Geo-Energy Research
Vol. 4, No. 4, p. 435-447, 2020

Original article

Modeling viscosity of crude oil using k-nearest neighbor algorithm

Mohammad Reza Mahdiani[1], Ehsan Khamehchi[1], Sassan Hajirezaie[2], Abdolhossein Hemmati – Sarapardeh[3]*

[1]*Department of Petroleum Engineering, Amirkabir University of Technology, Tehran, Iran*
[2]*Department of Civil and Environmental Engineering, Princeton University, NJ 08540, United States*
[3]*Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran*

Keywords:
Oil viscosity
machine learning
k-nearest neighbor
genetic programming
linear discriminant analysis

Cited as:
Mahdiani, M.R., Khamehchi, E., Hajirezaie, S., Hemmati-Sarapardeh, A. Modeling viscosity of crude oil using k-nearest neighbor algorithm. *Advances in Geo-Energy Research*, 2020, 4(4): 435-447, doi: 10.46690/ager.2020.04.08.

Abstract:
Oil viscosity is an important factor in every project of the petroleum industry. These processes can range from gas injection to oil reservoirs to comprehensive reservoir simulation studies. Different experimental approaches have been proposed for measuring oil viscosity. However, these methods are often time taking, cumbersome and at some physical conditions, impossible. Therefore, development of predictive models for estimating this parameter is crucial. In this study, three new machine learning based models are developed to estimate the oil viscosity. These approaches are genetic programing, k-nearest neighbor (KNN) and linear discriminant analysis. Oil gravity and temperature were the input parameters of the models. Various graphical and statistical error analyses were used to measure the performance of the developed models. Also, comparison study between the developed models and the well-known previously published models was conducted. Moreover, trend analysis was performed to compare the predictions of the models with the trend of experimental data. The results indicated that the developed models outperform all of the previously published models by showing negligible prediction errors. Among the developed models, the KNN model has the highest accuracy by showing an overall mean absolute error of 8.54%. The results show that the new developed models in this study can be potentially utilized in reservoir simulation packages of the petroleum industry.

1. Introduction

Oil consumption has significantly increased over the past few decades as a result of industrialization. In addition, the population growth and their need to energy is another reason for the increase in oil consumption. PVT properties such as viscosity, oil formation volume factor, and specific gravities of oil and gas have an important role in different aspects of petroleum engineering and fluid flow modeling (Khamehchi et al., 2019, 2020). One of the main aspects of crude oil is oil viscosity. Oil viscosity is a measure for the internal friction factor of the fluid flow (Zhang et al., 2017). The accurate value of viscosity is needed in various analyses in petroleum engineering including studying the fluid properties (Ilieva et al., 2016; Kleinhans et al., 2016), fluid flow (Abubakar et al., 2015; Al-Sarkhi et al., 2016; Norouzi et al., 2017; Zhang et al., 2017), mixing properties (Wen et al., 2016) and asphaltene precipitation (Ilyin et al., 2016).

One method for estimating oil viscosity is conducting experimental measurements, which is the most accurate approach. However, experiments are usually time consuming, expensive and not applicable in some cases such as simulation studies that need the estimation of the viscosity at different conditions of pressure and temperature (Barati-Harooni and Najafi-Marghmaleki, 2016; Hosseinifar and Jamshidi, 2016; Ershadnia et al., 2020; Xu et al., 2020). By this definition, it is very important to have an accurate model for determining the oil viscosity. Generally, two different groups of the models are used for estimating oil viscosity; one is based on the bulk properties of fluid such as pressure, temperature, bulk density (black oil) and the other one is based upon focusing on the properties of the individual components of fluid such as the

Yandy Scientific Press

*Corresponding author.
E-mail address: mahdiani@aut.ac.ir (M.R. Mahdiani); khamehchi@aut.ac.ir (E. Khamehchi); sassanh@princeton.edu (S. Hajirezaie); hemmati@uk.a.c.ir & aut.hemmati@gmail.com (A. Hemmati-Saraparde).

Received November 7, 2020; revised November 22, 2020; accepted November 23, 2020; available online November 25, 2020.

density of each component and composition. The first group is called black oil modeling and the second group is known as compositional modeling (Parsi et al., 2015; Sakthipriya et al., 2015; Shetty et al., 2016). The first one is simpler and faster than second one but the second one is more accurate. Black oil models are more commonly used in different studies of petroleum engineering (Mahdiani and Khamehchi, 2015a, 2015b). There are different studies that have tried to find a model for estimating crude oil viscosity in the literature. Most of these models are created by correlating a model with a database of a specific type of oil from a specific field with a specific range of thermodynamic parameters and it is clear that how much it can be limiting. There is no guarantee that the resulted model could be applicable for other oils at other conditions out of the range of the used data for creating that model. Pressure can have a large impact on the fluid state and based on that the fluid can be categorized into two saturated and under-saturated groups. The main effect of pressure on fluid is the amount of gas that can be dissolved in that fluid (Mahdiani and Khamehchi, 2016). An oil which does not have any dissolved solution gas is called dead oil. Usually, the behaviors of fluid properties for a fluid in saturated and under saturated states are very different. As an example, in saturated fluids, increasing pressure reduces the viscosity, while this behavior is inverse in under saturated ones (Daridon et al., 2016; Mahdiani and Kooti, 2016; Salehinia et al., 2016). As mentioned before, the viscosity of an oil with no solution gas (dead oil) is different than the viscosity of an oil with dissolved soluble gas (live oil). However, live oil viscosity is related to the viscosity of that oil with no solution gas. In other words, the viscosity of live oil is a function of the viscosity of the same oil in dead oil status (no dissolved gas). There are many studies that focus on the viscosity of dead oil as discussed in following paragraphs. Most of these studies have focused on dead oil viscosity as a basis. In this study, the dead oil viscosity is modeled too. In addition, in most studies, the resulted model is a function of oil gravity and temperature, while in some studies, molar mass and critical properties are considered as the input of the models as well (Mehrotra, 1991; Svrcek and Mehrotra, 1998; Hemmati-Sarapardeh et al., 2014; Dehaghani and Badizad, 2016).

There are various correlations for estimating oil viscosity in literature. Some of the most important ones are Beal (1946), Beggs et al. (1975), Glaso (1980), Labedi (1982, 1992), Kaye (1985), Al-Khafaji et al. (1987), Khan et al. (1987), Egbogah and Ng (1990), Kartoatmodjo and Sschmidt (1994), Petrosky and Farshad (1995), Bennison (1998), Dutt (1998), Elsharkawy and Alikhan (1999), Whitson and Brulé (2000), Barrufeta and Dexheimerb (2004), Hossain et al. (2005), Naseri et al. (2005), Omole and Deng (2009), Hemmati-Sarapardeh et al. (2014, 2016) and Khamehchi et al. (2020).

