018.

Berrar, D., 2019. Cross-validation. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, United Kingdom, pp. 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

Brieuc, M.S., Waters, C.D., Drinan, D.P., et al., 2018. A practical introduction to Random Forest for genetic association studies in ecology and evolution. Mol. Ecol. Resour 18 (4), 755–766. https://doi.org/10.1111/1755-0998.12773.

Broomhead, D.S., Lowe, D., 1988. Radial basis functions, multi-variable functional interpolation, and adaptive networks. In: Technical Report. Royal Signals and Radar Establishment, Malvern, United Kingdom.

Castro-Amoedo, R., Granacher, J., Maréchal, F., 2024. A combined genetic algorithm and active learning approach to build and test surrogate models in Process Systems Engineering. Comput. Chem. Eng. 181, 108517. https://doi.org/10.1016/j.compchemeng.2023.108517.

Chawla, N.V., Bowyer, K.W., Hall, L.O., et al., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

da Silva, M., Sad, C.M., Pereira, L.B., et al., 2018. Study of the stability and homogeneity of water in oil emulsions of heavy oil. Fuel 226, 278–285. https://doi.org/10.1016/j.fuel.2018.04.011.

de Souza, U.F., Quina, F.H., Guardani, R., 2007. Prediction of emulsion stability via a neural network-based mapping technique. Ind. Eng. Chem. Res. 46 (15), 5100–5107. https://doi.org/10.1021/ie070337a.

Deng, S., Pan, H.Y., Wang, H.G., et al., 2024. A hybrid machine learning optimization algorithm for multivariable pore pressure prediction. Pet. Sci. 21 (1), 535–550. https://doi.org/10.1016/j.petsci.2023.09.001.

Dobos, D., Nguyen, T.T., Dang, T., et al., 2023. A comparative study of anomaly detection methods for gross error detection problems. Comput. Chem. Eng. 175, 108263. https://doi.org/10.1016/j.compchemeng.2023.108263.

Dong, S.Q., Sun, Y.M., Xu, T., et al., 2023. How to improve machine learning models for lithofacies identification by practical and novel ensemble strategy and principles. Pet. Sci. 20 (2), 733–752. https://doi.org/10.1016/j.petsci.2022.09.006.

Foroutan, S., Moghadasi, J., 2013. A neural network approach to predict formation damage due to calcium sulphate precipitation. In: SPE European Formation Damage Conference and Exhibition. SPE. https://doi.org/10.2118/165157-MS.

Fredd, C., Fogler, H.S., 1998. Alternative stimulation fluids and their impact on carbonate acidizing. SPE J. 3 (1), 34–41. https://doi.org/10.2118/31074-PA.

Gao, L., Xie, R.H., Xiao, L.Z., et al., 2022. Identification of low-resistivity-low-contrast pay zones in the feature space with a multi-layer perceptron based on conventional well log data. Pet. Sci. 19 (2), 570–580. https://doi.org/10.1016/j.petsci.2021.12.012.

Garud, S.S., Karimi, I.A., Kraft, M., 2018. LEAPS2: learning based evolutionary assistive paradigm for surrogate selection. Comput. Chem. Eng. 119, 352–370. https://doi.org/10.1016/j.compchemeng.2018.09.008.

Geurts, P., Irrthum, A., Wehenkel, L., 2009. Supervised learning with decision tree-based methods in computational and systems biology. Mol. Biosyst. 5 (12), 1593–1605. https://doi.org/10.1039/B907946G.

Greene, E., Lybarger, J., Richardson, E., 1974. In-situ acid neutralization system solves facility upset problems. J. Pet. Technol. 26 (10), 1153–1155. https://doi.org/10.2118/4796-PA.

Hagan, M.T., Menhaj, M.B., 1994. Training feedforward networks with the Marquardt algorithm. IEEE Trans. Neural Netw. 5 (6), 989–993. https://doi.org/10.1109/72.329697.

Hancock, J.T., Khoshgoftaar, T.M., 2020. CatBoost for big data: an interdisciplinary review. J. Big Data 7 (1), 94. https://doi.org/10.1186/s40537-020-00369-8.

Hemmati-Sarapardeh, A., Dabir, B., Ahmadi, M., et al., 2019. Modelling asphaltene precipitation titration data: a committee of machines and a group method of data handling. Can. J. Chem. Eng. 97 (2), 431–441. https://doi.org/10.1002/cjce.23254.

