scientific reports

Check for updates

OPEN

Modeling interfacial tension of surfactant–hydrocarbon systems using robust tree-based machine learning algorithms

Ali Rashidi-Khaniabadi[1], Elham Rashidi-Khaniabadi[2], Behnam Amiri-Ramsheh[3], Mohammad-Reza Mohammadi[3] & Abdolhossein Hemmati-Sarapardeh[3,4]✉

Interfacial tension (IFT) between surfactants and hydrocarbon is one of the important parameters in petroleum engineering to have a successful enhanced oil recovery (EOR) operation. Measuring IFT in the laboratory is time-consuming and costly. Since, the accurate estimation of IFT is of paramount significance, modeling with advanced intelligent techniques has been used as a proper alternative in recent years. In this study, the IFT values between surfactants and hydrocarbon were predicted using tree-based machine learning algorithms. Decision tree (DT), extra trees (ET), and gradient boosted regression trees (GBRT) were used to predict this parameter. For this purpose, 390 experimental data collected from previous studies were used to implement intelligent models. Temperature, normal alkane molecular weight, surfactant concentration, hydrophilic–lipophilic balance (HLB), and phase inversion temperature (PIT) were selected as inputs of models and independent variables. Also, the IFT between the surfactant solution and normal alkanes was selected as the output of the models and the dependent variable. Moreover, the implemented models were evaluated using statistical analyses and applied graphical methods. The results showed that DT, ET, and GBRT could predict the data with average absolute relative error values of 4.12%, 3.52%, and 2.71%, respectively. The R-squared of all implementation models is higher than 0.98, and for the best model, GBRT, it is 0.9939. Furthermore, sensitivity analysis using the Pearson approach was utilized to detect correlation coefficients of the input parameters. Based on this technique, the results of sensitivity analysis demonstrated that PIT, surfactant concentration, and HLB had the greatest effect on IFT, respectively. Finally, GBRT was statistically credited by the Leverage approach.

Interfacial tension (IFT) is a parameter of interest in petroleum and chemical science and engineering[1–3]. It plays a vital role in multiphase flow, separation processes, formation and stability of emulsions, fluid transportation, and reservoir engineering processes like fluid contacts, fluid saturation distribution, recovery mechanisms, and enhanced oil recovery (EOR) processes[4–8].

Capillary pressure plays a critical role in oil recovery at all stages of production from oil reservoirs. The capillary number (N_c) concept and equation are used to investigate the effect of capillary pressure on oil recovery from the reservoir. The general form of the capillary number is defined as follows[9,10]:

$$N_c = \frac{(\text{viscous force})}{(\text{capillary force})} = \frac{\mu v}{\sigma \cos\theta} \quad (1)$$

where θ is the contact angle, σ is the IFT between the wetting and non-wetting phase, μ is the viscosity of the displacing phase, and v is the Darcy velocity. The amount of oil saturation remaining in the porous medium has a strong correlation with the capillary number[9–15]. Researchers concluded that increasing the capillary number increases oil recovery[16]. Also, capillary number was considered as the primary variable in several modeling

[1]Department of Petroleum Engineering, EOR Research Center, Omidiyeh Branch, Islamic Azad University, Omidiyeh, Iran. [2]Department of Mathematics, Yazd University, Yazd, Iran. [3]Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. [4]State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing, China. ✉email: hemmati@uk.ac.ir; aut.hemmati@gmail.com

and simulation studies related to IFT and wettability alteration[15,17,18]. According to Eq. (1), decreasing the IFT increases the capillary number. EOR techniques produce residual oil by optimizing the amount of capillary number[18–21]. Surfactants are amphiphilic molecules that are soluble in both organic solvents and water[22]. The surfactant reduces the IFT between oil and water by adsorbing at the liquid–liquid interface[23,24]. It was found that an oil droplet on the meniscus could be attracted to the wall when surfactant is added to water, while bubbles always move towards the walls; IFT and gravity play significant roles in these cases[25,26]. Researchers have conducted numerous laboratory studies to investigate the ability of surfactants to reduce the IFT between aqueous solution and oil for use in EOR techniques[21,27–29]. The measurement of IFT of a water-hydrocarbon interface in the presence of surfactants is of great interest for surfactant flooding. Various parameters affect the IFT between the solution containing surfactant and hydrocarbons that must be considered. Experimental studies have shown that the type[28,30] and the surfactant concentration[31], the temperature of the aqueous solution[32], and the hydrocarbon composition[33,34] can affect the IFT. Surfactants are divided into two general categories, which include ionic and nonionic surfactants. Ionic surfactants have a positive or negative electric charge, or both, classified into cationic, anionic, and amphoteric surfactants, respectively[35]. However, nonionic surfactants do not have an electric charge. The interfacial behavior of surfactants can vary depending on their structure[2,35,36]. Therefore, many researchers have investigated the role of surfactant structure in reducing the IFT between hydrocarbons and aqueous solutions. Strey[37] showed that with increasing temperature, the IFT behavior of the surfactant is curved and has a minimum point. It was found that before the minimum point of IFT, increasing the temperature which leads to an increase in the number of surfactant molecules at the interface between hydrocarbon and aqueous solution, reduces the IFT value[37]. Also, increasing the surfactant concentration to the critical micelle concentration (CMC) reduces the IFT[35,38].

The best way to measure the IFT between surfactant and hydrocarbon is performing laboratory methods. Laboratory methods for measuring IFT are the weight of drop method[39,40] pendant drop[41–43], spinning drop[44–46], etc. Time-consuming is one of the limitations and challenges of laboratory methods. Considering the price of the chemicals used to perform the IFT test and the cost of the tests, this method is costly. Therefore, developing a model for predicting the IFT between surfactants and hydrocarbons can be very attractive and practical. Previous studies have described the effect of surfactants on the interfacial boundary of two fluids with the surface equation of state[47,48]. The surface equation of state is a relationship between the surface concentration of surfactant and surface pressure[48]. The difference between the IFT without surfactant and after a surfactant to the solution is equal to the surface pressure[48]. Also, the concentration of surfactant molecules on the surface is defined by surface adsorption[49]. Different approaches to obtaining state equations were discussed in the literature. The Szyszkowski equation[50], the Frumkin equation[51], and the Langmuir model[52] are three examples of semi-empirical equations for the surface equation of state. The Langmuir model was used to describe the adsorption of nonionic surfactants at the interface between hydrocarbons and aqueous solution. This model cannot predict the effect of a surfactant mixture solution on IFT. It is also not suitable for describing the interfacial behavior of surfactants in the presence of inorganic ions[53]. Mulqueen and Blankschtein (2002)[54] developed a different molecular-thermodynamic approach to predict surfactant adsorption at the oil/water and air/water interfaces. They evaluated the validity of this model only on a limited number of laboratory data, and it was valid only for decane/water interface[54]. Bahramian and Zarbakhsh (2015)[48] performed studies to estimate the IFT of ionic surfactants. In their proposed model, they considered the size of the surfactant molecule and the CMC value of a surfactant in the aqueous solution as independent variables. In this method, a laboratory test set is required to obtain the CMC[48]. In a recent study, Nikseresht et al. (2019)[55] used the Butler equation to estimate the IFT between ionic surfactants and normal alkanes as the oil phase. This model is based on the surface state equation. For each case, two fitting parameters need to be set. In other words, the set model is not suitable for other conditions. They also examined Bahramian and Zarbakhsh's[48] equation for various surfactants and concluded that it could not be satisfactorily used to predict IFT in the presence of C_{10}TAB and C_{12}TAB[55]. In summary, thermodynamic models for estimating IFT have the following limitations: (1) they require laboratory testing to calculate the input parameters. (2) Each model fitted to a surfactant does not apply to other cases and conditions. (3) these models were evaluated for limited experimental data. (4) All parameters affecting IFT were not considered in these models. On the other hand, machine learning methods can model and solve complex numerical problems in industry[56–58]. Predicting the IFT between two immiscible fluids using intelligent methods has been considered by many researchers[59–62]. In previous studies, the ability of intelligent methods to estimate the IFT between hydrocarbons and the aqueous solution was evaluated[63–67]. It was found that intelligent methods are appropriate for this purpose. As the literature review shows, the IFT studies in recent years are mostly focused on the aqueous phase and hydrocarbons, and the role of surfactants is less considered. To the best of the authors' knowledge, no reliable and comprehensive model has been presented for this case. Therefore, there is a window to develop a reliable model for predicting the IFT of surfactants and hydrocarbon systems.

This study aims to develop accurate and reliable models to estimate the IFT between ionic surfactants and normal alkanes. Decision tree (DT), extra trees (ET), and gradient boosted regression trees (GBRT) models are implemented for this purpose. Temperature, normal alkane molecular weight, surfactant concentration, hydrophilic–lipophilic balance (HLB), and phase inversion temperature (PIT) are selected as inputs and independent variables. Also, the IFT between the surfactant solution and normal alkanes is selected as the output and the dependent variable. Sample data are collected from the literature to train and evaluate the models. The trial-and-error method is used to optimize the implemented models. In the present study, the implemented models are evaluated using statistical analysis and applied graphical methods. Furthermore, sensitivity analysis is performed on how changes in the model's inputs impact the IFT values. Finally, the leverage method is carried out to ensure the credibility of the gathered IFT databank and the accuracy and dependability of the best model for estimating the IFT between ionic surfactants and normal alkanes. Hence, the main contributions of this research are as follows:

- Collecting an extensive database of IFT of surfactant–hydrocarbon systems including important parameters such as HLB and PIT_x, which have a significant impact on the better characterization of surfactants.
- Development of accurate models with low error using robust tree-based machine learning algorithms.
- Performing sensitivity analysis to detect the relative effect of temperature, normal alkane molecular weight, surfactant concentration, HLB, and PIT_x on the IFT of surfactant–hydrocarbon systems.
- Implementation of leverage method to identify suspicious and outlier data related to IFT of surfactant–hydrocarbon systems reported in the literature.

Data gathering

In order to develop the models, 390 sample data were collected from previous studies[68–76]. The sample dataset contains temperature, normal alkane molecular weight, surfactant concentration, HLB, and PIT_x. In this paper, five different surfactants were used, and their specifications were presented in Table 1. Also, n-hexane, n-heptane, n-octane, n-nonane, n-decane, n-undecane, n-dodecane, n-tetradecane, and n-heptadecane were used as normal alkanes. HLB and PIT_x were used to represent the type of surfactants. The HLB value determines the hydrophilicity and lipophilic of a surfactant[77,78]. Researchers have developed various methods for calculating the amount of the HLB[79,80]. In this study, the method introduced by Davies (1957)[80] was used to calculate the HLB. Davis method calculates the value of the HLB based on the group number as follows:

$$\mathrm{HLB}_{\mathrm{Davies}} = 7 + \Sigma\,(\text{hydrophilic group numbers}) - \Sigma(\text{lipophilic group numbers}) \quad (2)$$

The hydrophilic group numbers and the lipophilic group numbers are obtained from tables provided by Davies (1957)[80]. The values of the HLB for the surfactants used in this study are presented in Table 1. Another parameter used to characterize the surfactant is the PIT. This parameter depends on the structure of the surfactant[76]. For better visualization, the structures of surfactants utilized in this work are depicted in Fig. 1. The amount of the PIT_x of the five surfactants used in this study are presented in Table 1. Moreover, the statistical parameters of the databank used in this work are represented in Table 2. The collected data were divided into training and

Surfactant		Chemical formula	dPIT/dx	HLB
Decyl trimethyl ammonium bromide	C_{10}TAB	$C_{10}H_{21}N(CH_3)_3Br$	338	21
Dodecyl trimethyl ammonium bromide	C_{12}TAB	$C_{12}H_{25}N(CH_3)_3Br$	486	19
Myristyl trimethyl ammonium bromide	C_{14}TAB	$C_{14}H_{29}N(CH_3)_3Br$	453	18
Hexadecyl trimethyl ammonium bromide	C_{16}TAB	$C_{16}H_{33}N(CH_3)_3Br$	426	17
Sodium dodecyl sulfate	SDS	$C_{12}H_{25}NaSO_4$	499	40

Table 1. Characteristics of surfactants used in this study.