Generally, the models for estimating crude oil viscosity are divided into two main categories. Some such as Bennison (1998), Dutt (1998) and Hossain et al. (2005) are applicable only for heavy oils (in which oil API $<$ 20) while other models such as Beggs et al. (1975), Glaso (1980), Labedi (1982, 1992), Khan et al. (1987), Petrosky and Farshad (1995), Elsharkawy and Alikhan (1999), Barrufeta and Dexheimerb (2004), Naseri et al. (2005) and Omole and Deng (2009) are developed for light oils. Dutt (1998) used 250 data points from different oil fields for estimating crude oil viscosity. In the same year, Bennison (1998) created a model by fitting it to experimental data points. Hossain et al. (2005) used the data points of heavy oils with API gravity from 10 to 22.3 and created a model for heavy oils. Beggs et al. (1975) developed a model for light oils using 460 light oil data points. Glaso (1980), created his model in 1980 based on the data of the North Sea oils. Two years later in 1982, Labedi (1982) used the data of Nigeria and Angola oils for creating his model. He (Labedi, 1992) used the data of Libya oil in 1992 to create a predictive model. Khan et al. (1987) used 75 data points to create a model for Saudi Arabian oils. Elsharkawy and Alikhan (1999) used the data of Middle East oils for developing their model. Barrufeta and Dexheimerb (2004) used a database to create a model for predicting crude oil viscosity. Naseri et al. (2005) used Iranian oils for making a model. Omole and Deng (2009) used the data of Nigerian oil and used artificial neural network to create an intelligent model. Hemmati-Sarapardeh et al. (2013) used 120 data points to create a model for estimating crude oil viscosity. Later in 2014 and 2016, Hemmati-Sarapardeh et al. (2014, 2016) used various intelligent approaches for creating new models for estimating crude oil viscosity. Li et al. (2018) used a JIT-based extreme earning machine for predicting the oil viscosity. Also, Talebkeikhah et al. (2020) using a compositional modeling approach, created a mode to predict the viscosity of various oils. Also, in this year Khamehchi et al. (2020), using more than 1,000 datapoints, developed models using various machine learning algorithms. Although their models (simulated annealing programing (SAP), decision tree (DT), and multilayer perceptron (MLP)) had good results, but they did not try to reduce the dimensions of the data to make their model more efficient. In addition, their database was very dense, so it seemed for that kind of database k-nearest neighbor (KNN) can make a very good model.

Most of the previous works suffer from limited applicable range because they are applicable only in the range of the data points in which they have been developed. In addition, those data points were from a specific field and thus the created model is accurate only for the same field. Here in this study, the objective is finding a universal model for estimating light and intermediate crude oils using the data bank of Hemmati-Sarapardeh et al. (2014, 2016). This data bank was also used by Khamechchi et al. (2020); however, we removed the outlier data from this data bank to develop a better model. Because of focusing on light and intermediate oils, considering only API gravity and temperature for estimating viscosity is sufficient.

The main method used in most of the previous studies is based on simple regression. Because of that, these models suffer from inaccuracy in estimating some data points and inflexibility for changing their shape when their input database is extended (Mahdiani and Khamehchi, 2014; Mahdiani and Kooti, 2016). Here, first an extensive database from literature is collected. This database contains the viscosity, API gravity and temperature of medium to light oils. Using artificial intelligence and heuristic methods is a powerful way to analyses

complicated (Khamehchi and Mahdiani, 2017). Thus, in this study, three intelligent methods, KNN, genetic programming (GP) and linear discriminant analysis (LDA) are used to develop flexible and accurate models for estimating crude oil viscosity. These three algorithms are representative of clustering and non-clustering algorithms. The resulted models are accurate and flexible enough to widen their range of usage by importing a more extended dataset. The accuracy of the created models is measured using statistical and graphical error analysis and their performance is compared with previously introduced models. In addition, the trend prediction of the models is compared with that of experimental data and the effect of different parameters on the estimation error of models is investigated.

2. Model description

In this work, by using three different intelligent modeling methods, three models for estimating crude oil viscosity were developed. For developing these models, the experimental data of temperature, oil API gravity and their corresponding viscosity were used; These data are collected from different geological locations all over the world (Everett and Weinaug, 1955; Glaso, 1980; Miadonye et al., 1992; De Ghetto et al., 1995; Degiorgis et al., 2001; Hossain et al., 2005; Naseri et al., 2005; Al-Maamari et al., 2006; Croft and Patzek, 2009; Naseri et al., 2012; Sadeghi et al., 2013; Alomair et al., 2014). The statistical parameters of these data are shown in Table 1. These modeling procedures are briefly explained here:

Table 1. The statistical parameters of the used data for developing the models.

	API	T (K)
Max	50.00	42.04
Average	34.39	313.06
Median	35.10	310.93
Min	20.00	273.15
SD	5.38	19.84

2.1 K-nearest neighbor (KNN)

KNN algorithm is a pattern recognition methods that is widely used in different scientific areas such as economics (Li et al., 2016), mechanical engineering (Baraldi et al., 2016), energy (Huang and Perry, 2016), and medical sciences (Chen et al., 2015). It is known for its easy methodology, good interpretation and its low computation time. This algorithm works based on the assumption that points with similar inputs have similar outputs. First, the points are classified into some clusters based on their similar properties in n-dimensional space. n is the number input parameters. Fig. 1 shows the classification of patients based on their respiration affected by a drug with various ratios of sodium to potassium based on

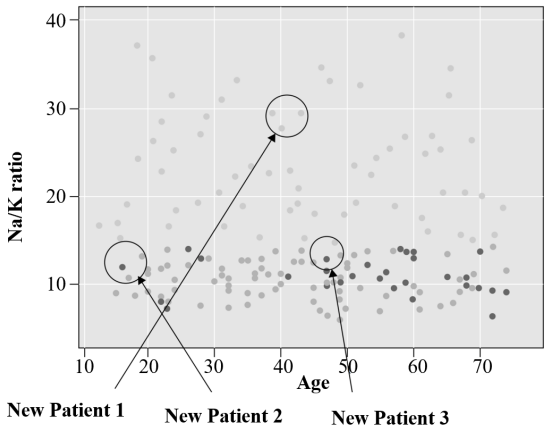

Fig. 1. Data classification; the points have three colors, the lighter color shows the people with no effect, the medium color shows people with medium effect and the dark color shows people with overdose (Larose, 2014).

their age (Tanveer et al., 2016; Xu, 2016). Then for a new point, the k nearest points to that point are selected and are analyzed to find the class which has the largest number of points near the new point. Usually, k is a small number and therefore, a hypercube can be assumed in which its center is on the new point and it grows bigger and bigger until k points fall inside it. Next, the points are counted and it is determined from which cluster more points exist among the points of the hypercube. Thus, the new point is assigned to that cluster and its output is predicted using the method of cluster prediction (Hu et al., 2020). Fig. 1 shows some examples of new points and their classification based on their distance to the members of clusters. The value of k has a great effect on the KNN performance, especially when some noisy points exist. Using heuristic methods to find the best k is an effective approach for developing an efficient machine learning model (Papadopoulos, 2006; Galdames, 2008; Mucherino et al., 2009).

There are different functions for distance measuring. The most common ones are Euclidean, Manhattan, and Minkowski distances defined as follows:

Elucidean:

$$d = \sqrt{\sum (x_i - y_i)^2} \tag{1}$$

Manhattan:

$$d = \sum |x_i - y_i| \tag{2}$$

Minkowski:

$$d = \left(\sum |x_i - y_i|^p\right)^{\frac{1}{p}} \tag{3}$$

where x_i and y_i are the attributes of the two points. Also, p is a real value between 1 and 2. For discrete values, the Hamming distance is used as follows:

$$\begin{aligned} D_H &= \sum_{i=1}^{k} |x_i - y_i| \\ x = y &\Rightarrow D = 0 \\ x \neq y &\Rightarrow D = 1 \end{aligned} \tag{4}$$

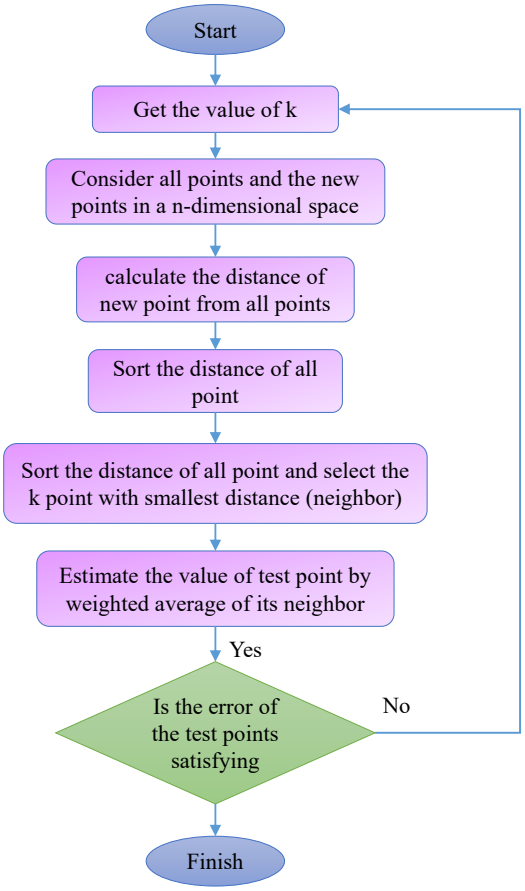

Fig. 2. A simple flowchart for the k-nearest neighbor modeling.

