

RESEARCH

Open Access



An assessment of ensemble learning approaches and single-based machine learning algorithms for the characterization of undersaturated oil viscosity

Theddeus T. Akano^{1,2*}  and Chinemerem C. James¹

Abstract

Background: Prediction of accurate crude oil viscosity when pressure volume temperature (PVT) experimental results are not readily available has been a major challenge to the petroleum industry. This is due to the substantial impact an inaccurate prediction will have on production planning, reservoir management, enhanced oil recovery processes and choice of design facilities such as tubing, pipeline and pump sizes. In a bid to attain improved accuracy in predictions, recent research has focused on applying various machine learning algorithms and intelligent mechanisms. In this work, an extensive comparative analysis between single-based machine learning techniques such as artificial neural network, support vector machine, decision tree and linear regression, and ensemble learning techniques such as bagging, boosting and voting was performed. The prediction performance of the models was assessed by using five evaluation measures, namely mean absolute error, relative squared error, mean squared error, root mean squared error and root mean squared log error.

Results: The ensemble methods offered generally higher prediction accuracies than single-based machine learning techniques. In addition, weak single-based learners of the dataset used in this study (for example, SVM) were transformed into strong ensemble learners with better prediction performance when used as based learners in the ensemble method, while other strong single-based learners were discovered to have had significantly improved prediction performance.

Conclusion: The ensemble methods have great prospects of enhancing the overall predictive accuracy of single-based learners in the domain of reservoir fluid PVT properties (such as undersaturated oil viscosity) prediction.

Keywords: Pressure volume temperature, Undersaturated oil viscosity, Machine learning algorithms, Evaluation measures, Ensemble method

1 Background

Among all the properties of a fluid, viscosity plays a key role in influencing and controlling flow behavior either in pipelines hydraulics or in porous media such

as a reservoir. The term “viscosity” is described as the internal resistance fluid encounters during flow [1]. In the petroleum industry, a reliable determination of oil viscosity is essential for the analysis of pressure drop which results from fluid flow through tubing or pipes, porous media and so on, for ascertaining well productivity through proper tubing, pipelines and pump size selection, for implementation of secondary recovery processes, enhanced oil recovery processes, reservoir simulation, sound design facilities and for optimum

*Correspondence: takano@unilag.edu.ng; akanott@ub.ac.bw

² Department of Mechanical Engineering, University of Botswana, Gaborone, Botswana

Full list of author information is available at the end of the article

reservoir management [2, 3, 4, 5]. Physical properties and thermodynamics of reservoir fluids such as temperature, pressure, bubble point pressure (Pb), gas/oil ratio (GOR), gas gravity and oil gravity are all dependent on oil viscosity [6] which is usually measured in the laboratory at constant temperature and different pressures considering its variability under changing operating conditions. Empirical correlations can be utilized to estimate qualities over a greater range of pressure and temperature in such instances. However, there are a lot of correlations in the literature, so it is hard to know which one to utilize in each circumstance. Standing [7] developed the first widely accepted correlation for predicting saturation pressure and crude oil formation volume factor. Four parameters influenced this relationship: the gas-to-oil ratio, temperature and the gravity of both oil and gas. Many more correlations were reported for other crude samples based on the same characteristics, with more experimental data used in general than in Standing's study [8–16].

Correlations are being employed in predicting viscosity especially when experimental data are not available at temperatures aside from reservoir temperature. These correlations are gotten at three different conditions, namely at Pb, above Pb and below Pb. The term “saturated oil viscosity” describes the viscosity of the oil when bubble point pressure is reached occurring at reservoir temperature (Tr). Any extended reduction in pressure after Pb is reached results in the release of gases which were initially dissolved in oil. This continues until pressure declines to the pressure of the atmosphere till no more gas initially dissolved in the oil is remaining. The viscosity of such oil is called dead oil viscosity. The undersaturated oil viscosity is the viscosity of the crude oil at a pressure above the bubble point and reservoir temperature [17]. If crude oil is undersaturated at the initial reservoir pressure, the viscosity will drop somewhat as reservoir pressure falls. At saturation pressure, the viscosity will be at its lowest. As the reservoir pressure is reduced below the bubble threshold, the evolution of gas from solution increases the density and viscosity of the crude oil.

Generally, the majority of the discovered oil reservoirs in the Niger-Delta deep water environments of Nigeria are often highly “undersaturated.” The majority of these reservoirs continue to exist in an undersaturated state during production because of aquifer support or water flooding projects done to sustain the reservoir pressure (Pr) to always exceed Pb. Hence, the need for a reliable predictive model or tool to predict undersaturated oil viscosity has led several authors such as Refs. [18–29] that developed empirical models currently in use in the petroleum industry. For instance, Shokir and Ibrahim

[30] presented a new undersaturated crude oil viscosity model using multi-gene genetic programming (MGGP). Their results specified that the new MGGP-based model is useful in the prediction of undersaturated oil viscosity. A computer-based model, MLP-NN, was designed by Moghadam et al. [31] for the estimation of the viscosity of dead oil, saturated oil and undersaturated oil. The result outcomes indicated the precision and reliability of the method. Sinha et al. [32] developed two different approaches to estimate the dead oil viscosity form. The first approach was based on a trainable explicit expression that uses a reference viscosity measured at any temperature. Our second method utilized a hybrid machine learning method. Their approach uses a richer dataset with very limited input parameters. Beal [19] proposed the use of a graphical correlation with a mean deviation of 2.7% used to estimate undersaturated oil viscosity been a dependent variable of Pr, Pb and viscosity oil at Pb. Vazquez and Beggs [25] utilized numerous sets of PVT-measured values for their model which had –7.541% as the mean percentage error. Petrosky and Farshad [13] utilized data from the Gulf of Mexico for their correlation which had an outcome of –0.19% for relative error and 4.22% for standard deviation. Kartootmodjo and Schmidt [33] correlation is a modification of Beal's correlation with –4.29% as the mean error. Elsharkawy and Alikhan [23] correlation had 1.2% for mean absolute relative error and 0.022 for its deviation [2]. These techniques of estimating crude oil viscosity were mostly based on applied mathematical approaches and are able to estimate the viscosity of the oil based on API gravity of oil, oil formation factor (Bo), solution GOR and Tr. In addition, the correlations had an error ranging from 25 to 40% and were mostly constrained to a particular location from which the sample data were collected; hence, their accuracy level was limited as such they could not be generalized [34].