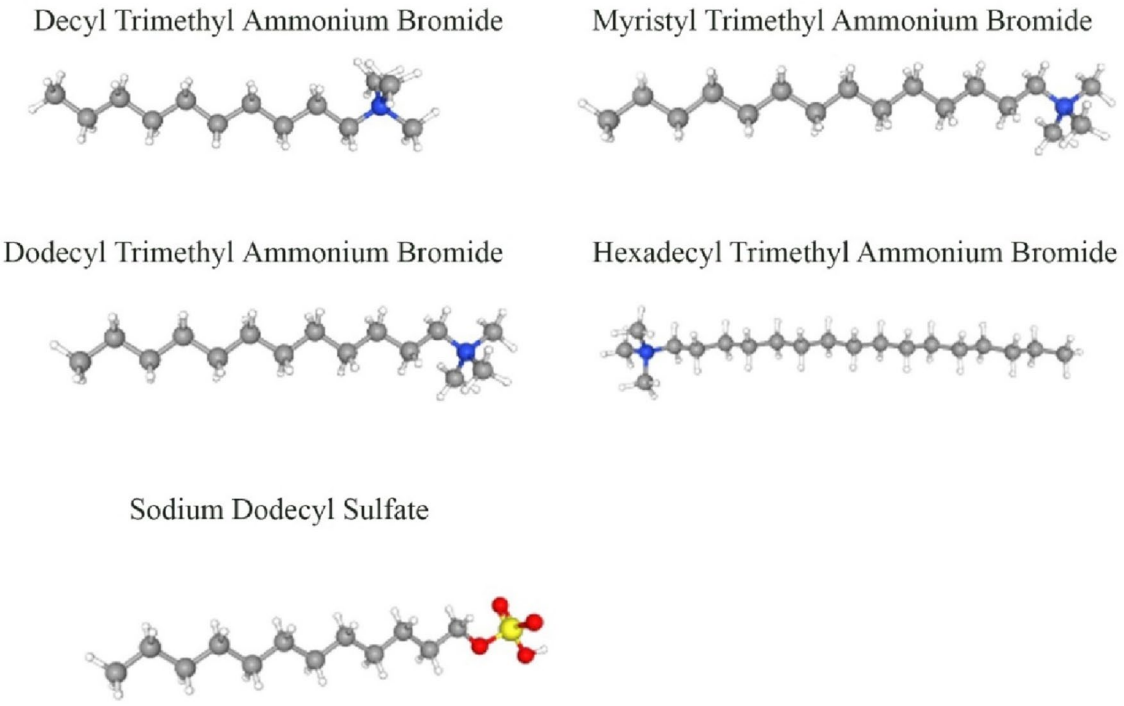


Figure 1. Structures of surfactants used in this study.

	HLB	dPIT/dx (°C)	Conc × 10^5 (mol/l)	T (K)	Mw (g/mol)	IFT (mN/m)
Mean	20.86	347.87	303.35	300.01	125.05	36.64
SD	14.45	200.32	714.44	7.68	40.80	14.39
Min	0	0	0	283.15	86.178	4.98
Skewness	−0.04	−1.02	4.89	2.07	1.03	−0.68
Kurtosis	−1.09	−0.71	28.44	5.81	0.45	−0.87
Max	40	499	6896.76	333.15	240.47	53.54
Status	Input	Input	Input	Input	Input	Output

Table 2. Statistical parameters of the databank.

testing categories for model development. In all modeling techniques, 80% of the sample dataset was used to train algorithms, and the remaining 20% was used to evaluate the models' performance. At this stage, the data were randomly divided.

Model development

Decision tree. The DT is used for regression and classification issues[81]. A simple structure of the DT is depicted in Fig. 2. A regression DT can predict numerical responses corresponding to independent variables. These types of algorithms are used in complex datasets. Decision trees are intuitive and interpretable[82,83]. In decision tree regression (DTR), it constantly divides the initial input space into smaller subsets and incrementally makes the final DT with decision and leaf nodes. ID3, C4.5, C5, and Classification and Regression tree (CART) are standard DT algorithms. C4.5 is an improved version based on ID3 and has the following advantages: (1) it can work with incomplete data, (2) it can use the pruning technique to prevent over-fitting, and (3) accepting discrete and continuous features. The CART is very similar to the C4.5. The difference between the two algorithms is that the CART does not calculate the rules and can also solve regression problems[84]. In this study, the optimized version of the CART algorithm was used.

Dividing nodes in the training process of the network is one of the most critical parts of implementing a DT algorithm. In CART, it uses a Gini coefficient to divide the nodes[85]. The DT implemented in that study consists of four stages[85]. In the first stage, the DT grows using the division of nodes. After dividing the training data into two parts, with the same logic, it divides these subsets again and so on. The DT greedily searches for an optimal division. This algorithm repeats the data segmentation in each step and does not check whether it leads to less impurity in the next steps. Each node is assigned to a predicted target based on the target's distribution in the sample data in nodes. This process continues until it cannot find a partition that reduces the impurity. Also, when the tree reaches its maximum state, the DT's growth process will stop. It is necessary to optimize the maximum depth value for this algorithm. In the second stage, after the tree's maximum growth, the process of building the tree stops. At this stage, the DT may not accurately predict the target value based on the test data. The third step involves pruning and simplifying the tree, which makes it better to predict. In the fourth step, the best tree is selected from the pruned trees.

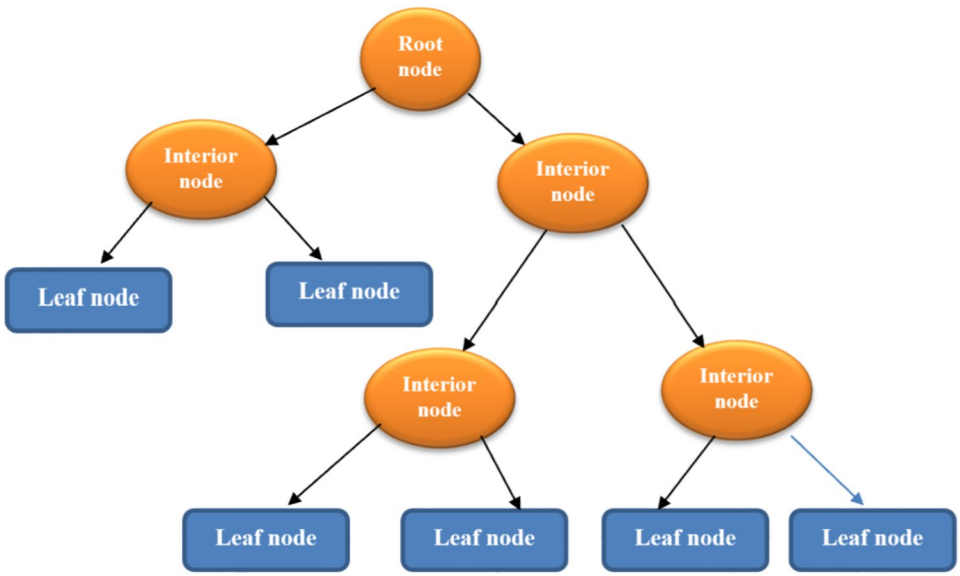


Figure 2. A simple structure of the DT.

The DT continuously divides the data into smaller sections during the same process to homogenize the data in the partition. The splitting rules may be set to optimize a criterion related to the target's predictive value, or the rules may be set to minimize local node impurity or dependent variables over the training set.

Identifying the number of training data points for the DT is a significant issue because it will cause over-fit if the sample data is low. Adding any level to the tree may lead to an increase in the number of samples required to learn the DT. The size of the tree should be controlled to prevent overfitting[86]. The main parameters for DT optimization include tree depth, minimum sample division, and minimum sample leaf. Ensemble methods can prevent over-fitting in the DT algorithm[87,88].

Extra trees. The ET creates a stronger model by combining several decision trees. The ET is one of the ensemble methods. In the ET method, the node is divided completely randomly by selecting the cutting points. Each DT grows independently, and all learning samples are used to grow the trees. The predicted target values are then added up for the final prediction. Finally, it predicts the final answer using the mathematical mean of the predicted values obtained from each base model[89]. The ET algorithm builds an ensemble model based on the explicit randomization of cutting points and feature combinations using averaging. Also, using all learning data to build base models can minimize the bias of the final model[90]. The tree growth method's complexity in the ET algorithm, assuming the trees are balanced, is similar to other tree growth methods[89]. This algorithm has three parameters including N_{min}, which denotes the minimum sample size to divide a node, K shows the number of randomly selected attributes in each node, and M illustrates the number of trees used as the base model. It is necessary to optimize these parameters to develop a more robust model based on additional trees. Each of these parameters has a different effect. The value of parameter N_{min} affects the average noise output of the model. The larger the value of N_{min}, the smaller the trees are made. As a result, the variance decreases, and the bias increases. The minimum size of sample data for node splitting should be optimized according to the model's amount of output noise. Obviously, in regression problems, high noise levels lead to overfitting. Geurts et al.[89] suggested that a higher value for parameter N_{min} should be used to build a stronger model when the data has more noise. In other words, to optimize ET in high noise conditions, it is necessary to increase the value of N_{min}. The number of selected attributes can also determine the strength of the attribute selection process. The maximum value that can be considered for K is the number of input features of the model. The low value of parameter K increases the randomness of the trees. It also makes the structure of the trees less dependent on the target value of the learning samples. Therefore, if we set the K to 1, the divisions are selected completely independent of the target variable. Also, if the value of this parameter is equal to the number of features in the learning data, the features are not explicitly selected randomly, and the randomization effect is applied only by selecting the cut points[89]. A schematic structure of the ET algorithm is illustrated in Fig. 3.

Gradient boosted regression trees (GBRT). Boosting is another method of an ensemble that combines several weak learners to create a stronger model for target prediction[85]. This method is used to solve regression and classification problems. The weak learners are trained one after the other, each focusing on correcting the previous step[91]. In this study, the GBRT was used, in which the DT is defined as the basic learner.