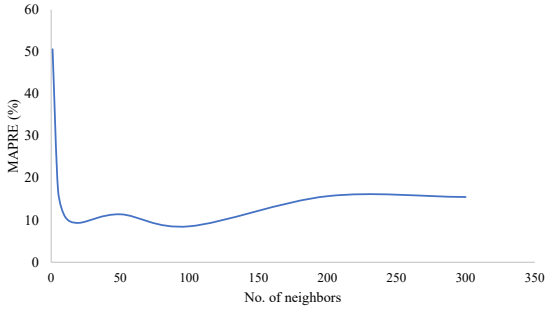

Fig. 3. The MAPRE (%) of the developed KNN models with various number of k.

in which D stands for distance and again x_i and y_i are the attributes of the two points.

Fig. 2 shows a flowchart for KNN algorithm. For selecting the most suitable number of neighbors (k), a sensitivity analysis for k is done which is shown in Fig. 3. When k is too small (lower than 10), the error of the model is very high. By increasing the number of k from 0 to 10, the model error decreases. After 10 to 100 the error decreases with slight slope and some fluctuations. k = 100 gives the least error and after 100 error increases again.

Also, the internal parameters of the used KNN of the current study are shown in Table 2.

Table 2. The internal parameters of the KNN model.

Parameters	Value
NumNeighbors	100
NSMethods	kd-tree
Distance	euclidean
Bucket size	50
include ties	0
distance weight	equal
break ties	smallest
standardize data	1
type	classification
mu	[34.40, 313.22]
sigma	[5.45, 20.12]
W	6.66E-04

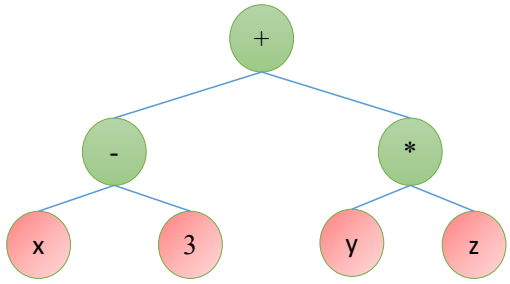

Fig. 4. Schematic of a tree structure. Green shows nodes and red shows terminals.

It should not be forgotten that before doing the calculations, all parameters should be normalized. Output estimation is performed after the classification of data points. One way of doing that is using the weighted average method. In this method, the points that are closer to the new point have a higher effect on the output of that the new point. In this case, the inverse of the distance can be considered as the weight (Cios et al., 2007).

2.2 Genetic programming (GP)

Changing the optimization method to be applicable in modeling is widely used in different problems (Mahdiani and Khamehchi, 2014; Hien et al., 2020). A rigorous optimization approach is genetic programing, which has been evolved from genetic algorithm. GP is an evolutionary method for modeling various problems, which is extensively used in different projects such as construction (Gandomi et al., 2016), mining (Faradonbeh et al., 2016), and chemical engineering (Kaydani et al., 2016). This algorithm is based on genetic algorithm, but instead of applying to a series of points, it is applied to a tree structure. A tree is a flexible structure that is used in different areas such as math, engineering, etc. Here, tree structure is used to represent the candidate models (equations) of the problem. Fig. 4 shows a sample tree. As this figure shows, a series of terminals (which can contain a variable or

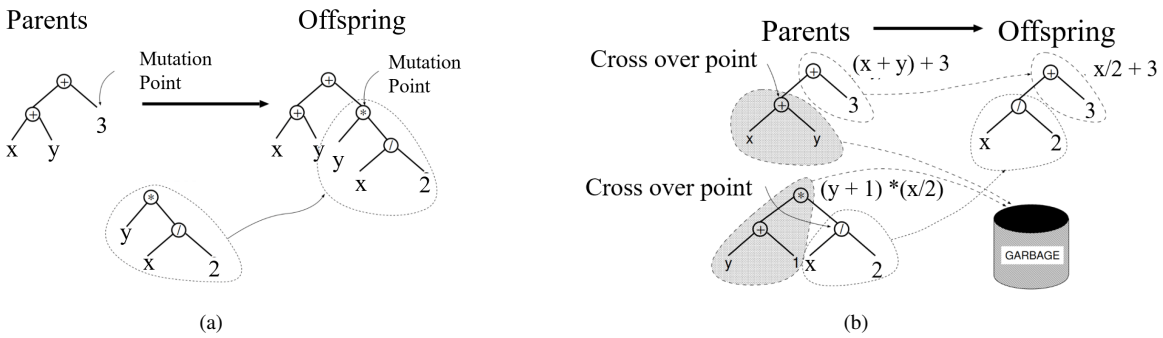

Fig. 5. Operation on trees in GP. (a) mutation, (b) crossover (Poli et al., 2008).

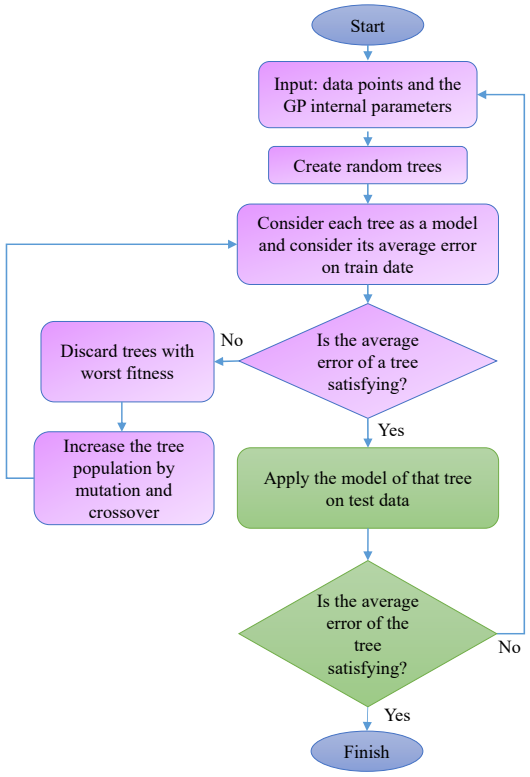

Fig. 6. Genetic programming flowchart.

a fixed value) are connected to nodes (which contain operators) and the nodes are connected to other nodes until the structure of the tree creates the equation (Mahdiani and Kooti, 2016).

Before utilizing GP, it is necessary to have some knowledge about genetic algorithm. In genetic algorithm, first a random possible population is created. This generation continues until an individual (possible solution) with satisfying fitness is found. In each generation, some individuals with better finesses are selected and the others are omitted. In addition, various operators are used for increasing the population such as crossover and mutation. In crossover, two individuals are mixed and a new solution is born, in which each part of it belongs to one of the parent individuals (Affenzeller et al., 2009; Chen, 2012). In GP, all of the mentioned operations are applied to trees.