To overcome most of these challenges in the correlation of oil viscosity, soft-computing and machine learning-based computational models were adopted in estimating the viscosity of crude oil. Linear regression, artificial neural networks, support vector machines, decision tree and so many other machine learning methods (see [35–44]) have been reported to have performed more efficiently than conventional empirical correlations. A study that involved predicting the viscosity of crude oil samples from Nigeria using a neural network with 0.99 as the coefficient of correlation presented an improved performance when compared to already developed empirical correlations. Also, studies on using neural networks to estimate the viscosity of crude oil samples in Iran and Oman have performed better than existing correlations. A report of studies on estimating the viscosity of crude

oil in Canada using neural networks and support vector machine was also accurately predicted, therefore affirming the applicability of these algorithms in estimating the viscosity of crude oil is reliable [45].

Consequently, learning techniques that combine different machine learning algorithms that ally to improve predictions (stacking) or decrease variance (bagging and bias boosting) were further adopted in view that the outcome will be better than single-based machine learning classifiers and regressors. These learning techniques were referred to as ensemble learning techniques. A machine learning paradigm in which numerous learners are trained to tackle the same issue is known as ensemble learning [46]. Unlike traditional machine learning algorithms that attempt to learn a single hypothesis from training data, ensemble methods attempt to create a number of hypotheses and aggregate them for usage. Ensemble methods were discovered to be more efficiently applied to datasets of points as over-fitting was greatly avoided in the course of applying these techniques [45–48]. Dietterich [49] outlined three key arguments for employing an ensemble-based system in his 2000 review article: i) statistical; ii) computational and iii) representational. Succinctly, ensemble learning techniques have found application in several fields and though has started gaining the attention of researchers in the petroleum engineering field. Santos et al. [50] utilized a neural network ensemble in identifying lithofacies. Gifford and Agah [51] utilized techniques that classify multiagents for lithofacies recognition. Masoudi [52] was able to identify Sarvak productive zone formation by integrating the outputs from fuzzy and Bayesian classifiers. Davronova and Adilovab [53] did a comparative analysis of the ensemble methods for drug design. Some recent works that have deployed ensemble learning could be found in Refs. [54–61].

Researchers in the petroleum industry have also found ensemble learning as a veritable tool to make phenomenal changes in the business. Anifowose et al. [62] characterized reservoir by applying the stacked generalization ensemble in order to improve supervised machine learning algorithms capabilities in prediction. Anifowose et al. [63] recommended that ensemble techniques find relevant applications in the oil and gas industry in his review work on the application of the ensemble method. Bestagini et al. [64] classified Kansas oil-field data lithofacies by applying the random forest ensemble method. Xie et al. [65], and Tewari and Dwivedi [66] presented a study which compared the efficiency of ensemble methods applied in recognizing lithofacies. Tewari and Dwivedi [67] also did a comparison work on lithofacies identification. Bhattacharya et al. [68] utilized ANN and RF learning techniques in the prediction of the daily production

of gas from the unconventional reservoir. Tewari [69, 70] utilized ensemble methods to estimate the recovery factors of various reservoirs. The accuracies reported in these applications supported the argument that the application of ensemble classifiers or regressors was more precise when compared to the outcome of the discrete classifiers or regressors. Not quite much has been known concerning the comparison of ensemble regressors in predicting the PVT property of reservoir fluids such as oil viscosity. Given these observations, this study seeks to compare single-based machine learning techniques and ensemble learning techniques in predicting undersaturated oil viscosity.

The outcome of this paper brings more clarity to the applications of diverse ensemble techniques in predicting oil viscosity. Again, it helps the petroleum/reservoir engineer with an optimum predictive modeling tool for their tasks as a wide range of methods have been developed and harnessed. It also exposes additional tools and techniques to improve accuracy in the prediction of oil viscosity. Finally, to the best of authors' knowledge, this work has presented a more robust comprehensive comparative analysis between single-based machine learning techniques and ensemble learning techniques than it has been in previous literature.

2 Methods

Several intelligent algorithms and optimization techniques [71–85] have been deployed to solve various boundary value problems in physical sciences. This section contains a brief explanation of machine learning methods implemented in estimating undersaturated oil viscosity.

2.1 Single-based machine learning algorithms

This study adopts four (4) baseline machine learning techniques, namely ANN, SVM, DT and LR, based on their superiority and widespread applications in estimating PVT properties of reservoir fluid in research.

- (1) Decision tree (DT) algorithm: This algorithm is in the form of a tree-like structure that makes use of a branching methodology in clarifying each likely single result for a particular prediction/decision. Decision tree in recent times has been increasingly applied for various classification tasks because of their simplicity, ease of interpretation, low-slung cost of computation and graphical representation ability. The appropriate property for each generated tree node uses an information gain approach and any attribute that has the maximum information gain is selected as the current node for the test attribute. The operation of a decision tree algorithm

on a dataset (DS) is expressed below. Firstly, the entropy value estimate of DS is presented mathematically in Eq. 1 as,

$$E(S) = \sum_{i=1}^m -p_i \log_2 p_i \tag{1}$$

where $E(S)$ is the entropy of a DS collection, p_i represents the number of proportion instances which belongs to the class i and m represent the number of classes in the system. Secondly, information gain (IG) for a particular attribute, for instance, K in a collection S is presented mathematically in Eq. 2 as,

$$G(S, K) = E(S) - \sum_{v \in \text{invalues}(K)} \frac{S_v}{S} E(S_v) \tag{2}$$

where S_u is the set of value v for attribute K instances.