In a modeling issue, one has a system consisting of a set of random "explanatory" variables or "input" $x = \{\times 1; \ldots; xn\}$ and random "response" variables or "output" y. The goal is to create a function $F^{*}(x)$ that relate y to x.

$$F^{*}(x) = \arg_{F(x)}{}^{\min} E_{y,x}\,\Psi\left(y, F(x)\right) \tag{3}$$

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m) \tag{4}$$

the expected value of some specified loss function $\Psi\left(y, F(x)\right)$ is minimized. Here, $h(x; a_m)$ is a simple function of x defined as a basic learner. The expansion coefficients $\{\beta_m\}_0^m$ and the parameters $\{a_m\}_0^m$ are jointly fit to the training data in a forward "stage-wise" manner.

$$(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i-1}^{N} \Psi\left(y_i, F_{m-1}(x_i) + \beta h(x_i; a)\right) \tag{5}$$

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m) \tag{6}$$

Equation (5) can be solved in two steps for a given cost function by the Gradient Boosting method[92,93]. First, the function $h(x; a)$ is fit by least squares:

$$a_m = \text{argmin}_{a,\rho} \sum_{i=1}^{N} \left[\widetilde{y}_{im} - \rho h(x_i; a)\right]^2 \tag{7}$$

$$\widetilde{y}_{im} = -\left[\frac{\partial \Psi\left(y_i, F(x_i)\right)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{8}$$

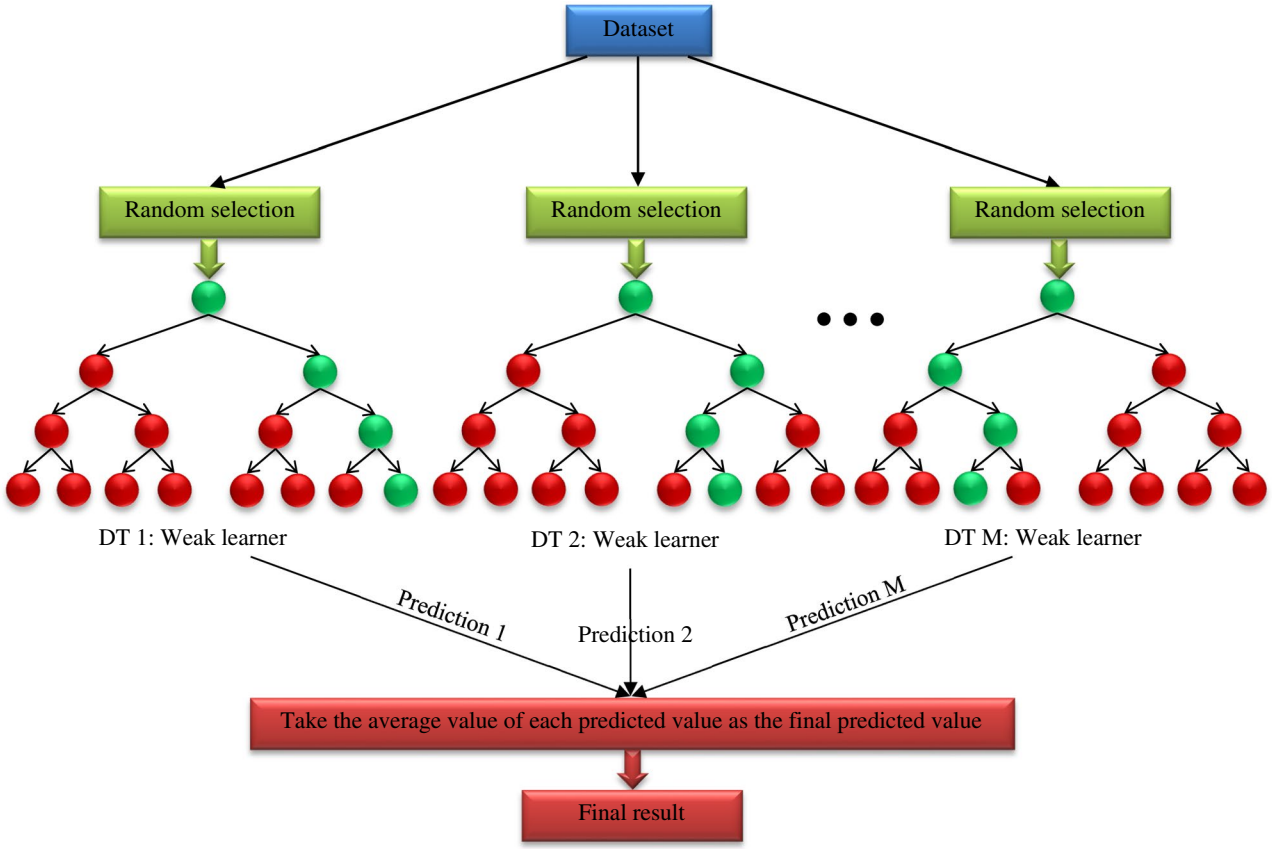


Figure 3. Schematic structure of the ET.

In the next step, according to $h(x; a_m)$, the optimal value of β_M is determined. GBRT specializes in this strategy to the case where the weak learner $h(x; a)$ is an L terminal node regression tree. At each iteration m, a regression tree is divided the x space into L-disjoint regions $\{R_{lm}\}_{l=1}^{L}$.

$$h\left(x; \{R_{lm}\}_1^L\right) = \sum_{l-1}^{L} \bar{y}_{lm} 1(x \in R_{lm}) \tag{9}$$

$$\bar{y}_{im} = \text{mean}_{xi \in R_{im}}\left(\tilde{y}_{im}\right) \tag{10}$$

the solution to Eq. (8) reduces to a simple "location" estimate based on the criterion Ψ.

$$y_{lm} = \arg\min_y \sum_{x_i \in R_{lm}} \Psi\left(y_i, F_{m-1}(x_i) + y\right) \tag{11}$$

It will be updated separately in each corresponding area.

$$F_m(x) = F_{m-1}(x) + \nu . y_{lm} 1(x \in R_{lm}) \tag{12}$$

The learning rate is controlled by the "shrinkage" parameter $0 < \nu \leq 1$. The small values of this parameter ($\nu \leq 0.1$) lead to a much better generalization error. Friedman (1999)[92] presented specific algorithms based on this template for several loss criteria, including least-absolute deviation, least squares, Huber, and for classification, K class multinomial negative log-likelihood. It should be noted that hyper-parameters should be considered to optimize the implemented model. These parameters such as the number of base estimators, subsample, loss function, maximum depth, the minimum number of leaf nodes, the maximum number of features, and the minimum number of sample split samples, define the structure of the network. A simple architecture of the GBRT algorithm is depicted in Fig. 4.

Results and discussion

Description of the models' development. In the present work, three different data-driven techniques, including DT, ET, and GBRT were developed to establish accurate models for the estimation of the IFT between ionic surfactants and normal alkanes. As mentioned, in order to create a more robust and faster model, the specific hyperparameters of each algorithm must be adjusted and optimized. As mentioned earlier, the trial-and-

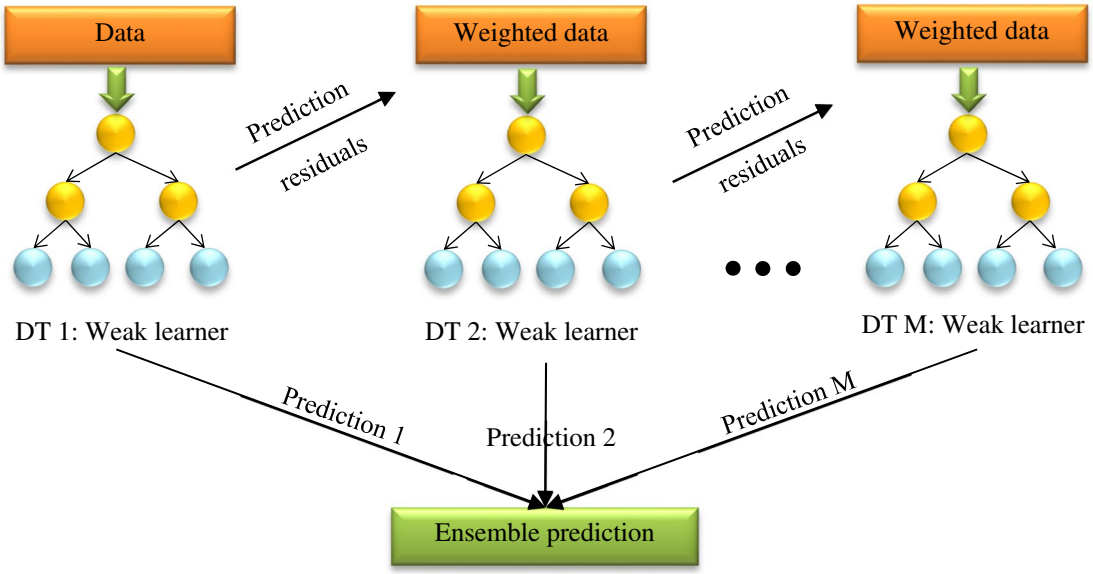


Figure 4. A simple architecture of the GBRT.

error method was employed to optimize the implemented models. The value of the maximum depth parameter of the DT strongly affects the speed and accuracy of the model. The depth of the tree should be carefully adjusted so as not to cause over-fitting and under-fitting. As reported in Table 3, the best value for this parameter is 7. The proposed control parameters for implementing the DT algorithm based on the sample data used in this study are reported in Table 3. In this study, two ensemble algorithms based on the DT were used. Ensemble methods were used to increase the stability and accuracy of the DT model. Ensemble models can also prevent over-fitting and create a robust and stable model based on the base estimator[94]. Due to the nature of ensemble models, the number of estimators is the most important parameter for optimization. To build an accurate model based on the GBRT algorithm, loss function, learning rate, number of estimators, subsample, maximum depth, and

Model	Parameter	Value
GBRT	learning rate	0.12
	loss	Huber
	n_estimators	60
	sub_sample	0.17
	criterion	Friedman mse
	min_sample_split	2
	min_sample_leaf	1
	max_depth	9
	alpha	0.97
DT	criterion	Friedman mse
	min_sample_split	2
	min_sample_leaf	1
	max_depth	7
	ccp_alpha	0.0075
	splitter	Best
	max_features	None
ET	criterion	Friedman mse
	min_sample_split	2
	min_sample_leaf	1
	max_depth	12
	n_estimators	70
	Bootstrap	True

Table 3. Internal parameters of the developed models.

alpha must be considered and adjusted. The adjusted parameters for the models implemented in this study were reported in Table 3.

Statistical evaluation. In this study, statistical criteria were used to evaluate the accuracy and ability of the developed models in predicting IFT. For this purpose, statistical parameters including determination coefficient (R^2), average percent relative error (APRE, %), root mean square error (RMSE), standard deviation (SD), and average absolute percent relative error (AAPRE, %), were used. The formulas of these statistical parameters are listed as follows[95]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(IFT_{exp_i} - IFT_{pred_i}\right)^2}{\sum_{i=1}^{n} \left(IFT_{exp_i} - \overline{IFT}\right)^2} \tag{13}$$

$$APRE = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{IFT_{exp_i} - IFT_{pred_i}}{IFT_{exp_i}} \right) \times 100 \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{1}^{n} \left(IFT_{exp_i} - IFT_{pred_i}\right)^2} \tag{15}$$

$$AAPRE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{IFT_{exp_i} - IFT_{pred_i}}{IFT_{exp_i}} \right| \times 100 \tag{16}$$

$$SD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{IFT_{exp_i} - IFT_{pred_i}}{IFT_{exp_i}} \right)^2} \tag{17}$$

If the value of R^2 is high and the values of AAPRE, APRE, RMSE, and SD are low, the model predicts the IFT with higher accuracy. The maximum value of R^2 is one, and the lowest value of the AAPRE value is zero. Statistical parameters for evaluating the models implemented in this study at different development stages are presented in Table 4. According to the RMSE values presented in Table 4, the accuracy of the models implemented in this study is as follows:

$$\text{GBRT} > \text{ET} > \text{DT}; \text{ for both training and testing phases.}$$

Graphical error analysis. In this section, graphical error analysis shows the models' validity and accuracy. Therefore, four graphs, including bar-plot, cross-plot, relative error distribution, and cumulative frequency diagram were investigated. Figure 5 is a cross-plot of DT, ET, and GBRT models. This diagram plots the predicted values in the training and testing phases versus the experimental values. In this type of diagram, if the train and test points are close to the unit slope line (X = Y), it indicates that the model can predict with high accuracy. As Fig. 5 shows, some of the test points of the DT and ET models are above or below the X = Y line, which indicates

		GBRT model	ET model	DT model
Train	SD	0.041	0.065	0.058
	RMSE	0.96	1.223	1.416
	APRE	−0.29	−1.32	−0.33
	AAPRE	2.47	3.17	3.53
	R^2	0.9957	0.9945	0.9906
Test	SD	0.053	0.074	0.100
	RMSE	1.628	1.769	2.268
	APRE	−1.14	−1.28	−1.43
	AAPRE	3.63	4.89	6.46
	R^2	0.9852	0.9827	0.9713
Total	SD	0.044	0.067	0.069
	RMSE	1.126	1.278	1.623
	APRE	−0.46	−1.32	−0.55
	AAPRE	2.71	3.52	4.12
	R^2	0.9939	0.9925	0.9873

Table 4. Statistical assessment of the developed models.