Other parts of GP are similar to genetic algorithm. Here, crossover and mutation, which are slightly different from genetic algorithm, are explained. Fig. 5(a) represents mutation and crossover over trees. This figure shows one of the most common types of mutations known as one-point mutation in which one node is selected and its corresponding subtree is replaced by a randomly generated subtree (Langdon, 2012). Another operation applied to trees is crossover. There are different kinds of crossovers, but the most common one is subtree crossover. In this kind of crossover, two nodes of the parents are selected randomly and then the subtree of one parent is replaced with that of the other one (Langdon and Poli, 2002). This is shown in Fig. 5(b). Finally, Fig. 6 shows the flowchart of using GP in the study of this paper; also, the internal parameters of the GP model of this paper are illustrated in Table 3.

Table 3. The internal parameters of the GP model.

Parameter	Values
Population size	300
Initial Population Generation	Random
Initial Score	Calculated by Fitness Function
Selection Function	Tournament
Mutation Function	Uniform
Crossover Function	Scattered (Two parents)
Type of replacement	elitist
Crossover Probability	0.7
Mutation Probability	0.05
Elite count	30
Hybrid	No
Max generation	500
Max Stall generation	100
Function tolerance	1.00E-06
Fraction of constants in initial	0.5
Tree maximum depth	15
Maximum mute depth	10
Tree nodes functions	4 basic operator plus exponential and logarithm
Maximum nodes	Inf

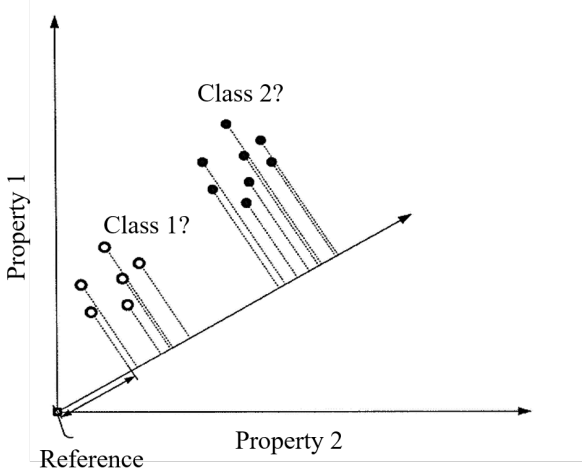

Fig. 7. Two classes of points on a 2D plot (Trifonov and Lalyko, 2010).

2.3 Linear discriminant analysis (LDA)

LDA is a method of pattern recognition and machine learning, which finds a linear combination of features to relate the inputs to outputs. It has been used in various area of science such as chemistry (Abbruzzo et al., 2016), hydrology (Close et al., 2016) and cereal science (Promchan et al., 2016). Finding the correct weight of each part is very important and is the learning part of this algorithm. LDA uses regression and variance analysis to state the outputs as a linear combination of the inputs. One of the main functions of LDA is to reduce the dimension of the problem. LDA should decrease the dimension of the problem such that the data can be classified separately into various classes (McLachlan, 2004).

One example of classification is the two classes of data points, which are in a two-dimensional (2D) sheet. The 2D points can be projected to the horizontal or vertical axis to make the problem one dimensional. However, the problem is that one axis is completely ignored in that method in addition to that the classes are not separated completely. Fig. 7 shows a line that all of the points can be projected on that such that none of the axes is ignored as well as the classes are separated completely (Gnanadesikan, 1988).

The equation of the line can be found by some statistical methods. Using the above procedure, the dimension of classes is reduced as many as possible. The new points are projected to the lower dimension to see their class and by using maximum likelihood estimation (MLE), their output can be estimated (Deng et al., 2011). Here, a method for finding the equation of the line is explained:

As mentioned earlier, LDA is based on linear combination of variables for separating the classes. For separating the classes. For separating the classes, the following functions are used:

$$Z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d \tag{5}$$

$$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \tag{6}$$

$$S(\beta) = \frac{\bar{Z}_1 - \bar{Z}_2}{\text{Variance of } Z\text{within groups}} \tag{7}$$

Knowing the score function, the objective is to find the linear coefficient that maximizes the score by solving the following equations:

Model coefficient:

$$\beta = C^{-1}(\mu_1 - \mu_2) \tag{8}$$

x_i: data points attributes

Z: linear combination of predictors

Pooled covariance matrix:

$$C = \frac{1}{n_1 + n_2}(n_1 C_1 + n_2 C_2) \tag{9}$$

where $S(\beta)$ is score function, β is linear model coefficient, C_1, C_2 are covariance matrices, μ_1, μ_2 are mean vectors.

Next, the Mahalanobis distance between two groups is calculated. If its value is more than three, it means that the classification is good.

$$\Delta^2 = \beta^T(\mu_1 - \mu_2) \tag{10}$$

where Δ is the Mahalanobis distance between two class.

A new point is classified into class C_1 if:

Model coefficient:

$$\beta^T \left(x - \left(\frac{\mu_1 + \mu_2}{2} \right) \right) > \log \frac{P(c_1)}{P(c_2)} \tag{11}$$

where β is coefficient vector, x is data vector, μ is Mean vector, P is class probability.

Fig. 8 shows a flowchart for LDA. Table 4 shows the internal parameters of LDA model of this study.

3. Results and discussion

In the current article, three machine-learning models (GP, LDA and KNN) were developed to use oil API gravity and temperature for predicting the oil viscosity. To achieve this objective, a large databank covering different ranges of input (oil API gravity and temperature) and output (oil viscosity) parameters was collected from the literature. The data was

Table 4. The internal parameters of the LDA model.

Parameter	Values
Discriminant Type	diagonal linear
Type	Classification
W	6.66E-04
Gamma	1
Delta	0
Sigma	[9.97, 282.86]
Between Sigma	[28.57, -6.79]
	[-6.79, 375.13]
Delta Predictor	[4.94, 6.47]
log det	7.94

Table 5. Statistical parameters of the models in the train, test and total data set.

		KNN	GP	LDA
Training set	MAPRE (%)	8.22	23.48	28.49
	MPRE (%)	1.6	10.42	-10.06
	SD	0.02	0.35	0.18
	RMSE, cP	2.33	64.96	26.16
	Number of data	1,000	1,000	1,000
Test set	MAPRE (%)	9.12	23.15	30.93
	MPRE (%)	2.15	10.42	-10.06
	SD	0.03	0.17	0.2
	RMSE, cP	1.52	74.21	26.02
	Number of data	433	433	433
All data	MAPRE (%)	8.54	23.3	29.51
	MPRE (%)	1.6	10.67	-10.94
	SD	0.02	0.29	0.19
	RMSE, cP	2.09	67.04	25.87
	Number of data	1,433	1,433	1,433

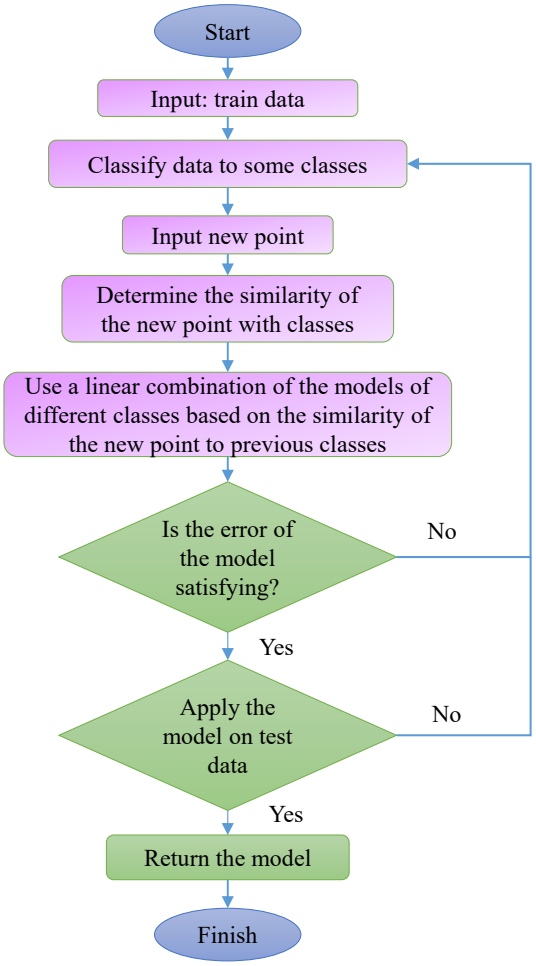

Fig. 8. A flowchart for the linear discriminant analysis.

divided into two sections of testing and training sets. Table 2 to Table 4 shows the model parameters. Afterward, using the three algorithms of GP, LDA and KNN three models were developed. The details of internal parameters of these algorithms and how these methods are used to create a model are completely explained in model description section and its sub-sections.