- (2) Support vector machine (SVM) algorithm: SVM is also a type of supervised machine learning technique which can be applied for both classification and regression tasks. It serves as the linear separator inserted between two data nodes to identify two different classes in environs composed of multiple dimensions. The implementation of SVM is as follows. Assume DS to be the training dataset, $DS = (x_i, y_i, \dots, (x_n, y_n)) \in X, R$. In support vector machine, DS is represented as points in an N -dimensional space and then attempts to develop a hyperplane that will divide the space into specific class labels with a right margin of error. Support vector machine optimization algorithm is represented mathematically in Eq. 3 as,

$$\begin{aligned} &\underset{d, \omega}{\text{minimize}} \quad \frac{1}{2} Y^T Y + C \sum_{i=1}^n \omega_i \\ &\text{subject to} \quad z_i \left(Y^T \theta(u_i + b) \geq 1 - \omega_i \right), \omega_i > 0 \end{aligned} \tag{3}$$

Mapping θ of vectors $u_i(DS)$ to a higher-dimensional space. The support vector machine then locates best-fit margin that divides the hyperplane linearly in this dimension. The formulation of the kernel function is presented as $K(u_i, u_j) \equiv \theta(u_i)^T \theta(u_j)$. The radial basis function (RBF) kernel adopted in this study can be expressed mathematically in Eq. 4 as,

$$\text{RBF} : K(u_i, u_j) \equiv \exp(-z \|u_i - u_j\|^2), z > 0 \tag{4}$$

where $(u_i - u_j)$ is the Euclidean distance between two data points.

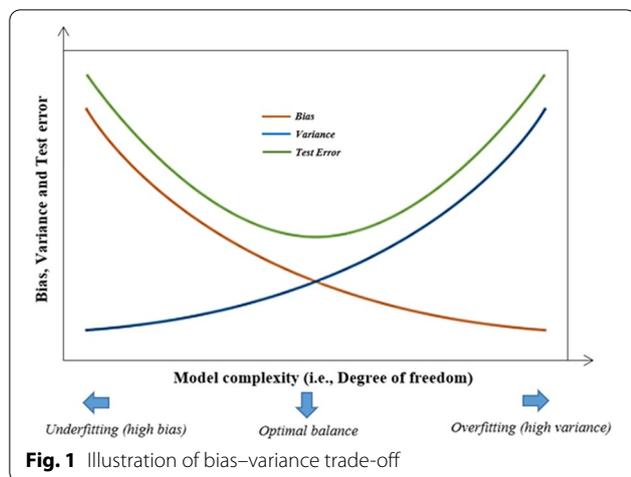
- (3) Artificial neural networks (ANN) algorithm: ANN also referred to as neural network (NN) is a connection of interdependent layers that receives input, initializes it and passes it over to the next available component. The multilayer perceptron (MLP) for the ANN was adopted in this work. Multilayer perceptron maps a function $f(\cdot) : R^D \rightarrow R^0$, and dataset (dS) is trained, where (d) stands for dS input dimension, and 0 represents the number of dimensions of required output. Assume U is a set of features, and z expected output, taking $U = \{u_1, u_2, u_3, \dots, u_D\}$, the MLP now learns the nonlinearity of the function approximators for classification and regression tasks. Stochastic gradient descent, limited-memory-Broyden Fletcher-Goldfarb-Shanno (*lbfgs*) or adaptive moment estimation (ADAM) can be used in training multilayer perceptron but the Tikhonov regularizer and Adam optimizer have been utilized in this work. In each layer, the activation function that was adopted in this work is rectified linear unit (ReLU) function. For each layer, the mapping functions can be expressed as Eq. 5, while the algorithm for back-propagation was utilized to train the *MLP*.

$$Z^l = \omega^{(l)T} \times \alpha^{(l-1)} + \beta^{(l)} \tag{5}$$

ω^l is the weight matrix, β^l is the bias and α is the weighted input sum.

2.2 Ensemble learning methods (EMs)

EM is an approach in machine learning in which combined models (usually referred to as “weak learners”) that have been trained to solve a particular problem with their results aggregated to yield more improved output [86]. Weak models when appropriately combined yields more robust and/or accurate models which is the primary hypothesis upon which ensemble learning is based on. Thus, ensemble methods accentuate the strength and reduce the weakness of the single classifiers or predictor. In ensemble method, diverse single classifiers or regressors are independently trained using same or differing dataset though not having the same variables [87]. The target output is derived by determining the mean of the individual single-based classifier/regressor output. Figure 1 shows an illustration of bias–variance trade-off. To be able to solve a problem, the degree of freedom of the proposed base learning model should be sufficient to solve any underlying complexity of the dataset been worked upon in addition, the degree of freedom required



should also avoid high variance so the model could be robust in both classification and regression tasks. This condition is referred to as bias–variance trade-off. The following factors need careful consideration when applying ensemble classifier and regressors models; difficulties in identifying the most suitable classifier or regressor for a particular application domain due to the numerous available methods, the amount of single regressors or classifiers to combined for improved accuracy as well as combination technique that will be most suitable for the different single classifiers and regressors to yield a better outcome. Therefore, presented below are discussions of few techniques for ensemble learning. Ensemble learning techniques can be grouped into diverse powerful methods such as bagging, boosting, voting, stacking and so on.

(1) *Bagging approach* Bagging stands for bootstrap aggregating which mostly finds its application in regression and classification tasks. Bagging increases models' accuracy by adopting decision trees techniques by reducing the variance to an optimum point. One of the major challenges of most predictive models is over-fitting but by reducing variance which invariably increases accuracy this challenge is eliminated. The advantage of bagging is that most often the learners with a weak base combine to form a more stable single learner while its high computational cost is one of its major limitations. In addition, when the appropriate procedure for bagging is ignored, it could result in more bias in models. There are two classifications of bagging, namely bootstrapping and aggregation. In bootstrapping, samples are gotten from the entire populace (set) by a procedure called the replacement method. The replacement method aids in making the selection process to be randomized.

To complete the procedure, the base learner runs on the samples. In aggregation, all possible prediction outcomes are usually incorporated and randomized. Aggregation is a function of the probability of the bootstrapping procedures or based on all the results of the models used in prediction.