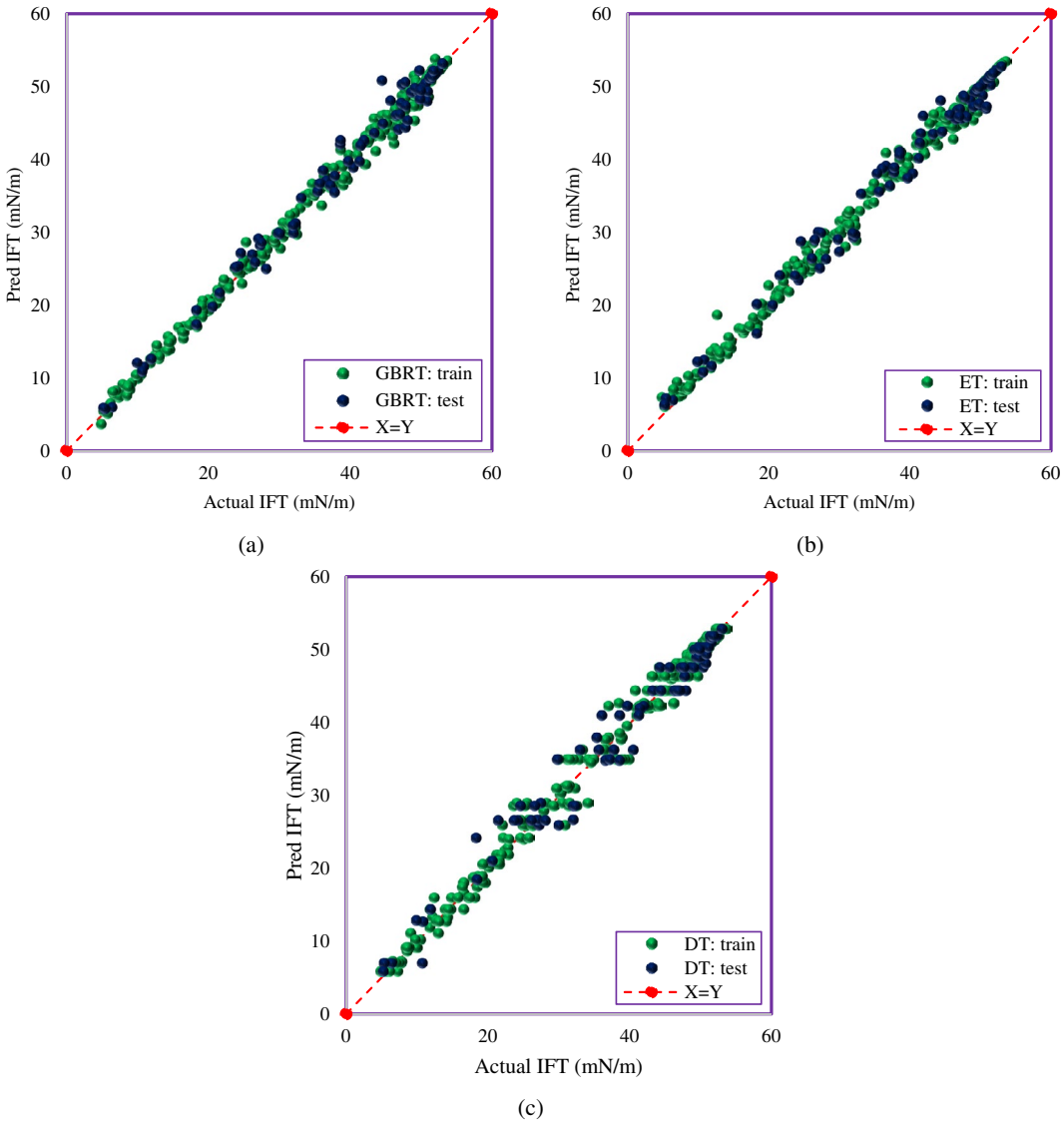


Figure 5. Cross plots of the developed models; (**a**) GBRT, (**b**) ET, (**c**) DT.

the lower accuracy of these models. Figure 5 shows that the points of the GBRT model are scattered around the unit slope line. This model estimates the IFT close to the experimental values. It can be seen that the GBRT model estimates the IFT with higher accuracy than the DT and ET models.

Furthermore, the relative error diagram is a practical tool to show the deviation of the value predicted by the model from the experimental value. Absolute error is the difference between the predicted and experimental values. The relative error is equal to the absolute error divided by the experimental value. The relative error is calculated as follows:

$$\text{Relative Error} = \frac{\text{IFT}_{\text{exp}} - \text{IFT}_{\text{pred}}}{\text{IFT}_{\text{exp}}} \tag{18}$$

In Fig. 6, the zero line indicates that the model predicts without error. Therefore, if the training and test points are close to the zero line, it indicates that a robust model has been developed. Figure 6 shows that from a value of 20 mN/m onwards, the relative error of all models implemented in this study is low. The points in the negative range of Fig. 6 with respect to the relative error Eq. (18) show that the model is overestimated. As explained in the model section, ensemble methods increase accuracy and improve overfitting[87,96]. According to the results presented in Figs. 6 and 7, the models implemented in this study, including ET and GBRT, have reduced variance and controlled overfitting, as seen in the test phase.

Next, Fig. 8 illustrates the cumulative frequency plot, which displays the proportion of predicted data that are less than or equal to a particular error value. This graph displays absolute relative errors (%) that are calculated using the next equation for different proportions of predicted data by models.

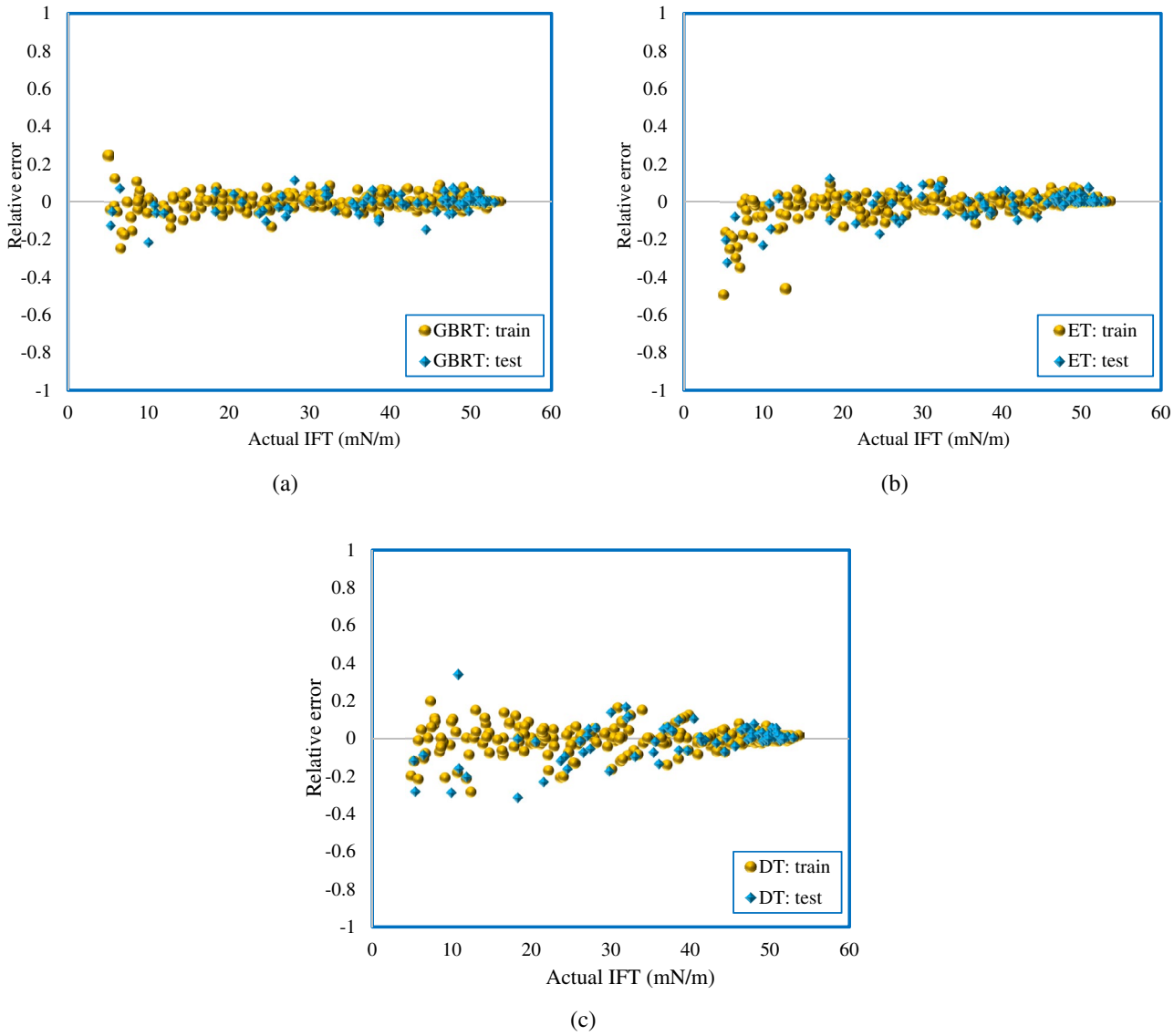


Figure 6. The relative deviation of the experimental results and predicted data using (**a**) GBRT, (**b**) ET, (**c**) DT models.

$$\text{Absolute Relative Error (\%)} = \left| \frac{\text{IFT}_{\text{exp}} - \text{IFT}_{\text{pred}}}{\text{IFT}_{\text{exp}}} \right| \times 100 \tag{19}$$

The closer a model gets to the vertical axis, the more data it can predict with lower error and consequently considered a more precise model. As Fig. 8 shows, the GBRT model estimates 70% of the IFT values with just less than 3.2% error. This is while the ET and DT models have predicted this proportion of the data with an error of 3.6% and 4.2%, respectively. In addition, about 90% of the data, estimated by the GBRT model, has an error lower than 6.2%, while it is 8.2% and 10.8% for ET and DT models, respectively. As a result, the superiority of the GBRT model over the ET and DT models can be recognized again.

Based on the presented results in this section, the GBRT model is proposed to estimate the IFT between the surfactant solution and the normal alkane with high accuracy. A part of IFT values predicted by GBRT model in the test phase is reported in Table 5, and no significant difference is observed in the prediction of experimental data by this model. The results of Fig. 7 show that the AAPRE and RMSE of the GBRT model in the test phase were 3.63% and 1.628, respectively, which indicates the high reliability of this model in predicting the IFT of surfactant–hydrocarbon systems.