Hemmati-Sarapardeh, A., Varamesh, A., Husein, M.M., et al., 2018. On the evaluation of the viscosity of nanofluid systems: modeling and data assessment. Renewable Sustainable Energy Rev. 81, 313–329. https://doi.org/10.1016/j.rser.2017.07.049.

Hutin, A., Argillier, J.F., Langevin, D., 2016. Influence of pH on oil-water interfacial tension and mass transfer for asphaltenes model oils. Comparison with crude oil behavior. Oil Gas Sci. Technol. 71 (4), 58. https://doi.org/10.2516/ogst/2016013.

Kalhori, P., Abbasi, A., Malayeri, M.R., et al., 2022. Impact of crude oil components on acid sludge formation during well acidizing. J. Pet. Sci. Eng. 215, 110698. https://doi.org/10.1016/j.petrol.2022.110698.

Kamari, A., Gharagheizi, F., Bahadori, A., et al., 2014. Rigorous modeling for prediction of barium sulfate (barite) deposition in oilfield brines. Fluid Phase Equilib 366, 117–126. https://doi.org/10.1016/j.fluid.2013.12.023.

Khamehchi, E., Mahdiani, M.R., Amooie, M.A., et al., 2020. Modeling viscosity of light and intermediate dead oil systems using advanced computational frameworks and artificial neural networks. J. Pet. Sci. Eng. 193, 107388. https://doi.org/10.1016/j.petrol.2020.107388.

Kharisov, R.Y., Folomeev, A.E., Sharifullin, A.R., et al., 2012. Integrated approach to acid treatment optimization in carbonate reservoirs. Energy Fuels 26 (5), 2621–2630. https://doi.org/10.1021/ef201388p.

Kiannejad Amiri, M., Ghorbanzade Zaferani, S.P., Sarmasti Emami, M.R., et al., 2023. Multi-objective optimization of thermophysical properties GO powders-DW/EG Nf by RSM, NSGA-II, ANN, MLP and ML. Energy 280, 128176. https://doi.org/