- (2) *Boosting approach* Boosting learns the mistakes of the past predictor for improved and enhanced future predictions. This approach combines many base learners that are weak to obtain one strong learner, which improves the model's predictability significantly. In boosting, weak learners are arranged in sequence so that each of them learns from the preceding one in the sequence, thereby resulting in a better predictive model. The different forms of boosting include gradient boosting, adaptive boosting (also referred to as AdaBoost) and extreme gradient boosting (also referred to as XGBoost). AdaBoost uses weak learners in the form of decision trees, with one split that is mostly referred to as decision stumps. This decision stump contains several observations of like weights. In gradient boosting, predictors are added sequentially so that leading predictors correct their upcoming predictors targeted at improving the model's accuracy by canceling out the effect of the predictor's errors in the previous ones, the new predictors are usually fitted for this purpose. The gradient booster identifies the problems in the predictions of the learners through the help of gradient descent, and counters them necessary. In XGBoost, decision trees are used alongside boosted gradient to improve the speed and performance of the learner. XGBoost depends mainly on the speed of computation and the target model's performance.
- (3) *Voting approach* A voting ensemble is also called a majority voting ensemble. It is an ensemble technique which combines multiple models' predictions and is used to improve the model performance of any single model. Voting can be applied for both regression and classification tasks. Regression tasks usually require computing the mean of the model's predictions; hence, predictions are the average of contributing models. For classification tasks, each label's prediction is usually presented so that the predicted one is the most voted label. The voting ensemble may be in the form of a meta-model or a model of models. The term "meta-model" describes a set of existing trained models while the existing models would not be aware of their usage in the ensemble. A voting ensemble is better applied if two or more models with good and aggregable predictions.

In a nutshell, ensemble approaches combine numerous models rather than using just one in order to increase the accuracy of outcomes in models. The combined models considerably improve the accuracy of the outcomes. This has increased the acceptance of ensemble techniques in machine learning. By integrating numerous models into one highly dependable model, ensemble approaches seek to increase predictability in models. The three most common ensemble techniques are stacking, bagging and boosting. For regression and classification, ensemble approaches work best because

they lower bias and variance and increase model accuracy.

2.3 Study framework

Figure 2 shows the framework adopted in this research work which implemented single-based machine learning algorithms (ANN, SVM, DT and LR) and ensemble learning algorithms (bagging, boosting and voting) using the ANN, SVM, DT and LR for predicting undersaturated oil viscosity followed by an evaluation of their accuracies and error measurement, respectively. The

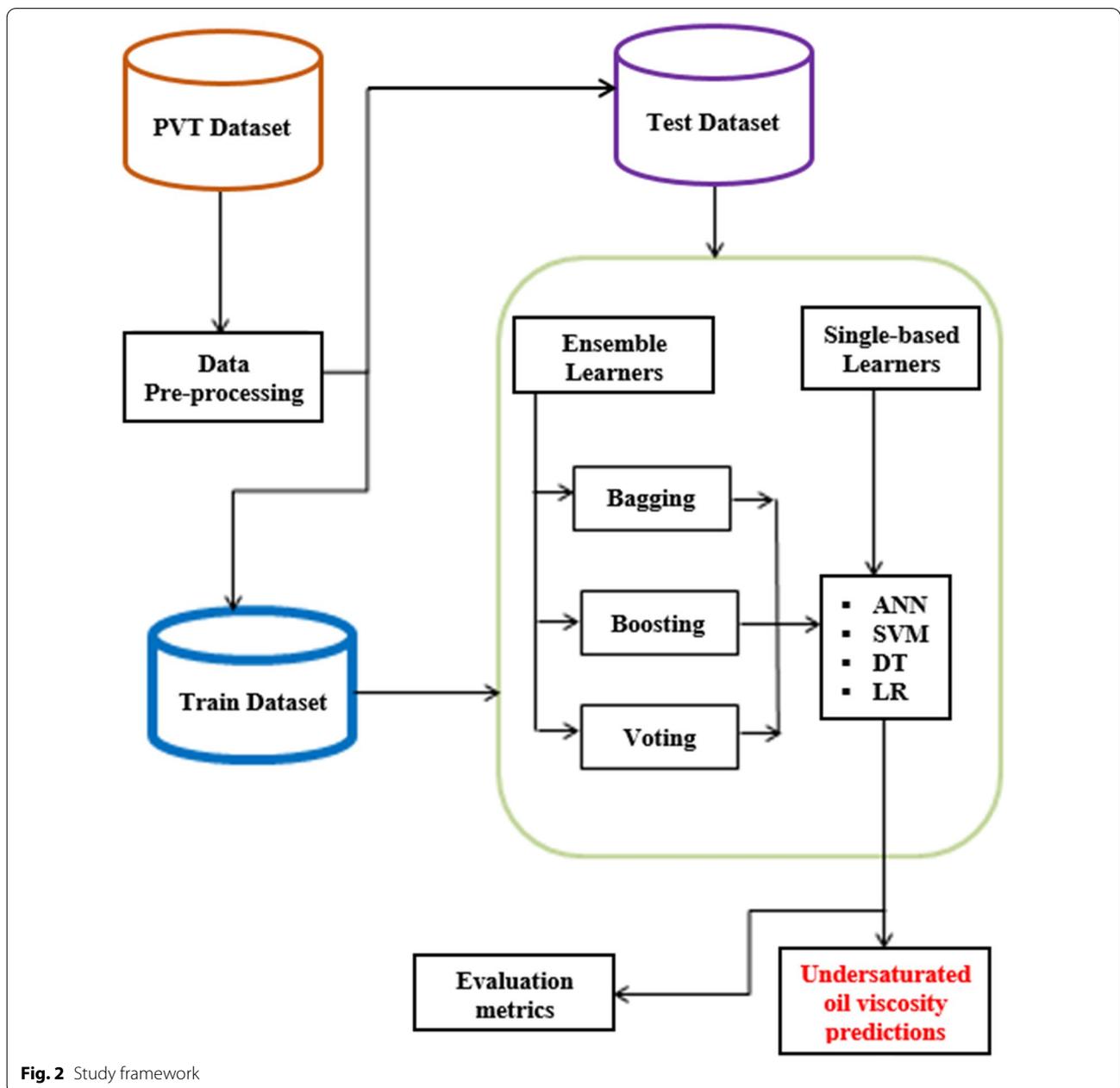


Fig. 2 Study framework

procedure for this study shown in Fig. 2 is classified into three main phases, namely the data preprocessing stage, the model building stage which includes both the single-based methods and the ensemble methods, respectively, and finally the comparison of the models' accuracies and error metrics.