The statistical analysis of the testing and training data sets are shown in Table 5. These results show that the developed models can efficiently predict oil viscosity by showing a total mean absolute percent relative error (MAPRE) %, standard deviation (SD) and root mean square error (RMSE) of 8.54%, 0.02 cP and 2.09 cP for the KNN model, 23.48%, 0.35 and 64.96 cP for the GP model and 28.49%, 0.18 cP and 26.16 cP for the LDA model, respectively. In 3 model that has been developed in this study, KNN has the best performance and accuracy. Fig. 9 shows a 3-D scatter plot the experimental and predicted by KNN viscosity, versus API and temperature.

Different statistical and graphical error analyzing methods were employed to compare the accuracy of the developed models and to compare them with the well-known previously previous models.

The results of MAPRE% analysis for the developed models as well as the previously published models are summarized in Table 6. As this table shows, KNN model outperforms the previously published models in predicting reservoir oil viscosity. It can be observed in this table that Beggs et al. (1975) and Glaso (1980) models are of the least accurate models in estimating reservoir oil viscosity.

As it was mentioned earlier, graphical error analysis was utilized to investigate the performance of the developed models. A graphical analysis of the MAPRE% of the created models and the previously published models is illustrated in Fig. 10. It is obvious that the KNN model has a better performance than all of other models by showing the smallest amount of error. The largest error belongs to the Glaso (1980),

Table 6. Statistical analysis of different literature models in estimating crude oil viscosity.

Models	MAPRE (%)	MPRE (%)	SD	RMSE, cP
Beggs and Robinson	1,651.05	-1,643.47	444.13	595,573.92
Glaso	99.97	99.97	1	69.29
Bennison	88.84	84.8	0.91	40.32
Hossain	87.68	85.65	0.9	39.73
Labedi-Nigeria and Angol	41.49	-17.96	0.53	57.82
labedi-Libya	36.86	16.08	0.61	61.52
Hemmati-Sarapardeh	32.95	31.09	0.38	56.68
Kaye	32.78	21.08	0.55	543.62
Naseri	32.24	29.13	0.38	55.68
Petrosky	30.38	25.6	0.37	59.93
Egbogah and Ng	26.42	17.64	0.36	42.58
Elsharkawy	23.08	-6.93	0.4	321.41
Kartoatmodjo and Schimidt	22.81	15.12	0.29	48.69
Al-Khafaji	21.55	5.18	0.29	54.96
LSSVM	18.33	-35.01	11.18	24.65
SAP	18.16	3.91	0.06	24.92
DT	12.75	-1.81	0.08	28.31
MLP	21.01	-6.89	0.10	31.46
LDA	29.51	-10.94	0.19	25.87
KNN	8.54	1.6	0.02	3.35
GP	23.3	10.67	0.29	65.5

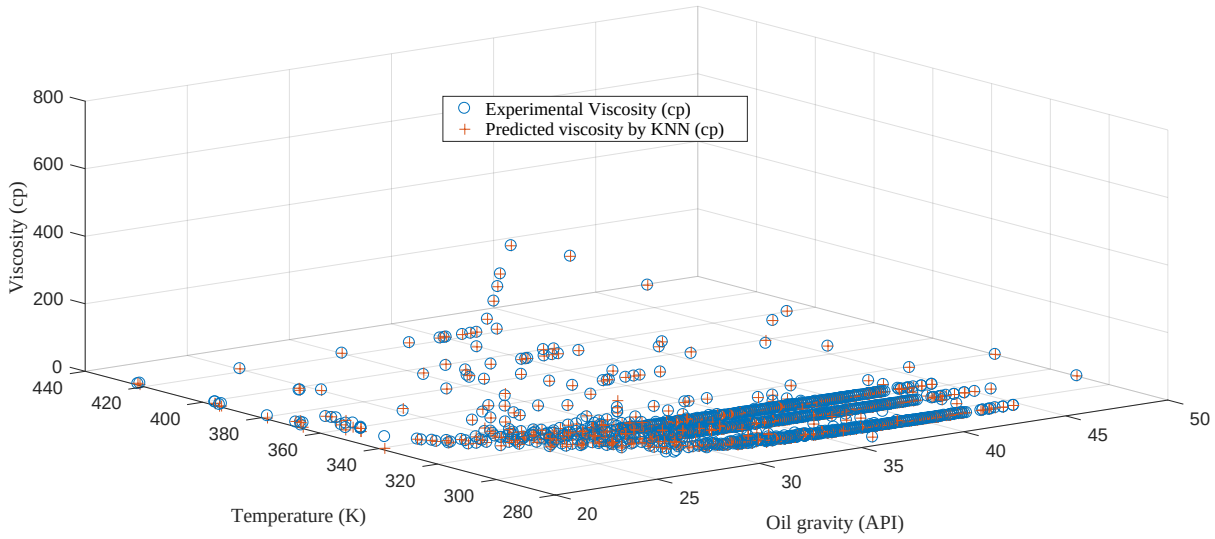

Fig. 9. 3D scatter plot of viscosity predicted by KNN and measured experimental viscosity.

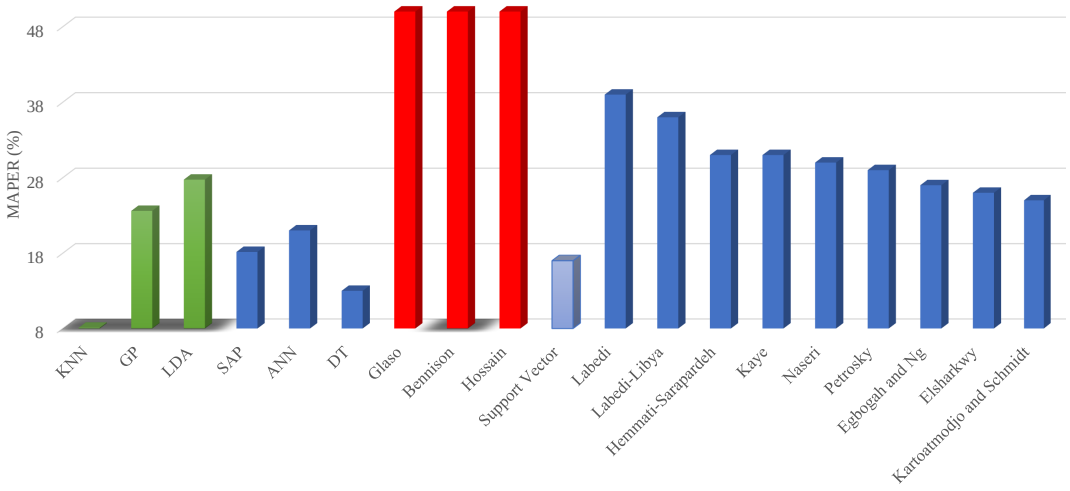

Fig. 10. 3D scatter plot of viscosity predicted by KNN and measured experimental Comparison of the MAPRE (%) of the models developed in this paper and some other most common models of the literature for predicting crude oil viscosity.