Sensitivity analysis. In this study, the sensitivity analysis was performed to determine the extent and type of relationship between the independent variables presented in Table 2 and the amount of the IFT (output). Different methods of sensitivity analysis are introduced for regression models[97,98]. In this section, the Pearson coefficient was used to calculate the relevancy factor[99]:

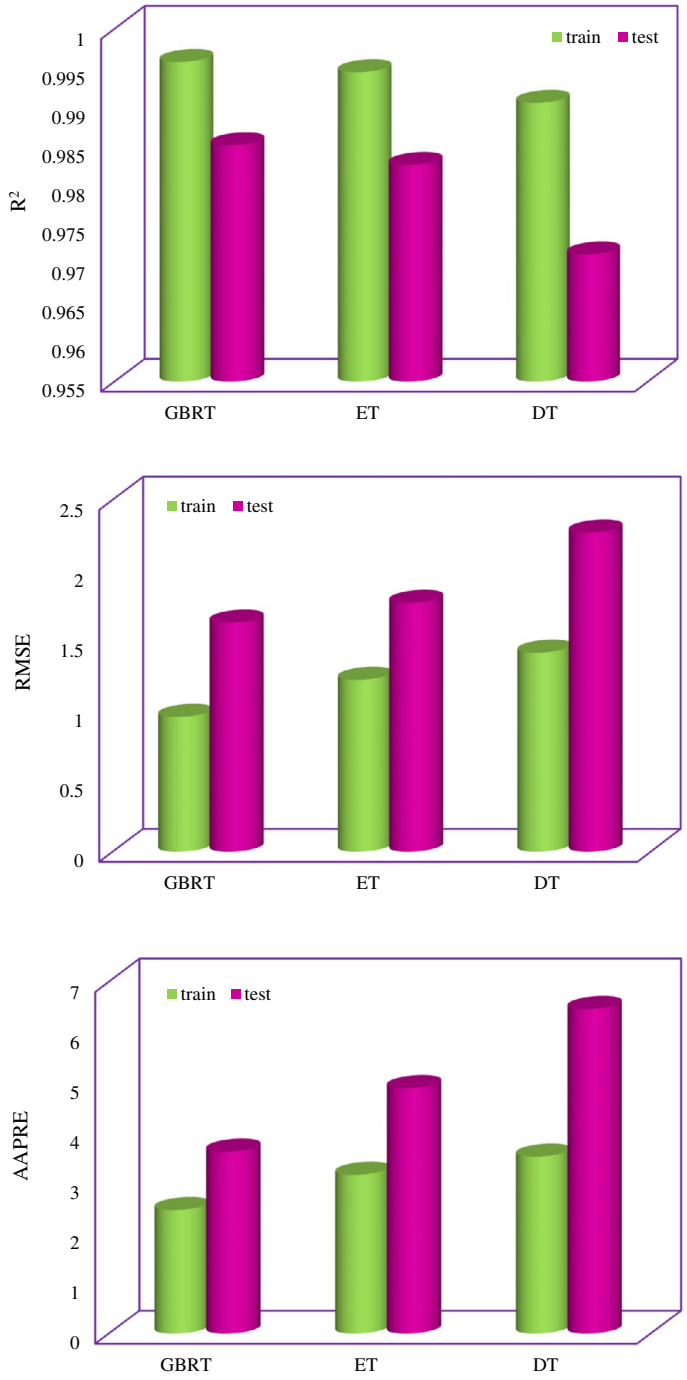


Figure 7. Statistical evaluation of the developed models.

$$RF = \frac{\sum_{i=1}^{n} (\overline{x_k} - x_{k.\,i})\left(\overline{y} - y_i\right)}{\sqrt{\sum_{i=1}^{n} (\overline{x_k} - x_{k.\,i})^2 \sum_{i=1}^{n} \left(\overline{y} - y_i\right)^2}} \tag{20}$$

where n and k represent the number of data points and the type of input variable. The symbols y_i, $\overline{y}$, $x_{k.i}$, and x_k denote the target, the average of the target value, the input value, and the k_{th} input value average. Figure 9 shows the absolute values of sensitivity analysis results of the proposed GBRT model. In this data set, PITx, the surfactant concentration, and HLB had the greatest effect on IFT, respectively. At concentrations lower than the CMC of surfactants, the IFT decreases with increasing surfactant concentration[100]. Therefore, the surfactant concentration was expected to have large effect on IFT along with the type of surfactants. The molecular weight of alkanes has the lowest value of the Pearson coefficient compared to other input parameters.

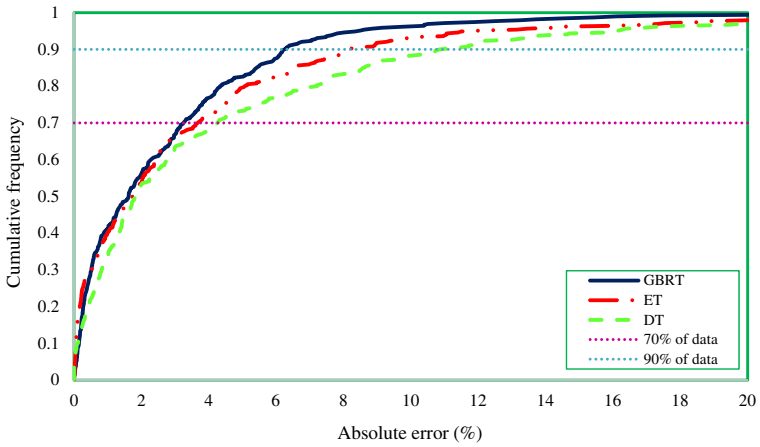


Figure 8. Cumulative frequency diagram of three proposed models for estimating the IFT of ionic surfactants and normal alkanes.

Trend analysis. In this section, the ability of the proposed GBRT model to predict IFT behavior in different conditions was investigated. Figure 10 shows the predicted values of IFT by the GBRT model and the experimental data[68, 69,72]. In this figure, the IFT of anionic (SDS) and cationic (C_{10}TAB) surfactants with n-hexane as hydrocarbon phase was plotted. The temperature of the surfactant solution was considered constant (298.2 K), and IFT data was plotted as a function of surfactant concentration. Based on the results presented in Fig. 10, it can be seen that the GBRT model accurately predicts the IFT of the surfactants and n-hexane systems. The IFT between the surfactant solution and the hydrocarbon depends on the surfactant concentration[101–103]. At concentrations lower than the CMC, by increasing the surfactant concentration, the IFT value will reduce[38]. Surfactant molecules are adsorbed on the liquid–liquid interface and reduce IFT[104]. Therefore, increasing the adsorbed surfactant molecules at the interfacial boundary further reduces the IFT.

In the next step, Fig. 11 shows the IFT between C_{10}TAB and C_{12}TAB in n-octane and n-nonane hydrocarbon phases[69] along with the prediction of the GBRT model. For this analysis, the temperature was considered a constant value of 298.2 K. As can be seen, the heavier the hydrocarbon phase, the greater the IFT of surfactant–hydrocarbon. Again, the GBRT model renders great predictions for the IFT of the surfactants and hydrocarbons considered in this figure.

Another parameter affecting the IFT of the surfactant that was considered in this study is temperature. The IFT values were predicted for different concentrations of SDS corresponding to a temperature range. Using the GBRT model, a comparison of the predicted trend of IFT changes in the temperature range with experimental data[70] is shown in Fig. 12. The predicted trend is similar to the experimental data and also shows good agreement with the real trend of IFT variation. As mentioned earlier, the IFT decreases with increasing temperature. It also shows that the proposed GBRT model predicts the interfacial behavior of the surfactant well under defined conditions.

Outlier detection using Leverage approach. The Leverage approach[105–107] is a reliable technique to discover outliers that may exist in a databank due to a variety of circumstances, including experimental errors. These points are located at an improper distance from the majority of data. Therefore, catching the inappropriate data noted above is critical for preventing model inaccuracy and unreliability. According to the Leverage method, the values of the standardized residuals (R) as well as a matrix named the Hat matrix, which is made up of the exploratory and predicted values obtained from the model, are needed to conduct this analysis. The leverage or Hat indexes (H) are determined using the following formula[108–110]:

$$H = X(X^T X)^{-1} X^T \tag{21}$$

Here, X represents the matrix of explanatory variables, and T represents the transpose matrix operator. Moreover, the critical Leverage (H^*) is calculated according to the following formula[111]:

$$H^* = \frac{3(number\ of\ inputs + 1)}{number\ of\ data\ points} \tag{22}$$

In this work, the databank includes four inputs and 390 data points, leading to $H^* = 0.0461$. On the other hand, considering MSE as the mean square of error, e_j as the error value of the jth data, and Hj as the jth Leverage value, R values can be determined as follows[90, 112]:

$$R_j = \frac{e_j}{[MSE(1 - H_j)]^{0.5}} \tag{23}$$

Number	HLB	dPIT/dx (°C)	Surfactant concentration × 10^5 (mol/l)	T (K)	Mw (g/mol)	IFT (mN/m)	GBRT	Relative error (%)
1	0	0	0	305.65	149.29	51.26	51.05222	0.41
2	40	499	286.472	298.15	128.2	24.2583	25.39824	−4.70
3	0	0	0	310.65	170.33	51.43	51.51684	−0.17
4	40	499	837.628	298.15	240.471	10.9684	11.62199	−5.96
5	0	0	0	305.65	86.18	49.7	49.50517	0.39
6	21	338	0.100012	298.15	100.21	50.2605	49.15622	2.20
7	19	486	69.96725	298.15	114.23	24.5995	27.15528	−10.39
8	19	486	713.515	298.15	198.39	18.3707	17.41083	5.23
9	21	338	306.963	298.15	170.33	36.6211	37.08537	−1.27
10	19	486	0.010001	298.15	128.2	49.6681	52.18449	−5.07
11	19	486	9.878343	298.15	100.21	40.4008	38.8913	3.74
12	18	453	0.013639	295.15	86.18	47.2076	50.27795	−6.50
13	40	499	34.9646	293.2	86.18	39.6676	39.73671	−0.17
14	19	486	0.078221	298.15	100.21	50.7415	47.95615	5.49
15	19	486	1.91964	298.15	142.29	41.5223	42.11541	−1.43
16	19	486	98.41191	298.15	128.2	27.249	28.22018	−3.56
17	21	338	10.50997	298.15	170.33	46.6652	44.09608	5.51
18	21	338	3.009322	298.15	128.2	44.3392	50.7895	−14.55
19	40	499	8	293.2	86.178	46.4	46.08001	0.69
20	21	338	6729.21	298.15	142.29	10.0208	12.15356	−21.28
21	19	486	74.11994	298.15	170.33	31.9658	29.92372	6.39
22	40	499	14	303.2	86.178	43.3	43.61458	−0.73
23	21	338	31.75242	298.15	170.33	41.8647	42.63048	−1.83
24	0	0	0	313.15	100.21	49.38	49.48231	−0.21
25	19	486	0.971186	298.15	114.23	47.7117	44.35099	7.04
26	21	338	102.32	298.15	86.18	36.0927	38.4876	−6.64
27	40	499	0	288.2	86.178	51.4	51.62253	−0.43
28	40	499	672.085	298.15	142.29	11.9565	12.68166	−6.07
29	0	0	0	310.65	114.23	50.09	49.88087	0.42
30	40	499	793.492	298.15	86.18	6.51163	6.071484	6.76
31	18	453	42.27935	295.15	86.18	18.3688	19.35079	−5.35
32	0	0	0	328.15	128.2	49.09	49.23298	−0.29
33	40	499	679.088	298.15	114.23	10.7708	10.98328	−1.97
34	21	338	1030.93	298.15	170.33	27.0355	29.14769	−7.81
35	40	499	96.628	298.15	86.18	27.4419	28.60547	−4.24
36	21	338	298.274	298.15	128.2	35.342	35.62043	−0.79
37	21	338	9.56027	298.15	142.29	47.9878	45.3272	5.54
38	40	499	169.492	298.15	114.23	29.8419	29.91685	−0.25
39	0	0	0	298.15	156.31	52.25	52.14065	0.21
40	40	499	83.55376	298.15	114.23	37.1542	36.40291	2.02

Table 5. The IFT data predicted by GBRT models in the test phase.