3 Research data

Experimental PVT dataset of fifteen (15) bottom-hole samples collected from different Niger-Delta oil fields in Nigeria yielding a total of one hundred and three (103) data points were used in the models. After running the training and testing scenarios on the dataset for both single-based learners and a sample of the ensemble methods, for model training, 70% of the total dataset was utilized while the remaining 30% was utilized to predict undersaturated oil viscosity values. This research involves experimental PVT data of fifteen (15) bottom-hole samples taken from different Niger-Delta oil fields in Nigeria. These reservoirs were still in the "undersaturated" condition as some were been produced with the support of water flooding while others were at the early stage of initial production when the sample was collected for analysis in the laboratory. The bottom-hole samples were flash separated to derive the API gravity of oil, relative density of gas and solution GOR values. The viscosity data were obtained by rolling ball viscometer at varying pressures. Data from four reservoirs were used in this study. The total selected data points are one hundred and three (103). The range of values for the data is outlined in Table 1.

4 Model setup and evaluation

Scikit-learn library [88] and Python programming language was used to build a total of twenty-three (23) different models in this study of predicting undersaturated oil viscosity which comprises of four (4) using single-based machine learning algorithms, five (5) using bagging ensemble method, six (6) using boosting ensemble method and eight (8) using voting ensemble method. The base

learners' parameters were set as follows: For the ANN, multilayer perceptron was adopted with thirteen (13) hidden layers (HL). The maximum iteration was set to 2000, *optimizer = lbfgs*, *activation = ReLU* activation function. For SVM, the RBF kernel was utilized. The decision tree method setting, *criterion = MSE*. These were implemented on an intel corei7 64 bits with 64 GB memory laptop.

The determination of the performance between ensemble methods and single-based machine learning techniques involved using these diverse approaches to estimate the measured dataset for the undersaturated oil viscosity property. Followed by a comparative analysis between the outcome of ensemble methods and single-based machine learning methods to measure their agreement with experimental data to ascertain the efficiency of each of the approaches. The model performances were assessed under the design format of 70% training dataset and 30% testing data. To evaluate the model performance, five (5) widely known evaluation metrics were used, namely five evaluation measures, namely MAE, R^2 , MSE, RMSE and RMSLE and their formulas are shown in Eqs. 6–10.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |z_i - u_i| \quad (6)$$

$$R^2 = \frac{\sum_{i=1}^n ((u_i - \hat{u})(z_i - \hat{z}))}{\left(\left(\sum_{i=1}^n (u_i - \hat{u}) \right)^2 \left(\sum_{i=1}^n (z_i - \hat{z}) \right)^2 \right)^{\frac{1}{2}}} \quad (7)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n (z_i - u_i)^2 \quad (8)$$

$$\text{RMAE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (z_i - u_i)^2} \quad (9)$$

$$\text{RMSLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(u_i + 1) - \log(z_i + 1))^2} \quad (10)$$

z_i are predicted values, u_i are observed values, \hat{z} is the mean predicted values, \hat{u} is the mean observed values, and n is the sum of the instances. A regression model's accuracy will be higher if its MAE, MSE and RMSE values are lower. However, a greater R^2 value is seen as preferable. The effectiveness of a linear regression model in fitting a dataset is quantified by all evaluation metrics, including MAE, MSE, RMSE, RMSLE and R^2 . While R^2 indicates how well the predictor variables can account for

Table 1 Statistical description of the PVT data

PVT property	Minimum value	Maximum value
Oil viscosity above bubble point (cp)	0.15200	0.36390
Pb (psia)	2588	4042
GOR (SCF/STB)	500	2073
Gas gravity	0.354	0.972
API	38	43
Temperature, $^{\circ}F$	177	240

the variation in the response variable, MAE, MSE, RMSE and RMSLE indicate how well a regression model can predict the value of a response variable in absolute terms. The ability of a system or model to maintain stability and undergo only little (or no) changes when subjected to noisy or inflated inputs is known as robustness. Therefore, an outlier's impact on a robust system or measure must be reduced. Since the squaring of the errors will place higher importance on outliers in this scenario, it is simple to conclude that some evaluation measures, like MSE, may be less robust than others, like MAE. Then, using RMSE, the MSE error is square-rooted to return it to its original unit while keeping the characteristics of penalizing higher errors. As such, the five measures were deployed to take care of both the robust and non-robust systems.

The entropy value estimate of the dataset (DS) as shown in Eq. 1 essentially informs us how impure a set of data is. Non-homogeneity is described here by the term "impure." In other words, homogeneity is measured by entropy. It provides information on an arbitrary dataset's impureness and non-homogeneity. The entropy of a set of instances or dataset, DS , which includes examples of a target idea in both positive and negative light with respect to this Boolean categorization is given in Eq. 11 as,

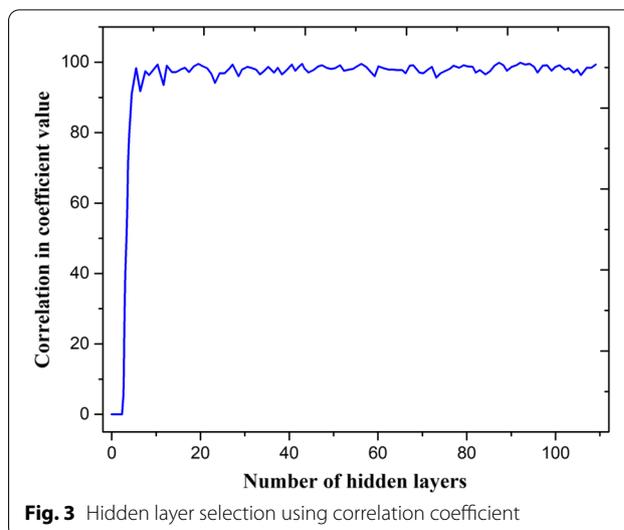
$$\text{Entropy}(DS) = -(P_{\oplus} \log_2 P_{\oplus} + P_{\ominus} \log_2 P_{\ominus}) \quad (11)$$

P_{\oplus} is the portion of positive examples and P_{\ominus} is the portion of negative examples in DS . Summarily, when the data collection is entirely non-homogeneous, the entropy is highest; when it is homogeneous, it is lowest.