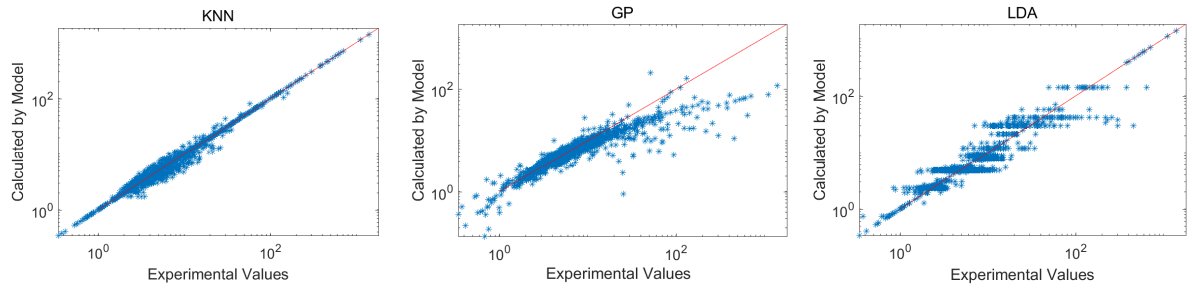

Fig. 11. Cross plot of the experimental data versus different models estimations: (a) KNN, (b) GP, (c) LDA.

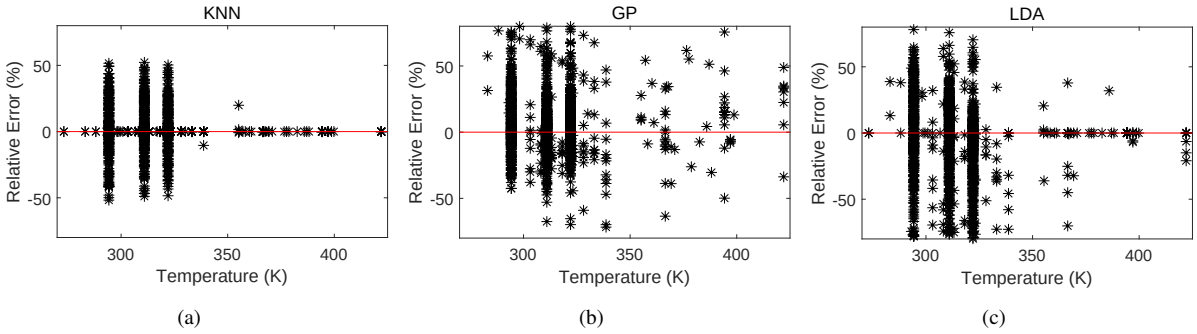

Fig. 12. Trend analysis of the developed models in this paper: (a) KNN, (b) GP, (c) LDA.

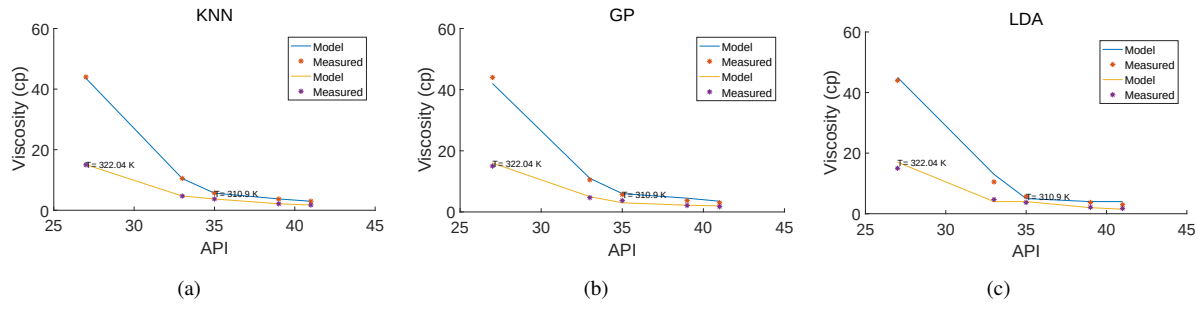

Fig. 13. Different values of viscosity estimated by the different models developed in this paper versus experimental values: (a) KNN, (b) GP, (c) LDA.

Table 7. Some sample viscosity estimations by the different models of this paper.

API	T (K)	Exp. Oil Viscosity (cP)	KNN	GP	LDA
36.2	322.03889	3.77	3.77	3.424604	4.85
39	322.03889	2.696	2.653	2.438253	2.53
41.4	322.03889	2.175	2.259	1.824858	2.38
31	294.26111	13.258	13.258	13.85964	30
38.4	294.26111	5.366	4.014	5.346208	4.941
40.8	322.03889	1.741	1.741	1.962904	2.38
20.9	323.15	53.8	53.8	27.77395	42
31.2	322.03889	5.523	5.523	6.649256	8.9
29.3	310.92778	14.6	14.6	11.71741	30

Bennison (1998) and Hossain et al. (2005) models.

In the next step, the predicted data by the developed models were plotted against the experimental data in the format of cross plots as shown in Fig. 11 to assess the robustness of each model. As can be observed, the KNN model shows the closest cloud of data points to the unit slope line meaning that this model's predictions match the experimental data better than the other models.

Error distribution was another graphical approach employed in this paper to evaluate the performance of the models. The results of error distribution analysis based on temperature are reported in Fig. 12. As can be seen, all of the intelligent models illustrate a good performance by having a negligible error trend over the chosen temperature zone. It can be seen in this figure that the KNN model has the smallest error cloud around the zero-error line. Also, this figure shows that this model has a great performance at high temperatures as well as low temperatures. The plots show that most of the data points are located in the low temperature region (< 330K).

Finally, in order to get a better understanding of the models predictions as well as to compare their results with the physical real trend of data, the predicted and experimental viscosity data as a function of oil gravity at two temperatures was plotted in Fig. 13. As the experimental trends show, viscosity decreases as the oil API increases. It can also be observed that at a specific API, viscosity decreases with temperature. The plots at higher oil APIs show that there is not a significant change in viscosity when temperature changes. This means that temperature has a larger impact on heavy oils as it eases the movement of liquid molecules when they have less motion compared to higher temperatures. Also, to be able to see the exact number of each model predication, the prediction of viscosity on some sample API and temperature is shown in Table 7.

4. Conclusions

In this study, oil viscosity is modeled using three intelligent models called discriminant analysis (LDA), KNN and GP. A large data bank was collected from various sources of literature to cover a wide range of oil gravity and temperature conditions. The results of statistical and graphical error analysis uncovered that the developed models outperform all of the previously published models for oil viscosity estimation. Among the developed models, KNN was found to be the most accurate model by showing a total mean absolute error of 8.54%. GP was found to be the next model by showing an estimation error of 23.48% followed by LDA with an error of 28.49%. Error distribution curves showed that the models follow the real trend of experimental data with no significant error over a wide range of temperatures. Trend analysis illustrated that the developed models follow the experimental data trend with high accuracy over a wide range of API conditions. In addition, cross plots of the models indicated that the models of this paper can effectively and accurately predict oil viscosity.

Conflict of interest

The authors declare no competing interest.

References

Abbruzzo, A., Tambur, E., Varrica, D., et al. Penalized linear discriminant analysis and discrete adaboost to distinguish human hair metal profiles: The case of adolescents residing near mt. Etna. Chemosphere 2016, 153: 100-106.

Abubakar, A., Al-Wahaibi, Y., Al-Wahaibi, T., et al. Effect of low interfacial tension on flow patterns, pressure gradients and holdups of medium-viscosity oil/water flow in horizontal pipe. Exp. Therm. Fluid Sci. 2015, 68: 58-67.

Affenzeller, M., Wagner, S., Winkler, S., et al. Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications. Boca Raton, USA, Crc Press, 2009.