10.1016/j.energy.2023.128176.
Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980.
Kumar, K., Panpalia, G., Priyadarshini, S., 2011. Application of artificial neural networks in optimizing the fatty alcohol concentration in the formulation of an O/W emulsion. Acta Pharm. 61 (2), 249–256. https://doi.org/10.2478/v10007-011-0013-7.
Lashkarbolooki, M., Hezave, A.Z., Ayatollahi, S., 2012. Artificial neural network as an applicable tool to predict the binary heat capacity of mixtures containing ionic liquids. Fluid Phase Equilib. 324, 102–107. https://doi.org/10.1016/j.fluid.2012.03.015.
Lee, W.H., Park, C.Y., Diaz, D., et al., 2022. Predicting bilgewater emulsion stability by oil separation using image processing and machine learning. Water Res. 223, 118977. https://doi.org/10.1016/j.watres.2022.118977.
Liu, B., Mohammadi, M.-R., Ma, Z., et al., 2023a. Experimental investigation and intelligent modeling of pore structure changes in type III kerogen-rich shale artificially matured by hydrous and anhydrous pyrolysis. Energy 282, 128799. https://doi.org/10.1016/j.energy.2023.128799.
Liu, B., Mohammadi, M.-R., Ma, Z., et al., 2023b. Pore structure characterization of solvent extracted shale containing kerogen type III during artificial maturation: experiments and tree-based machine learning modeling. Energy 283, 128885. https://doi.org/10.1016/j.energy.2023.128885.
Liu, H., Ren, Y.L., Li, X., et al., 2022. Rock thin-section analysis and identification based on artificial intelligent technique. Pet. Sci. 19 (4), 1605–1621. https://doi.org/10.1016/j.petsci.2022.03.011.
Liu, W., Chen, Z., Hu, Y., et al., 2023c. A systematic machine learning method for reservoir identification and production prediction. Pet. Sci. 20 (1), 295–308. https://doi.org/10.1016/j.petsci.2022.09.002.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. NIPS.
Lv, Q., Zheng, R., Guo, X., et al., 2023. Modelling minimum miscibility pressure of CO_2-crude oil systems using deep learning, tree-based, and thermodynamic models: application to CO_2 sequestration and enhanced oil recovery. Sep. Purif. Technol. 310, 123086. https://doi.org/10.1016/j.seppur.2022.123086.
Meng, J., Zhou, Y.J., Ye, T.R., et al., 2023. Hybrid data-driven framework for shale gas production performance analysis via game theory, machine learning, and optimization approaches. Pet. Sci. 20 (1), 277–294. https://doi.org/10.1016/j.petsci.2022.09.003.
Minakov, A.V., Mikhienkova, E.I., Pryazhnikov, M.I., et al., 2022. Experimental study of the rheological properties and stability of highly-concentrated oil-based emulsions. J. Mol. Liq. 349, 118125. https://doi.org/10.1016/j.molliq.2021.118125.
Mirvakili, A., Rahimpour, M.R., Jahanmiri, A., 2012. Effect of a cationic surfactant as a chemical destabilization of crude oil based emulsions and asphaltene stabilized. J. Chem. Eng. Data 57 (6), 1689–1699. https://doi.org/10.1021/je2013268.
Mohammadi, M.-R., Hemmati-Sarapardeh, A., Schaffie, M., et al., 2021. Application of cascade forward neural network and group method of data handling to modeling crude oil pyrolysis during thermal enhanced oil recovery. J. Pet. Sci. Eng. 205, 108836. https://doi.org/10.1016/j.petrol.2021.108836.
Mohammadzadeh Shirazi, M., Ayatollahi, S., Ghotbi, C., 2019. Damage evaluation of acid-oil emulsion and asphaltic sludge formation caused by acidizing of asphaltenic oil reservoir. J. Pet. Sci. Eng. 174, 880–890. https://doi.org/10.1016/j.petrol.2018.11.051.
Pang, H.W., Wang, H.Q., Xiao, Y.T., et al., 2023. Machine learning for carbonate formation drilling: mud loss prediction using seismic attributes and mud loss records. Pet. Sci. 21 (2), 1241–1256. https://doi.org/10.1016/j.petsci.2023.10.024.
Patel, N., Upadhyay, S., 2012. Study of various decision tree pruning methods with their empirical comparison in WEKA. Int. J. Comput. Appl. 60, 20–25. https://doi.org/10.5120/9744-4304.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html.
Pei, Z.J., Song, X.Z., Wang, H.T., et al., 2024. Interpretation and characterization of rate of penetration intelligent prediction model. Pet. Sci. 21 (1), 582–596. https://doi.org/10.1016/j.petsci.2023.10.011.
Pourakaberian, A., Ayatollahi, S., Shirazi, M.M., et al., 2021. A systematic study of asphaltic sludge and emulsion formation damage during acidizing process: experimental and modeling approach. J. Pet. Sci. Eng. 207, 109073. https://doi.org/10.1016/j.petrol.2021.109073.
Prokhorenkova, L., Gusev, G., Vorobev, A., et al., 2018. CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems (NeurIPS 2018). Montreal, Canada.
Qiao, C., Yu, X., Song, X., et al., 2020. Enhancing gas solubility in nanopores: a combined study using classical density functional theory and machine learning. Langmuir 36 (29), 8527–8536. https://doi.org/10.1021/acs.langmuir.0c01160.