5 Results

To ascertain an optimal number of hidden layers for multilayer perceptron/ANN to be used, a range of experiments was carried out by testing 1 to 109 hidden layers. Figure 3 shows that the ANN achieved the highest accuracy ($R^2 = 0.990$) at different hidden layers. However, in this study, thirteen (13) hidden layers were adopted because the error values obtained using thirteen (13) hidden neurons were considerably low (MAE = 0.0009, MSE=1.20-e06, RMSE=0.001095) when compared to other numerous hidden layers that had the same value for Coefficient correlation with it.

Model proficiency evaluation formed the first part of the analysis. The correlation coefficient (R^2) is calculated for each of the models under different train/test conditions. Based on the prediction performance gotten, as well as consideration of the recent application of some specific single-based machine techniques in the petroleum industry in predictive modeling, only LR, SVM,



ANN and DT out of five regressors were chosen to ascertain the optimal training/testing scenario. Altogether, five different training and testing scenarios based on different training and testing dataset ratios were tried and tested for each single-based machine learning algorithm and bagging ensemble method as shown in Figs. 4 and 5. As can be observed in Fig. 4, the representation of MLP in almost all scenarios has been good, with the highest correlation coefficient equal to 0.99946 under 70–30% scenario which no other single-based machine learning technique could achieve. On the other hand, the performance of SVM was the lowest with correlation coefficient equal to -0.09743 under 90–100% scenario. The correlation coefficients obtained by each of the techniques were high except for SVM which was extremely low. However, the consistency of ANN to stand out in all the scenarios is quite noticeable followed by LR and then DT which were at a close range to ANN. Figure 5 reveals that the correlation coefficient values of bagged-LR, bagged-DT and bagged-ANN are considerably close ties with just a minute difference across all the five different training and testing scenarios based on different training and testing dataset ratios. It is also worthy of noting that the correlation coefficient values of DT, LR and ANN of the single-based algorithm were improved upon in the bagging ensemble method with the most conspicuous being the DT. Yet 70–30% of training and testing scenarios gave the highest correlation coefficient equal to 0.99964. It is based on this result that the dataset ratio adopted throughout this work is 70% for training and 30% for testing.

After training all the 23 models built in this work using both the single-based machine learning techniques and ensemble method techniques with 70% of the dataset, the

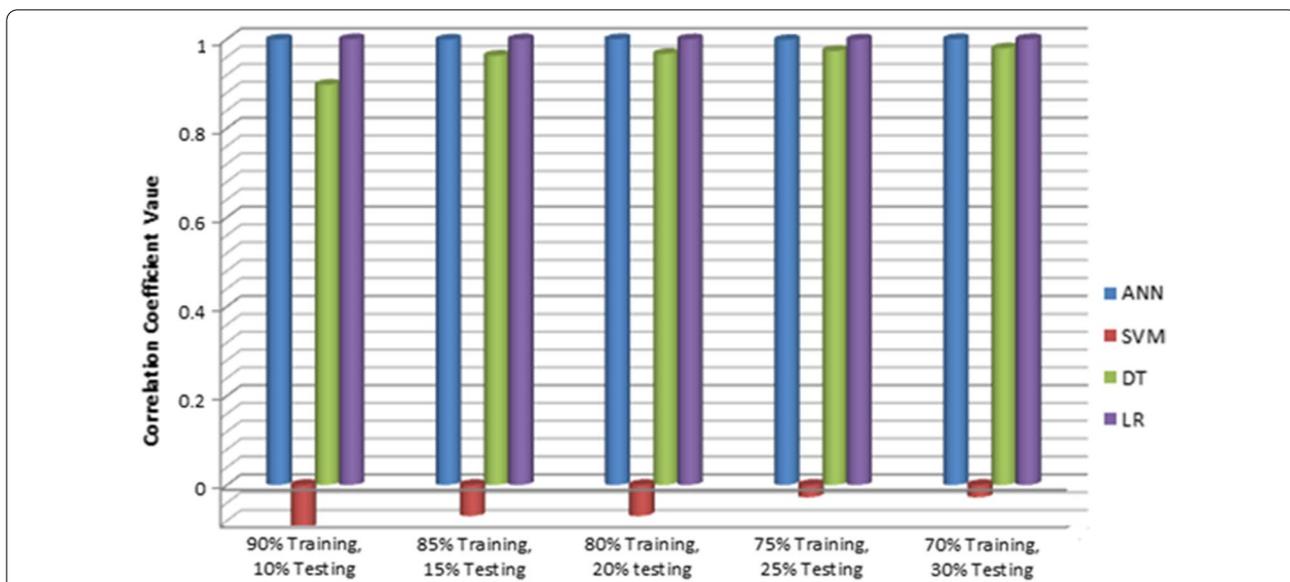


Fig. 4 Training and testing scenario for single-based machine learning algorithms