Al-Khafaji, A.H., Abdul-Majeed, G.H., Hassoon, S.F., et al. Viscosity correlation for dead, live and undersaturated crude oils. J. Pet. Res. 1987, 6(2): 1-16.

Al-Maamari, R.S., Houache, O., Abdul-Wahab, S.A. New correlating parameter for the viscosity of heavy crude oils. Energy Fuels 2006, 20(6): 2586-2592.

Al-Sarkhi, A., Pereyra, E., Sarica, C., et al. Positive frictional pressure gradient in vertical gas-high viscosity oil slug flow. Int. J. Heat Fluid Flow 2016, 59: 50-61.

Alomair, O., Elsharkawy, A., Alkandari, H. A viscosity prediction model for kuwaiti heavy crude oils at elevated temperatures. J. Pet. Sci. Eng. 2014, 120: 102-110.

Baraldi, P., Cannarile, F., Di Maio, F., et al. Hierarchical k-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions. Eng. Appl. Artif. Intell. 2016, 56: 1-13.

Barati-Harooni, A., Najafi-Marghmaleki, A. An accurate rbf-nn model for estimation of viscosity of nanofluids. J. Mol. Liq. 2016, 224: 580-588.

Barrufeta, M.A., Dexheimerb, D. Use of an automatic data quality control algorithm for crude oil viscosity data. Fluid Phase Equilib. 2004, 219(2): 113-121.

Beal, C. The viscosity of air, water, natural gas, crude oil and its associated gases at oil field temperatures and pressures. Trans. AIME 1946, 165(1): 94-115.

Beggs, D.H., Robinson, J.R. Estimating the viscosity of crude oil systems. J. Pet. Technol. 1975, 27(9): 1140-1141.

Bennison, T. Prediction of heavy oil viscosity. Paper Presented at IBC Heavy Oil Field Development Conference, 1998.

Chen, C.-H., Huang, W.-T., Tan, T.-H., et al. Using k-nearest neighbor classification to diagnose abnormal lung sounds. Sensors 2015, 15(6): 13132-13158.

Chen, S.-H. Genetic Algorithms and Genetic Programming in Computational Finance. Berlin, Germany, Springer, 2012.

Cios, K.J., Pedryc, W., Swin, R.W. Data Mining: A knowledge Discovery Approach. Berlin, Germany, Springer, 2007.

Close, M.E., Abraham, P., Humphries, B., et al. Predicting groundwater redox status on a regional scale using linear discriminant analysis. J. Contam. Hydrol. 2016, 191: 19-32.

Croft, G.D., Patzek, T.W. The future of california's oil supply. Paper SPE 120174 Presented at SPE Western Regional Meeting, San Jose, California, 24-26 March, 2009.

Daridon, J.L., Orlandi, E., Carrier, H. Measurement of bubble point pressure in crude oils using an acoustic wave sensor. Fluid Phase Equilib. 2016, 427: 152-160.

De Ghetto, G., Paone, F., Villa, M. Pressure-volume-temperature correlations for heavy and extra heavy oils. Paper SPE 30316 Presented at SPE International Heavy Oil Symposium, Calgary, Alberta, Canada, 19-21 June, 1995.

Degiorgis, G., Maturano, S., Garay, M., et al. Oil mixture viscosity behavior: Use in pipeline design. Paper SPE 69420 Presented at SPE Latin American and Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina, 25-28 March, 2001.

Dehaghani, A.H.S., Badizad, M.H. Experimental study of iranian heavy crude oil viscosity reduction by diluting with heptane, methanol, toluene, gas condensate and naphtha. Petroleum 2016, 2(4): 415-424.

Deng, H., Miao, D., Lei, J. Artificial Intelligence and Computational Intelligence. Berlin, Germany, Springer, 2011.

Dutt, N.V.K. A simple method of estimating the viscosity of petroleum crude oil and fractions. Chem. Eng. J. 1990, 45(2): 83-86.

Egbogah, E.O., Ng, J.T. An improved temperature-viscosity correlation for crude oil systems. J. Pet. Sci. Eng. 1990, 4(3): 197-200.

Elsharkawy, A.M., Alikhan, A.A. Models for predicting the viscosity of middle east crude oils. Fuel 1999, 78(8): 891-903.

Ershadnia, R., Amooie, M.A., Shams, R., et al. Non-newtonian fluid flow dynamics in rotating annular media: Physics-based and data-driven modeling. J. Pet. Sci. Eng. 2020, 185: 106641.

Everett, J., Weinaug, C.F. Physical properties of eastern kansas crude oils. Kansas Geological Survey, 1955.

Faradonbeh, R.S., Armaghani, D.J., Monjezi, M., et al. Genetic programming and gene expression programming for flyrock assessment due to mine blasting. Int. J. Rock Mech. Min. Sci. 2016, 88: 254-264.

Galdames, P. Managing continuous k-nearest neighbor queries in mobile peer-to-peer networks. Lowa State, Lowa State University, 2008.

Gandomi, A.H., Sajedi, S., Kiani, B., et al. Genetic programming for experimental big data mining: A case study on concrete creep formulation. Autom. Constr. 2016, 70: 89-97.

Glaso, O. Generalized pressure-volume-temperature correlations. J. Pet. Technol. 1980, 32(5): 785-795.

Gnanadesikan, R. Discriminant Analysis and Clustering. Washington, USA, National Achademy Press, 1988.

Hemmati-Sarapardeh, A., Aminshahidy, B., Pajouhandeh, A., et al. A soft computing approach for the determination of crude oil viscosity: Light and intermediate crude oil systems. J. Taiwan Inst. Chem. Eng. 2016, 59: 1-10.

Hemmati-Sarapardeh, A., Khishvan, M., Naseri, A., et al. Toward reservoir oil viscosity correlation. Chem. Eng. Sci. 2013, 90: 53-68.

Hemmati-Sarapardeh, A., Majidi, S., Mahmoudi, B., et al. Experimental measurement and modeling of saturated reservoir oil viscosity. Korean J. Chem. Eng. 2014, 31(7): 1253-1264.

Hien, N.T., Tran, C.T., Nguye, X.H. Genetic programming for storm surge forecasting. Ocean Eng. 2020, 215: 107812.

Hossain, M.S., Sarica, C., Zhang, H.-Q., et al. Assessment and development of heavy oil viscosity correlations. Paper SPE 97907 Presented at SPE International Thermal Operations and Heavy Oil Symposium, Calgary, Alberta, Canada, 1-3 November, 2005.

Hosseinifar, P., Jamshidi, S. A new correlative model for viscosity estimation of pure components, bitumens, size-asymmetric mixtures and reservoir fluids. J. Pet. Sci. Eng. 2016, 147: 624-635.

Hu, J., Peng, H., Wang, J., et al. Knn-p: A knn classifier optimized by p systems. Theor. Comput. Sci. 2020, 817: 55-65.

Huang, J., Perry, M. A semi-empirical approach using gradient boosting and kk-nearest neighbors regression for gefcom2014 probabilistic solar power forecasting. Int. J. Forecast. 2016, 32(3): 1081-1086.

Ilieva, P., Kilzer, A., Weidner, E. Measurement of solubility, viscosity, density and interfacial tension of the systems tristearin and CO_2 and rapeseed oil and CO_2. J. Supercrit. Fluids 2016, 117: 40-49.

Ilyin, S., Arinina, M., Polyakova, M., et al. Asphaltenes in heavy crude oil: Designation, precipitation, solutions, and effects on viscosity. J. Pet. Sci. Eng. 2016, 147: 211-217.

Kartoatmodjo, T., Schmidt, Z. Large data bank improves crude physical property correlations. Oil Gas J. 1994, 92(27): 51-55.

Kaydani, H., Mohebbi, A., Hajizadeh, A. Dew point pressure model for gas condensate reservoirs based on multi-gene genetic programming approach. Appl. Soft Comput. 2016, 47: 168-178.