Rezaian, A., Kordestany, A., Haghighat Sefat, M., 2010. An artificial neural network approach to formation damage prediction due to Asphaltene deposition. In: SPE Nigeria Annual International Conference and Exhibition. SPE. https://doi.org/10.2118/208185-MS.
Rietjens, M., Nieuwpoort, M., 2001. An analysis of crude oil–acid reaction products by size-exclusion chromatography. Fuel 80 (1), 33–40. https://doi.org/10.1016/S0016-2361(00)00073-9.
Samuel, A.L., 1959. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 3 (3), 210–229. https://doi.org/10.1147/rd.441.0206.
Shaikhina, T., Lowe, D., Daga, S., et al., 2019. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed. Signal Process Control 52, 456–462. https://doi.org/10.1016/j.bspc.2017.01.012.
Shakouri, S., Mohammadzadeh-Shirazi, M., 2023. Modeling of asphaltic sludge formation during acidizing process of oil well reservoir using machine learning methods. Energy 129433. https://doi.org/10.1016/j.energy.2023.129433.
Shi, J.X., Zhao, X.Y., Zeng, L.B., et al., 2023. Identification of reservoir types in deep carbonates based on mixed-kernel machine learning using geophysical logging data. Pet. Sci. 21 (3), 1632–1648. https://doi.org/10.1016/j.petsci.2023.12.016.
Singh, P., Singh, N., Singh, K.K., et al., 2021. Diagnosing of disease using machine learning. In: Machine Learning and the Internet of Medical Things in Healthcare. Academic Press, Oxford, United Kingdom, pp. 89–111. https://doi.org/10.1016/B978-0-12-821229-5.00003-3.
Song, Y.Y., Lu, Y., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 27 (2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.
Taylor, K.C., Nasr-El-Din, H.A., Al-Alawi, M.J., 1999. Systematic study of iron control chemicals used during well stimulation. SPE J. 4 (1), 19–24. https://doi.org/10.2118/54602-PA.
Umar, A.A., Saaid, I.B.M., Sulaimon, A.A., et al., 2018. A review of petroleum emulsions and recent progress on water-in-crude oil emulsions stabilized by natural surfactants and solids. J. Pet. Sci. Eng. 165, 673–690. https://doi.org/10.1016/j.petrol.2018.03.014.
Wasserman, P.D., Schwartz, T., 1988. Neural networks. II. What are they and why is everybody so interested in them now? IEEE Expert 3 (1), 10–15. https://doi.org/10.1109/64.2091.
Wojnar, T.J., 2018. 2018 outlook for energy: a view to 2040. In: 2018 AIChE Annual Meeting. AIChE, Pittsburgh, USA.
Wu, Y., Misra, S., 2019. Intelligent image segmentation for organic-rich shales using random forest, wavelet transform, and hessian matrix. IEEE Geosci. Remote Sens. Lett. 17 (7), 1144–1147. https://doi.org/10.1109/LGRS.2019.2943849.
Xia, W.H., Zhao, Z.X., Li, C.X., et al., 2023. Intelligent risk identification of gas drilling based on nonlinear classification network. Pet. Sci. 20 (5), 3074–3084. https://doi.org/10.1016/j.petsci.2023.04.003.
Xue, L., Xu, S., Nie, J., et al., 2024. An efficient data-driven global sensitivity analysis method of shale gas production through convolutional neural network. Pet. Sci. https://doi.org/10.1016/j.petsci.2024.02.010.
Yan, T., Xu, R., Sun, S.H., et al., 2023. A real-time intelligent lithology identification method based on a dynamic felling strategy weighted random forest algorithm. Pet. Sci. 21 (2), 1135–1148. https://doi.org/10.1016/j.petsci.2023.09.011.
Yang, H., Fong, S., 2013. Incremental optimization mechanism for constructing a decision tree in data stream mining. Math. Probl Eng. 2013 (1), 580397. https://doi.org/10.1155/2013/580397.
Yao, Y., Qiu, Y., Cui, Y., et al., 2023. Insights to surfactant huff-puff design in carbonate reservoirs based on machine learning modeling. Chem. Eng. J. 451, 138022. https://doi.org/10.1016/j.cej.2022.138022.
Yetilmezsoy, K., Fingas, M., Fieldhouse, B., 2011. An adaptive neuro-fuzzy approach for modeling of water-in-oil emulsion formation. Colloids Surf., A. 389 (1), 50–62. https://doi.org/10.1016/j.colsurfa.2011.08.051.
Zhang, C., Zhang, R., Zhu, Z., et al., 2023. Bottom hole pressure prediction based on hybrid neural networks and Bayesian optimization. Pet. Sci. 20 (6), 3712–3722. https://doi.org/10.1016/j.petsci.2023.07.009.
Zhang, J., Tian, D., Lin, M., et al., 2016. Effect of resins, waxes and asphaltenes on water-oil interfacial properties and emulsion stability. Colloids Surf., A. 507, 1–6. https://doi.org/10.1016/j.colsurfa.2016.07.081.
Zhang, T., Bhatia, A., Pandya, D., et al., 2020. Industrial text analytics for reliability with derivative-free optimization. Comput. Chem. Eng. 135, 106763. https://doi.org/10.1016/j.compchemeng.2020.106763.
Zuluaga, E., Alvarez, H.D., Alvarez, J.D., 2002. Prediction of permeability reduction by external particle invasion using artificial neural networks and fuzzy models. J. Can. Pet. Technol. 41 (6), 19–24. https://doi.org/10.2118/02-06-01.