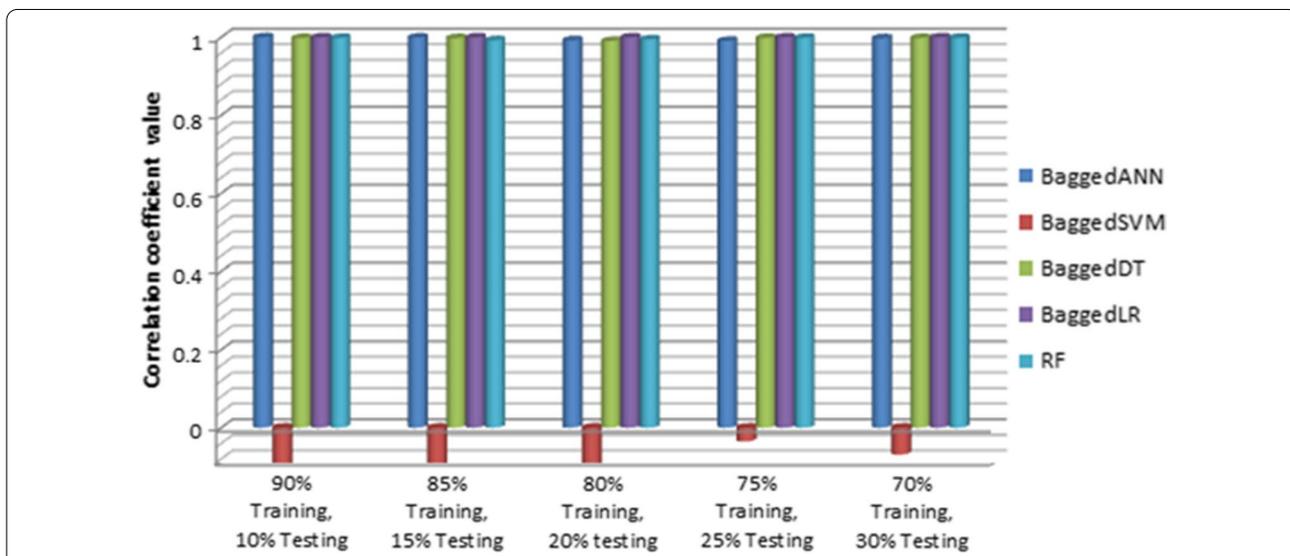


Fig. 5 Training and testing scenarios for bagging ensemble approach

models were up to be tested and evaluated. The second part of the analysis is the comparison of individual performance of the single-based machine learning methods. This was achieved by utilizing the last data group comprising 30% of the total dataset (29 data points) gathered for this work, which was not seen by any of these developed models during training was used. Table 2 presents the prediction performance of the single-based machine learning algorithm on the 30% test data used in this work.

Table 2 Prediction performance of different single-based machine learning algorithms

Regressors	R ²	MAE	MSE	RMSE	RMSLE
ANN	0.99665	0.00338	1.81914e-05	0.00042	0.00346
SVM	-0.01955	0.06963	0.00523	0.07237	0.05713
DT	0.96826	0.00806	0.00015	0.01239	0.01033
LR	0.99589	0.00209	5.77396e-06	0.00240	0.00188

From the information provided in Table 2 and comparing the evaluation metrics of each of the methods, SVM has the worst prediction performance for the dataset used in this work, while the other three algorithms performed significantly well with ANN prediction performance topping every other single-based algorithm having the highest (R^2) value (0.99665) which is very close to 1.00 depicting interdependence of variables as well as the lowest RMSE value (0.00042).

For an in-depth comparative analysis, the different ensemble methods (bagging, boosting and voting) were applied using the single-based machine algorithms (ANN, SVM, DT and LR) under study as base learners to further reveal the difference in prediction performance of the single-based machine learning techniques and the ensemble learning techniques. Tables 3, 4 and 5 present the prediction performance of the various ensemble methods under study. From the information provided in Table 3, the application of the bagging approach of ensemble technique resulted in significant improvement in the R^2 value and reduction of MAE, MSE and RMSE values, respectively. However, R^2 value of the bagged-SVM slightly improved when compared to its corresponding value in Table 2 when it is a single-based algorithm. Still, bagged-SVM performed the poorest when compared to other bagged base learners. This is obvious from the negative value of the R^2 value (-0.14575) which depicts that this model could decipher the interdependent relationship of the variables in the dataset. Also under the bagging procedure of ensemble learning, bagged-LR stands out with the highest value of R^2 value (0.99894) which is close to 1.00 and the lowest error evaluation metrics values (MAE = 0.00204, MSE = 5.723E-06, RMSE = 0.00456) across board; all indicating an excellent prediction. Its comparison with the prediction performance of single-based LR also reveals significant improvement.

From Table 4, the application of the boosting approach of ensemble technique on the base learners resulted in AdaBoost-LR having the best prediction performance with R^2 value (0.99930) and very low errors evaluation metric values (MAE = 0.00165, MSE = 3.76101e-

Table 3 Prediction performance of bagging approach type of ensemble technique

Bagged regressors	R^2	MAE	MSE	RMSE	RMSLE
Bagged-ANN	0.99744	0.00325	1.39161e-05	0.000373	0.00293
Bagged-SVM	-0.14575	0.03888	0.00153	0.03920	0.03288
Bagged-DT	0.99616	0.00386	2.08562e-05	0.00456	0.00366
Bagged-LR	0.99894	0.00204	5.72352e-06	0.00239	0.00187
Random-Forest	0.98310	0.00607	9.19762e-06	0.00959	0.00783

Table 4 Prediction performance of boosting techniques of ensemble method

Boosting regressors	R^2	MAE	MSE	RMSE	RMSLE
Gradient boosting	0.97881	0.006871	0.00011	0.01073	0.00875
XGBoosting	0.97864	0.00683	0.00011	0.01078	0.00879
AdaBoost-MLP	0.99943	0.00143	3.08633e-06	0.00175	0.00136
AdaBoost-SVM	-0.03092	0.07292	0.00561	0.07490	0.05894
AdaBoost-DT	0.97782	0.00704	0.00012	0.01098	0.00895
AdaBoost-LR	0.99930	0.00165	3.76101e-05	0.00193	0.00150