If the most of the data points are positioned in the ranges of $-3 \le R \le 3$ and $0 \le H_j \le H^*$, both the experimental data and the model's estimations will be statistically trustworthy and precise[113]. William's plot shows R values versus H values to identify outliers. Figure 13 displays the outcomes of the leverage approach utilizing the GBRT model's results. In this case, only 6 points are recognized as suspected data, which are located outside of the model application scope. Also, it was found just 8 points as outliers. This confirms that the experimental database of IFT between surfactants and hydrocarbon is highly reliable, and the GBRT model is statistically dependable and valid.

Conclusions

The aim of this study was to develop accurate and reliable models to estimate the IFT of ionic surfactants and normal alkanes. In this study, three intelligent computer-aided algorithms namely DT, ET, and GBRT models were implemented for this purpose. A databank containing 390 experimental IFT data points presented in the literature was used to develop these models. The models considered the following parameters as input: temperature, normal alkane molecular weight, surfactant concentration, HLB, and PIT. Based on this work, the following conclusions are drawn:

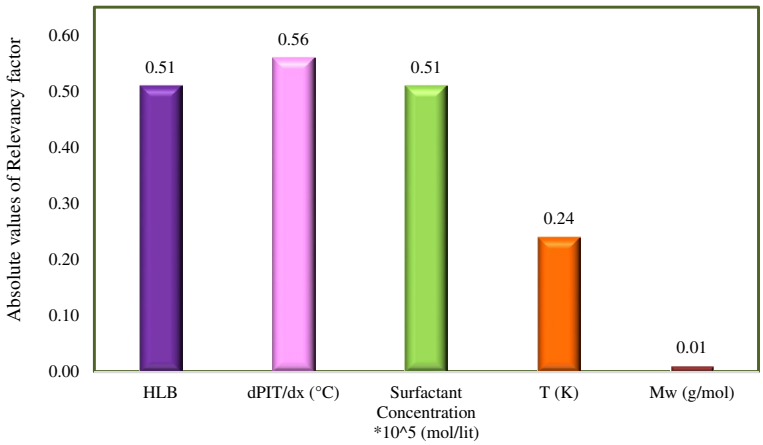


Figure 9. The absolute effect of each input parameter on the IFT of surfactant–hydrocarbon systems based on Pearson correlation coefficient.

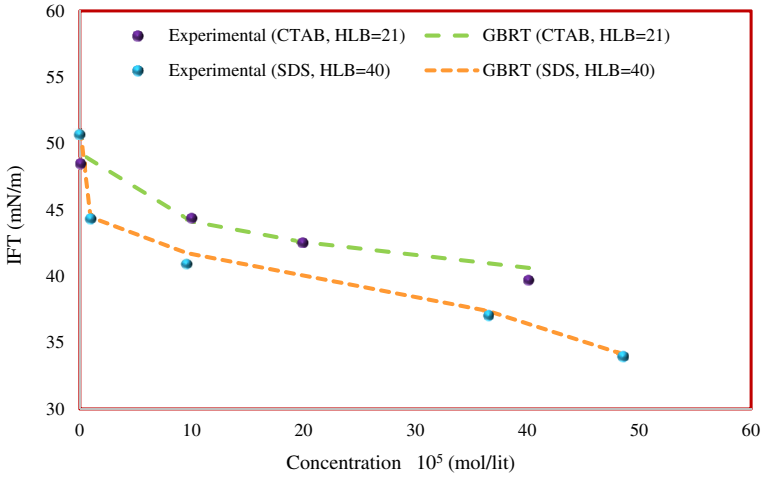
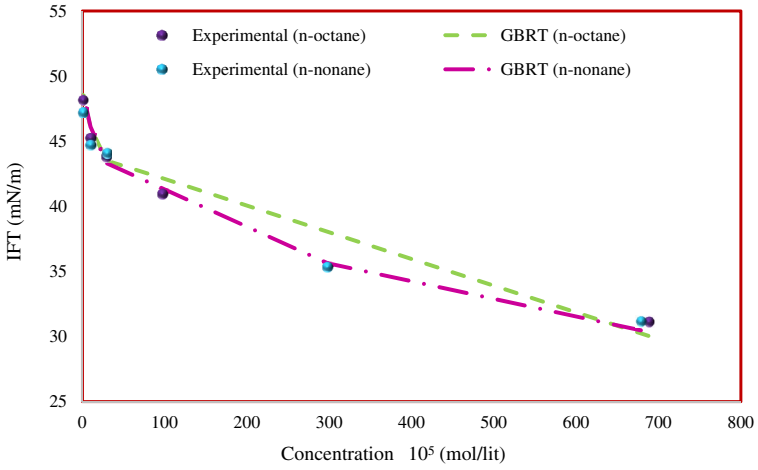


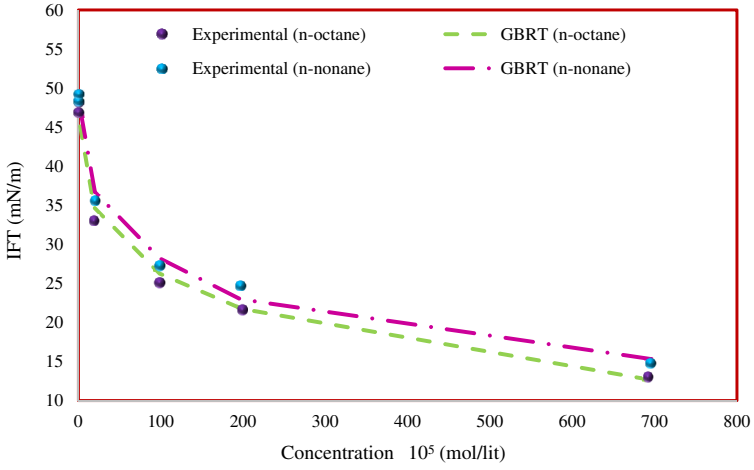
Figure 10. Comparison of predicted and experimental IFT trends for the systems of *n*-hexane and surfactants C_{10}TAB and SDS at 298.2 K.

1. The ensemble methods implemented in this study, GBRT and ET, were able to reduce the variance of the DT model.
2. Among all the models developed in this study, the GBRT was the best model for predicting the IFT between the surfactant and normal alkanes.
3. Statistical evaluation in the test phase showed that the AAPRE% and RMSE of the GBRT model are 3.63% and 1.628, respectively.
4. The trend analysis demonstrated that the predictions of the GBRT model follow the expected variations in terms of the independent variables.
5. The cumulative error distribution of the GBRT model was very satisfactory, with approximately 90% of the predicted data having a relative error of less than 6.2%.
6. According to the results of the sensitivity analysis, the effect of input parameters on the IFT is as follows: PIT > surfactant concentration > HLB > temperature > molecular weight of normal alkane.
7. The Leverage method demonstrated that the majority of the data points (almost 96.5%) are valid and both the IFT databank and the GBRT model seem to be highly trustworthy.

As a suggestion for future studies, it can be mentioned that the simultaneous involvement of the aqueous phase containing surfactant, the organic hydrocarbon phase and the solid phase including reservoir rock or its minerals can correlate the governing equations in the area of surface wetting.



(a)



(b)

Figure 11. Comparison of predicted and experimental IFT trends for pure hydrocarbons at 298.2 K; (**a**) C_{10}TAB and (**b**) C_{12}TAB.

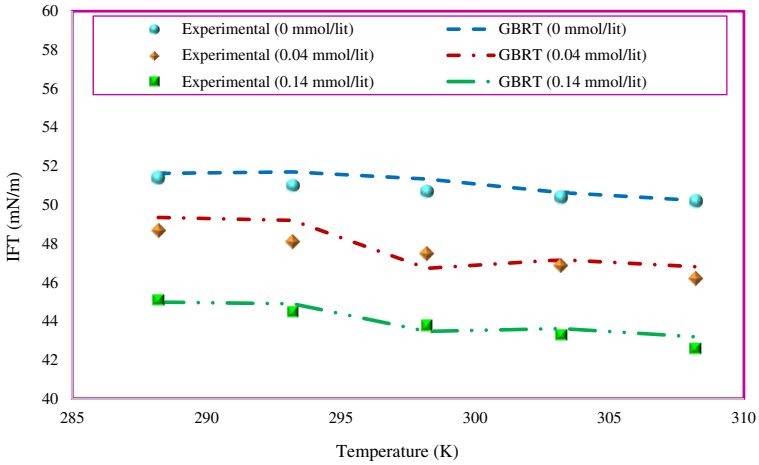


Figure 12. Comparison between experimental and predicted IFT values by the GBRT model for SDS surfactant in n-hexane mixture at different temperatures and concentrations.

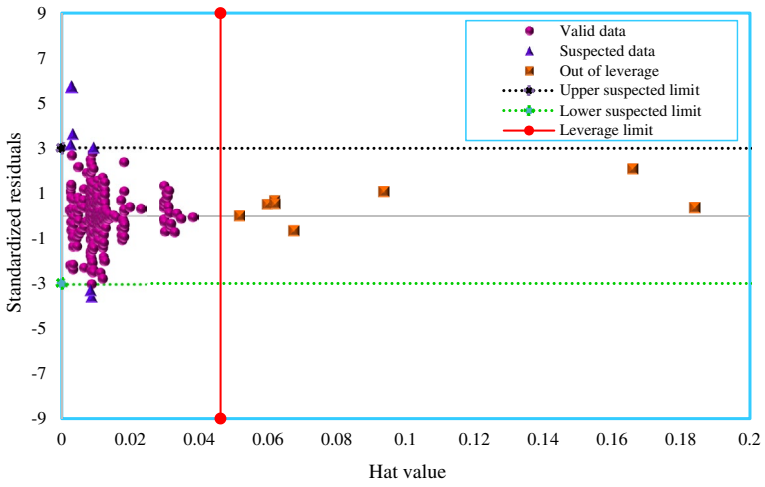


Figure 13. William's plot of the developed GBRT model.