Kaye, S.E. Offshore california viscosity correlations. COFRC, 1985.

Khamehchi, E., Mahdiani, M.R. Optimization Algorithms, in Gas Allocation Optimization Methods in Artificial Gas Lift. Berlin, Germany, Springer, 2017.

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

Khamehchi, E., Mahdiani, M.R., Suratgar, A.A. Optimizing and stabilizing the gas lift operation by controlling the lift gas specific gravity. J. Pet. Sci. Technol. 2019, 9(3): 46-63.

Khamehchi, E., Mohammad, Z., Mahdiani, M.R. A robust method for estimating the two-phase flow rate of oil and gas using wellhead data. J. Pet. Explor. Prod. Technol. 2020, 10(6): 2335-2347.

Khan, S.A., Al-Marhoun, M.A., Duffuaa, S.O., et al. Viscosity correlations for saudi arabian crude oils. Paper SPE 15720 Presented at Society of Petroleum Engineers, Bahrain, 7-10 March, 1987.

Kleinhans, A., Hornfischer, B., Gaukel, V., et al. Influence of viscosity ratio and initial oil drop size on the oil drop breakup during effervescent atomization. Chem. Eng. Process. 2016, 109: 149-157.

Labedi, R. Improved correlations for predicting the viscosity of light crudes. J. Pet. Sci. Eng. 1992, 8(3): 221-234.

Labedi, R.M. Pvt correlations of the african crudes. Colorado, Colorado School of Mines, 1982.

Langdon, W.B. Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming. Berlin, Germany, Springer, 2012.

Langdon, W.B., Poli, R. Foundations of Genetic Programming. Berlin, GER, Springer, 2002.

Larose, D.T., Larose, C.D. Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, USA, John Wiley & Sons, 2014.

Li, H., Li, Q., Liu, R. Consistent model specification tests based on kk-nearest-neighbor estimation method. J. Econom. 2016, 194(1): 187-202.

Li, Z., Hao, K., Lei, C., et al. Pet viscosity prediction using jit-based extreme learning machine. IFAC-PapersOnLine 2018, 51(18): 608-613.

Mahdiani, M.R., Khamehchi, E. A new method for building proxy models using simulated annealing. Middle-East J. Sci. Res. 2014, 22(3): 324-328.

Mahdiani, M.R., Khamehchi, E. Preventing instability phenomenon in gas-lift optimization. Iran. J. Oil Gas Sci. Technol. 2015a, 4(1): 49-65.

Mahdiani, M.R., Khamehchi, E. Stabilizing gas lift optimization with different amounts of available lift gas. J. Nat. Gas Sci. Eng. 2015b, 26: 18-27.

Mahdiani, M.R., Khamehchi, E. A novel model for predicting the temperature profile in gas lift wells. Petroleum 2016, 2(4): 408-414.

Mahdiani, M.R., Kooti, G. The most accurate heuristic-based algorithms for estimating the oil formation volume factor. Petroleum 2016, 2(1): 40-48.

McLachlan, G.J. Discriminant Analysis and Statistical Pattern Recognition. Hoboken, USA, John Wiley & Sons, 2004.

Mehrotra, A.K. Generalized one-parameter viscosity equation for light and medium liquid hydrocarbons. Ind. Eng. Chem. Res. 1991, 30(6): 1367-1372.

Miadonye, A., Singh, B., Puttagunta, V.R. One-parameter correlation in the estimation of crude oil viscosity. SPE 1992.

Mucherino, A., Papajorgji, P.J., Pardalos, P.M. Data Mining in Agriculture. Berlin, Germany, Springer, 2009.

Naseri, A., Nikazar, M., Mousavi Dehghani, S.A. A correlation approach for prediction of crude oil viscosities. J. Pet. Sci. Eng. 2005, 47(3-4): 163-174.

Naseri, A., Yousefi, S., Sanaei, A., et al. A neural network model and an updated correlation for estimation of dead crude oil viscosity. Braz. J. Pet. Gas 2012, 6(1): 31-41.

Norouzi, M., Panjalizadeh, H., Rashidi, F., et al. Dpr polymer gel treatment in oil reservoirs: A workflow for treatment optimization using static proxy models. J. Pet. Sci. Eng. 2017, 153: 97-110.

Omole, O., Falode, O.A., Deng, A.D.A. Prediction of nigerian crude oil viscosity using artificial neural network. Pet. Coal. 2009, 51(3): 181-188.

Papadopoulos, A.N. Nearest Neighbor Search: A Database Perspective. Berlin, Germany, Springer, 2006.

Parsi, M., Vieira, R.E., Torres, C.F., et al. On the effect of liquid viscosity on interfacial structures within churn flow: Experimental study using wire mesh sensor. Chem. Eng. Sci. 2015, 130: 221-238.

Petrosky, J., Farshad, F.F. Viscosity correlations for gulf of mexico crude oils. Paper SPE 29468 Presented at SPE Production Operations Symposium, Oklahoma, USA, 2-4 April 1995.

Poli, R., Langdon, W.B., McPhee, N.F., et al. A Field Guide to Genetic Programming. Lulu. com, 2008.

Promchan, J., Günther, D., Siripinyanond, A., et al. Elemental imaging and classifying rice grains by using laser ablation inductively coupled plasma mass spectrometry and linear discriminant analysis. J. Cereal Sci. 2016, 71: 198-203.

Sadeghi, M.B., Ramazani S.A., A., Taghikhani, V., et al. Experimental investigation of rheological and morphological properties of water in crude oil emulsions stabilized by a lipophilic surfactant. J. Dispersion Sci. Technol. 2013, 34(3): 356-368.

Sakthipriya, N., Doble, M., Sangwai, J.S. Fast degradation and viscosity reduction of waxy crude oil and model waxy crude oil using bacillus subtilis. J. Pet. Sci. Eng. 2015, 134: 158-166.

Salehinia, S., Salehinia, Y., Alimadadi, F., et al. Forecasting density, oil formation volume factor and bubble point pressure of crude oil systems based on nonlinear system identification approach. J. Pet. Sci. Eng. 2016, 147: 47-55.

Shetty, N., Deshannavar, U.B., Marappagounder, R., et al. Improved threshold fouling models for crude oils. Energy 2016, 111: 453-467.

Svrcek, W.Y., Mehrotra, A.K. One parameter correlation for bitumen viscosity. Chem. Eng. Res. Des. 1998, 66(4): 323-327.

Talebkeikhah, M., Amar, M.N., Naseri, A., et al. Experimental measurement and compositional modeling of crude oil viscosity at reservoir conditions. J. Taiwan Inst. Chem. Eng. 2020, 109: 35-50.

Tanveer, M., Shubham, K., Aldhaifallah, M., et al. An efficient regularized k-nearest neighbor based weighted twin support vector regression. Knowledge-Based Syst. 2016, 94: 70-87.

Trifonov, M.I., Lalyko, L.B. Iterative fisher linear discriminant analysis. Patent, U.S., 2010.

Wen, J., Zhang, J., Wei, M. Effective viscosity prediction of crude oil-water mixtures with high water fraction. J. Pet. Sci. Eng. 2016, 147: 760-770.

Whitson, C.H., Brulé, M.R. Phase Behavior. Texas, USA, Society of Petroleum Engineers Inc., 2000.

Xu, Y. K-nearest neighbor-based weighted multi-class twin support vector machine. Neurocomputing 2016, 205: 430-438.

Xu, Y., Ayala-Orozco, C., Chiang, P.-T., et al. Understanding the role of iron (iii) tosylate on heavy oil viscosity reduction. Fuel 2020, 274: 117808.

Zhang, J., Yuan, H., Zhao, J., et al. Viscosity estimation and component identification for an oil-water emulsion with the inversion method. Appl. Therm. Eng. 2017, 111: 759-767.