Data availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Received: 29 August 2022; Accepted: 29 June 2023
Published online: 05 July 2023

References

1. Sagir, M., Mushtaq, M., Tahir, M. S., Tahir, M. B. & Shaik, A. R. *Surfactant in Petroleum Industry BT—Surfactants for Enhanced Oil Recovery Applications*. (eds. Sagir, M., Mushtaq, M., Tahir, M. S., Tahir, M. B. & Shaik, A. R.).13–40 (Springer, 2020). https://doi.org/10.1007/978-3-030-18785-9_2.
2. Hirasaki, G. J., Miller, C. A. & Puerto, M. Recent advances in surfactant EOR. *SPE J.* **16**, 889–907 (2011).
3. Puerto, M., Hirasaki, G. J., Miller, C. A. & Barnes, J. R. Surfactant systems for EOR in high-temperature, high-salinity environments. *SPE J. (Society of Petroleum Engineers)* **17**, 11–19 (2012).
4. Aranda-Bravo, C. G., Romero-Martínez, A., Trejo, A. & Águila-Hernández, J. Interfacial tension and density of water + branched hydrocarbon binary systems in the range 303–343 K. *Ind. Eng. Chem. Res.* **48**, 1476–1483 (2009).
5. Bui, B. T. & Tutuncu, A. N. Interfacial tension induced-transport in shale: A pore-scale study. *J. Pet. Sci. Eng.* **171**, 1409–1419 (2018).
6. Kim, H. & Burgess, D. J. Prediction of interfacial tension between oil mixtures and water. *J. Colloid Interface Sci.* **241**, 509–513 (2001).
7. Reis, P. K. P. & Carvalho, M. S. Pore-scale compositional modeling of gas-condensate flow: Effects of interfacial tension and flow velocity on relative permeability. *J. Pet. Sci. Eng.* **202**, 108454 (2021).
8. Iyi, D., Balogun, Y., Oyeneyin, B. & Faisal, N. Numerical modelling of the effect of wettability, interfacial tension and temperature on oil recovery at pore-scale level. *J. Pet. Sci. Eng.* **201**, 108453 (2021).
9. Fulcher, R. A. Jr., Ertekin, T. & Stahl, C. D. Effect of capillary number and its constituents on two-phase relative permeability curves. *J. Pet. Technol.* **37**, 249–260 (1985).
10. Tang, J. S. Interwell tracer tests to determine residual oil saturation to waterflood at Judy Creek Bhl'a'pool. *J. Can. Pet. Technol.* **31**, 133 (1992).
11. Chatzis, I. & Morrow, N. R. Correlation of capillary number relationships for sandstone. *Soc. Pet. Eng. J.* **24**, 555–562 (1984).
12. Chatzis, I., Kuntamukkula, M. S. & Morrow, N. R. Effect of capillary number on the microstructure of residual oil in strongly water-wet sandstones. *SPE Reserv. Eng.* **3**, 902–912 (1988).
13. Garnes, J. M., Mathisen, A. M., Scheie, A. & Skauge, A. Capillary number relations for some north, sea reservoir sandstones. in *SPE/DOE Enhanced Oil Recovery Symposium SPE-20264-MS*. https://doi.org/10.2118/20264-MS (1990).
14. Johannesen, E. B. & Graue, A. Mobilization of remaining oil—Emphasis on capillary number and wettability. in *International Oil Conference and Exhibition in Mexico SPE-108724-MS*. https://doi.org/10.2118/108724-MS (2007).
15. Chukwudeme, E. A., Fjelde, I., Abeysinghe, K. & Lohne, A. Effect of interfacial tension on water/oil relative permeability on the basis of history matching to coreflood data. *SPE Reserv. Eval. Eng.* **17**, 37–48 (2014).
16. Guo, H. *et al.* Review of capillary number in chemical enhanced oil recovery. in *Society of Petroleum Engineers—SPE Kuwait Oil Gas Show Conference* (2015). https://doi.org/10.2118/175172-ms.
17. Delshad, M., Najafabadi, N. F., Anderson, G. A., Pope, G. A. & Sepehrnoori, K. Modeling wettability alteration by surfactants in naturally fractured reservoirs. *SPE Reserv. Eval. Eng.* **12**, 361–370 (2009).
18. Lohne, A. & Fjelde, I. Surfactant flooding in heterogeneous formations. in *SPE Improved Oil Recovery Symposium SPE-154178-MS*. https://doi.org/10.2118/154178-MS (2012).
19. Arab, D., Kantzas, A. & Bryant, S. L. Water flooding of oil reservoirs: Effect of oil viscosity and injection velocity on the interplay between capillary and viscous forces. *J. Pet. Sci. Eng.* **186**, 106691 (2020).
20. Dang, C. *et al.* Modeling and optimization of alkaline-surfactant-polymer flooding and hybrid enhanced oil recovery processes. *J. Pet. Sci. Eng.* **169**, 578–601 (2018).
21. Nowrouzi, I., Mohammadi, A. H. & Manshad, A. K. Water-oil interfacial tension (IFT) reduction and wettability alteration in surfactant flooding process using extracted saponin from Anabasis Setifera plant. *J. Pet. Sci. Eng.* **189**, 106901 (2020).
22. Halliday, H. L. Surfactants: Past, present and future. *J. Perinatol.* **28**, S47–S56 (2008).
23. Beverung, C. J., Radke, C. J. & Blanch, H. W. Protein adsorption at the oil/water interface: Characterization of adsorption kinetics by dynamic interfacial tension measurements. *Biophys. Chem.* **81**, 59–80 (1999).

24. Campanelli, J. R. & Wang, X. Dynamic interfacial tension of surfactant mixtures at liquid–liquid interfaces. *J. Colloid Interface Sci.* **213**, 340–351 (1999).
25. Li, S., Liu, J., Hou, J. & Zhang, G. Meniscus-induced motion of oil droplets. *Colloids Surf. A Physicochem. Eng. Asp.* **469**, 252–255 (2015).
26. Liu, J., Li, S. & Hou, J. Near-post meniscus-induced migration and assembly of bubbles. *Soft Matter* **12**, 2221–2230 (2016).
27. Li, Z. *et al.* Ultra-low interfacial tension biobased and catanionic surfactants for low permeability reservoirs. *J. Mol. Liq.* **309**, 113099 (2020).
28. Zhou, H. *et al.* Systematic investigation of ionic liquid-type gemini surfactants and their abnormal salt effects on the interfacial tension of a water/model oil system. *J. Mol. Liq.* **249**, 33–39 (2018).
29. Zhang, L. *et al.* Effect of different acidic fractions in crude oil on dynamic interfacial tensions in surfactant/alkali/model oil systems. *J. Pet. Sci. Eng.* **41**, 189–198 (2004).
30. Xu, J. *et al.* Effect of surfactant headgroups on the oil/water interface: An interfacial tension measurement and simulation study. *J. Mol. Struct.* **1052**, 50–56 (2013).
31. Moradi, S., Isari, A. A., Bachari, Z. & Mahmoodi, H. Combination of a new natural surfactant and smart water injection for enhanced oil recovery in carbonate rock: Synergic impacts of active ions and natural surfactant concentration. *J. Pet. Sci. Eng.* **176**, 1–10 (2019).
32. Mosayebi, A., Angaji, M. T. & Khadiv-Parsi, P. The effect of temperature on the interfacial tension between crude oil and ethoxylated nonylphenols. *Pet. Sci. Technol.* **34**, 1315–1322 (2016).
33. Hjelmeland, O. S. & Larrondo, L. E. Experimental investigation of the effects of temperature, pressure, and crude oil composition on interfacial properties. *SPE Reserv. Eng.* **1**, 321–328 (1986).
34. Farhadi, H., Ayatollahi, S. & Fatemi, M. The effect of brine salinity and oil components on dynamic IFT behavior of oil-brine during low salinity water flooding: Diffusion coefficient, EDL establishment time, and IFT reduction rate. *J. Pet. Sci. Eng.* **196**, 107862 (2021).
35. Negin, C., Ali, S. & Xie, Q. Most common surfactants employed in chemical enhanced oil recovery. *Petroleum* **3**, 197–211 (2017).
36. Sheng, S. S. *et al.* Structure-activity relationship of anionic–nonionic surfactant for reducing interfacial tension of crude oil. *J. Mol. Liq.* **313**, 112772 (2020).
37. Strey, R. Phase behavior and interfacial curvature in water–oil–surfactant systems. *Curr. Opin. Colloid Interface Sci.* **1**, 402–410 (1996).
38. Kamal, M. S., Hussein, I. A. & Sultan, A. S. Review on surfactant flooding: Phase behavior, retention, IFT, and field applications. *Energy Fuels* **31**, 7701–7720 (2017).
39. Lee, B. B., Ravindra, P. & Chan, E. S. A critical review: Surface and interfacial tension measurement by the drop weight method. *Chem. Eng. Commun.* **195**, 889–924. https://doi.org/10.1080/00986440801905056 (2008).
40. Yildirim, O. E., Xu, Q. & Basaran, O. A. Analysis of the drop weight method. *Phys. Fluids* **17**, 062107 (2005).
41. Berry, J. D., Neeson, M. J., Dagastine, R. R., Chan, D. Y. C. & Tabor, R. F. Measurement of surface and interfacial tension using pendant drop tensiometry. *J. Colloid Interface Sci.* **454**, 226–237 (2015).
42. Touhami, Y., Neale, G. H., Hornof, V. & Khalfalah, H. A modified pendant drop method for transient and dynamic interfacial tension measurement. *Colloids Surf. A Physicochem. Eng. Asp.* **112**, 31–41 (1996).
43. Garandet, J. P., Vinet, B. & Gros, P. Considerations on the pendant drop method: A new look at Tate's law and Harkins' correction factor. *J. Colloid Interface Sci.* **165**, 351–354 (1994).
44. Viades-Trejo, J. & Gracia-Fadrique, J. Spinning drop method: From Young–Laplace to Vonnegut. *Colloids Surf. A Physicochem. Eng. Asp.* **302**, 549–552 (2007).
45. Joseph, D. D. *et al.* A spinning drop tensioextensometer. *J. Rheol. (N. Y. N. Y)* **36**, 621 (1998).
46. Cayias, J. L., Schechter, R. S. & Wade, W. H. Measurement of low interfacial tension via the spinning drop technique. *ACS Symp. Ser.* https://doi.org/10.1021/BK-1975-0008.CH017 (1974).
47. Fainerman, V. B., Zholob, S. A., Lucassen-Reynders, E. H. & Miller, R. Comparison of various models describing the adsorption of surfactant molecules capable of interfacial reorientation. *J. Colloid Interface Sci.* **261**, 180–183 (2003).
48. Bahramian, A. & Zarbakhsh, A. Interfacial equation of state for ionized surfactants at oil/water interfaces. *Soft Matter* **11**, 6482–6491 (2015).
49. Kairaliyeva, T. *et al.* Surface tension and adsorption studies by drop profile analysis tensiometry. *J. Surfactants Deterg.* **20**, 1225–1241 (2017).
50. Ross, S. & Morrison, I. D. On the alleged ideality of Szyszkowski–Langmuir adsorption. *J. Colloid Interface Sci.* **91**, 244–247 (1983).
51. Rusanov, A. I. On the thermodynamics of thin films. The Frumkin equation. *Colloid J.* **81**, 741–746 (2019).
52. Markin, V. S., Volkova-Gugeshashvili, M. I. & Volkov, A. G. Adsorption at liquid interfaces: The generalized Langmuir isotherm and interfacial structure. *J. Phys. Chem. B* **110**, 11415–11420 (2006).
53. Mulqueen, M. & Blankschtein, D. Theoretical and experimental investigation of the equilibrium oil–water interfacial tensions of solutions containing surfactant mixtures. *Langmuir* **18**, 365–376 (2002).
54. Mulqueen, M. & Blankschtein, D. Theoretical and experimental investigation of the equilibrium oil–water interfacial tensions of solutions containing surfactant mixtures. *Langmuir* **18**, 365–376 (2002).
55. Nikseresht, S., Riazi, M., Amani, M. J. & Farshchi Tabrizi, F. Prediction of oil/water interfacial tension containing ionic surfactants. *Colloids Interface Sci. Commun.* **34**, 100217 (2020).
56. Chen, H. *et al.* A new prediction model of CO_2 diffusion coefficient in crude oil under reservoir conditions based on BP neural network. *Energy* **239**, 122286 (2022).
57. Zhang, L. *et al.* Prediction of coal self-ignition tendency using machine learning. *Fuel* **325**, 124832 (2022).
58. Tabasi, S. *et al.* Optimized machine learning models for natural fractures prediction using conventional well logs. *Fuel* **326**, 124952 (2022).
59. Ameli, F., Hemmati-Sarapardeh, A., Schaffie, M., Husein, M. M. & Shamshirband, S. Modeling interfacial tension in N2/n-alkane systems using corresponding state theory: Application to gas injection processes. *Fuel* **222**, 779–791 (2018).
60. Ameli, F., Hemmati-Sarapardeh, A., Tatar, A., Zanganeh, A. & Ayatollahi, S. Modeling interfacial tension of normal alkane-supercritical CO_2 systems: Application to gas injection processes. *Fuel* **253**, 1436–1445 (2019).
61. Amooie, M. A. *et al.* Data-driven modeling of interfacial tension in impure CO_2-brine systems with implications for geological carbon storage. *Int. J. Greenh. Gas Control* **90**, 102811 (2019).
62. Mehrjoo, H., Riazi, M., Nait Amar, M. & Hemmati-Sarapardeh, A. Modeling interfacial tension of methane-brine systems at high pressure and high salinity conditions. *J. Taiwan Inst. Chem. Eng.* **114**, 125–141 (2020).
63. Nait Amar, M., Shateri, M., Hemmati-Sarapardeh, A. & Alamatsaz, A. Modeling oil-brine interfacial tension at high pressure and high salinity conditions. *J. Pet. Sci. Eng.* **183**, 106413 (2019).
64. Rostami, A., Ebadi, H., Arabloo, M., Meybodi, M. K. & Bahadori, A. Toward genetic programming (GP) approach for estimation of hydrocarbon/water interfacial tension. *J. Mol. Liq.* **230**, 175–189 (2017).
65. Rouhibakhsh, K. & Darvish, H. Utilization of fuzzy C-means algorithm as a novel predictive tool for estimation of interfacial tension of hydrocarbon and brine. *Pet. Sci. Technol.* **36**, 1107–1112 (2018).

66. Kiomarsiyan, A. & Esfandiarian, A. Applying grid partitioning based fuzzy inference system method to estimate interfacial tension of brine and hydrocarbon. *Pet. Sci. Technol.* **37**, 1620–1625 (2019).
67. Abooali, D., Sobati, M. A., Shahhosseini, S. & Assareh, M. A new empirical model for estimation of crude oil/brine interfacial tension using genetic programming approach. *J. Pet. Sci. Eng.* **173**, 187–196 (2019).
68. Pradines, V. *et al.* Adsorption of alkyl trimethylammonium bromides at the water/air and water/hexane interfaces. *Colloids Surf. A Physicochem. Eng. Asp.* **371**, 22–28 (2010).
69. Mucic, N., Kovalchuk, N. M., Aksenenko, E. V., Fainerman, V. B. & Miller, R. Adsorption layer properties of alkyltrimethylammonium bromides at interfaces between water and different alkanes. *J. Colloid Interface Sci.* **410**, 181–187 (2013).
70. Saien, J., Rezvani Pour, A. & Asadabadi, S. Interfacial tension of the *n*-hexane–water system under the influence of magnetite nanoparticles and sodium dodecyl sulfate assembly at different temperatures. *J. Chem. Eng. Data* **59**, 1835–1842 (2014).
71. Fainerman, V. B. *et al.* Particular behavior of surface tension at the interface between aqueous solution of surfactant and alkane. *Langmuir* **35**, 15214–15220 (2019).
72. Biswal, N. R., Rangera, N. & Singh, J. K. Effect of different surfactants on the interfacial behavior of the *n*-hexane–water system in the presence of silica nanoparticles. *J. Phys. Chem. B* **120**, 7265–7274 (2016).
73. Zeppieri, S., Rodríguez, J. & López De Ramos, A. L. Interfacial tension of alkane + water systems. *J. Chem. Eng. Data* **46**, 1086–1088 (2001).
74. Rehfeld, S. J. Adsorption of sodium dodecyl sulfate at various hydrocarbon–water interfaces. *J. Phys. Chem.* **71**, 738–745 (1967).
75. Saien, J. & Bahrami, M. Understanding the effect of different size silica nanoparticles and SDS surfactant mixtures on interfacial tension of *n*-hexane–water. *J. Mol. Liq.* **224**, 158–164 (2016).
76. Ontiveros, J. F. *et al.* Structure–interfacial properties relationship and quantification of the amphiphilicity of well-defined ionic and non-ionic surfactants using the PIT-slope method. *J. Colloid Interface Sci.* **448**, 222–230 (2015).
77. Kondo, S. *et al.* Effect of the hydrophilic–lipophilic balance (HLB) of surfactants included in the post-CMP cleaning chemicals on porous SiOC direct CMP. in *2007 IEEE International Interconnect Technology Conference*. 172–174 (2007). https://doi.org/10.1109/IITC.2007.382381.
78. Reham, S. S. *et al.* Study on stability, fuel properties, engine combustion, performance and emission characteristics of biofuel emulsion. *Renew. Sustain. Energy Rev.* **52**, 1566–1579 (2015).
79. Casford, M. T. L., Davies, P. B. & Neivandt, D. J. Adsorption of sodium dodecyl sulfate at the hydrophobic solid/aqueous solution interface in the presence of poly(ethylene glycol): Dependence upon polymer molecular weight. *Langmuir* **22**, 3105–3111 (2006).
80. Davies, J. T. A quantitative kinetic theory of emulsion type, I. Physical chemistry of the emulsifying agent. in *Gas/Liquid and Liquid/Liquid Interface. Proceedings of the International Congress of Surface Activity*. Vol. 42. 6–438 (1957).
81. Loh, W.-Y. Classification and regression trees. *WIREs Data Min. Knowl. Discov.* **1**, 14–23 (2011).
82. Wu, M. *et al.* Beyond sparsity: Tree regularization of deep models for interpretability. *Proc. AAAI Conf. Artif. Intell.* **32**, 1670–1678 (2018).
83. Bibal, A. & Frénay, B. *Interpretability of Machine Learning Models and Representations: An Introduction Interpretability and Explanations of Nonlinear Dimensionality Reduction Mappings View Project Machine Learning and Formal Verification View Project Interpretability of Machin*e (2016).
84. Rokach, L. & Maimon, O. *Decision Trees BT—Data Mining and Knowledge Discovery Handbook* (eds. Maimon, O. & Rokach, L.). 165–192 (Springer, 2005). https://doi.org/10.1007/0-387-25465-X_9.
85. Quinlan, J. R. *Bagging, Boosting, and C4.5*.
86. Khoshgoftaar, T. M. & Allen, E. B. Controlling overfitting in classification—Tree models of software quality. *Empir. Softw. Eng.* **6**, 59–79 (2001).
87. Zhou, Z.-H., Wu, J. & Tang, W. Ensembling neural networks: Many could be better than all. *Artif. Intell.* **137**, 239–263 (2002).
88. Opitz, D. & Maclin, R. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999).
89. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
90. Zheng, H., Mahmoudzadeh, A., Amiri-Ramsheh, B. & Hemmati-Sarapardeh, A. Modeling viscosity of CO_2–N_2 gaseous mixtures using robust tree-based techniques: Extra tree, random forest, GBoost, and LightGBM. *ACS Omega* (2023).
91. Schapire, R. E. The strength of weak learnability. *Mach. Learn.* **5**, 197–227 (1990).
92. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
93. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **23**, 1189–1232 (2001).
94. Utkin, L. V. & Wiencierz, A. Improving over-fitting in ensemble regression by imprecise probabilities. *Inf. Sci. (NY)* **317**, 315–328 (2015).
95. Hagan, M. T. & Menhaj, M. B. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Netw.* **5**, 989–993 (1994).
96. Bauer, E. & Kohavi, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn.* **36**, 105–139 (1999).
97. Frey, H. C. & Patil, S. R. Identification and review of sensitivity analysis methods. *Risk Anal.* **22**, 553–578 (2002).
98. Castillo, E., Hadi, A. S., Conejo, A. & Fernández-Canteli, A. A general method for local sensitivity analysis with application to regression models and other optimization problems. *Technometrics* **46**, 430–444 (2004).
99. Benesty, J., Chen, J., Huang, Y. & Cohen, I. *Pearson Correlation Coefficient BT—Noise Reduction in Speech Processing*. (eds. Cohen, I., Huang, Y., Chen, J. & Benesty, J.). 1–4 (Springer, 2009). https://doi.org/10.1007/978-3-642-00296-0_5.
100. Chakraborty, T., Chakraborty, I. & Ghosh, S. The methods of determination of critical micellar concentrations of the amphiphilic systems in aqueous medium. *Arab. J. Chem.* **4**, 265–270 (2011).
101. Al-Sahhaf, T., Elkamel, A., Ahmed, A. S. & Khan, A. R. The influence of temperature, pressure, salinity, and surfactant concentration on the interfacial tension of the *n*-octane–water system. *Chem. Eng. Commun.* **192**, 667–684 (2005).
102. Akhlaghi, N., Riahi, S. & Parvaneh, R. Interfacial tension behavior of a nonionic surfactant in oil/water system; salinity, pH, temperature, and ionic strength effects. *J. Pet. Sci. Eng.* **198**, 108177 (2021).
103. Karnanda, W., Benzagouta, M. S., AlQuraishi, A. & Amro, M. M. Effect of temperature, pressure, salinity, and surfactant concentration on IFT for surfactant flooding optimization. *Arab. J. Geosci.* **6**, 3535–3544 (2013).
104. Li, Y. *et al.* Mesoscopic simulation study on the efficiency of surfactants adsorbed at the liquid/liquid interface. *Mol. Simul.* **31**, 1027–1033 (2005).
105. Rousseeuw, P.J. & Leroy, A.M. *Robust Regression and Outlier Detection* (Google Books).
106. Goodall, C. R. 13 Computation using the QR decomposition. *Handb. Stat.* **9**, 467–508 (1993).
107. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **26**, 694–701 (2007).
108. Amiri-Ramsheh, B., Safaei-Farouji, M., Larestani, A., Zabihi, R. & Hemmati-Sarapardeh, A. Modeling of wax disappearance temperature (WDT) using soft computing approaches: Tree-based models and hybrid models. *J. Pet. Sci. Eng.* **208**, 109774 (2022).
109. Amiri-Ramsheh, B., Zabihi, R. & Hemmati-Sarapardeh, A. Modeling wax deposition of crude oils using cascade forward and generalized regression neural networks: Application to crude oil production. *Geoenergy Sci. Eng.* **1**, 211613 (2023).
110. Mohammadi, M. R. *et al.* Modeling hydrogen solubility in alcohols using machine learning models and equations of state. *J. Mol. Liq.* **346**, 117807 (2022).

111. Ansari, S. *et al.* Prediction of hydrogen solubility in aqueous solutions: Comparison of equations of state and advanced machine learning-metaheuristic approaches. *Int. J. Hydrogen Energy* **47**, 37724–37741 (2022).
112. Mohammadi, M.-R. *et al.* Toward predicting SO_2 solubility in ionic liquids utilizing soft computing approaches and equations of state. *J. Taiwan Inst. Chem. Eng.* **133**, 104220 (2022).
113. Nakhaei-Kohani, R., Taslimi-Renani, E., Hadavimoghaddam, F., Mohammadi, M.-R. & Hemmati-Sarapardeh, A. Modeling solubility of CO_2–N_2 gas mixtures in aqueous electrolyte systems using artificial intelligence techniques and equations of state. *Sci. Rep.* **12**, 1–23 (2022).

Author contributions

A.R.-K.: Investigation, Methodology, Data curation, Writing-Original Draft, E.R.-K.: Writing-Original Draft, Visualization, B.A.-R.: Validation, Modeling, Writing-Review & Editing, M.-R.M.: Writing-Original Draft, Validation, Modeling, A.H.-S.: Methodology, Validation, Supervision, Writing-Review & Editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.H.-S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.