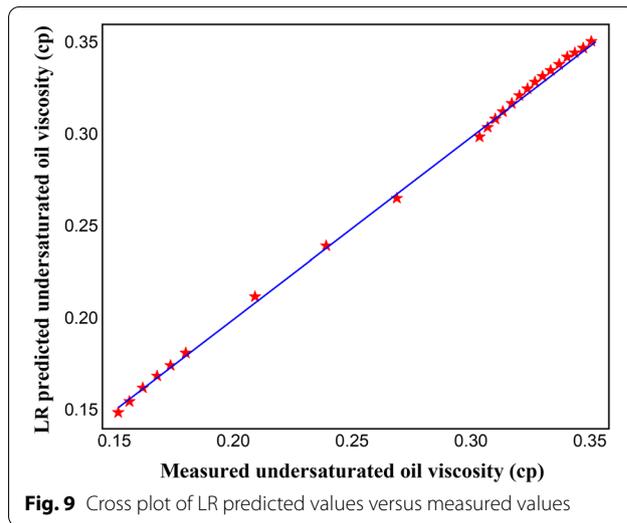
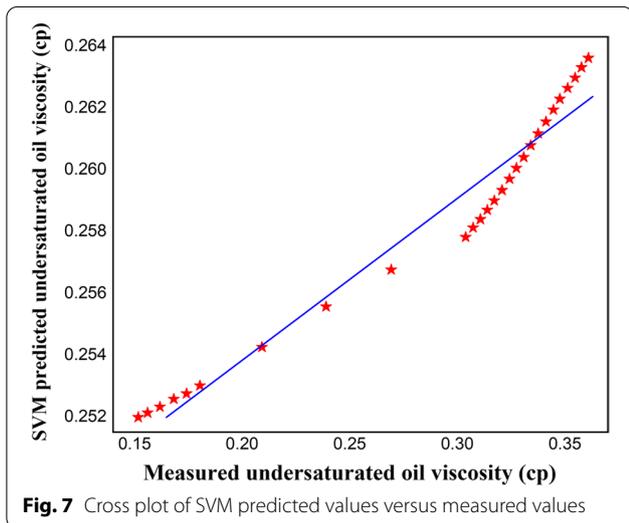
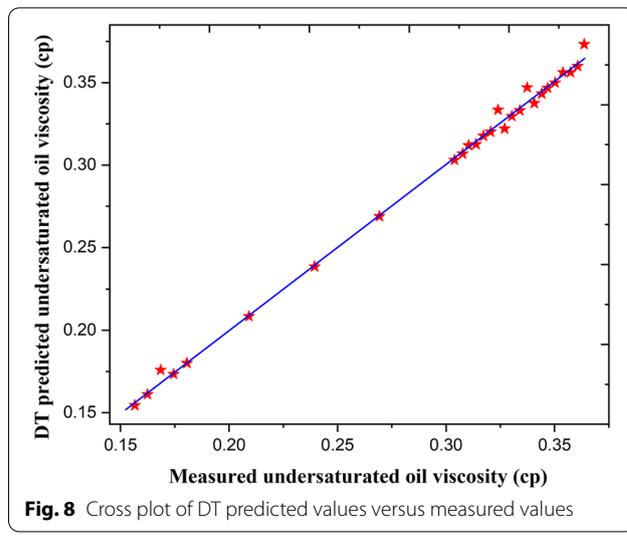
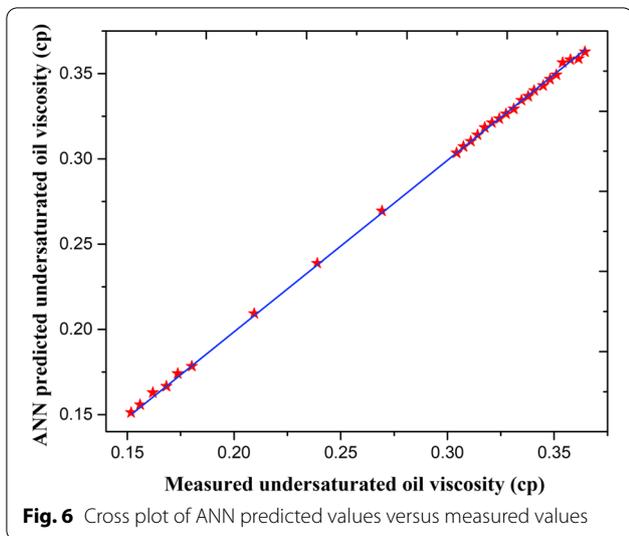
05, RMSE = 0.00193). It is also observed that boosting approach performed better than the bagging approach with the LR, ANN and SVM single-based algorithm as base learners.

From Table 5, the application of the voting approach of ensemble technique on the combination base learners further enhanced improvement in prediction performance. The most conspicuous of them all is SVM, which is a single-based algorithm with very low and negative R^2 values, indicating its weakness in prediction. When combined with other single-based algorithm in voting, significantly rose with R^2 values (0.73399, 0.87540, 0.88078, 0.88015 and 0.92915) for each of the combinations in voting approach.

Figures 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 and 27 illustrate cross plots of the predicted undersaturated oil viscosity values versus the experimental/measured undersaturated oil viscosity values. The cross plot reveals the level of coherence between the measured and the estimated values. A perfect agreement should have all the points lie on the 450 line on the cross plot. In every instance, all the graphs showed the cloud of tightest points around the 450 lines. This signifies an accepted agreement apart from Figs. 7, 11 and 17 representing the cross plots for

Table 5 Prediction performance of voting approach type of ensemble technique

Voting	R^2	MAE	MSE	RMSE	RMSLE
ANN,SVM	0.73399	0.03704	0.00144	0.03804	0.02977
ANN,DT	0.99446	0.00312	3.01229e-05	0.00548	0.00448
ANN,LR	0.99899	0.00189	5.52212e-06	0.00234	0.00193
ANN,SVM,DT	0.87540	0.02556	0.00067	0.02604	0.02043
ANN,SVM,LR	0.88078	0.02490	0.00064	0.25472	0.01993
SVM,LR,DT	0.88015	0.25085	0.00065	0.02553	0.02007
ANN,DT,LR	0.99710	0.00196	1.57588e-05	0.00396	0.00326
ANN,SVM,DT,LR	0.92915	0.01932	0.00038	0.01963	0.01542



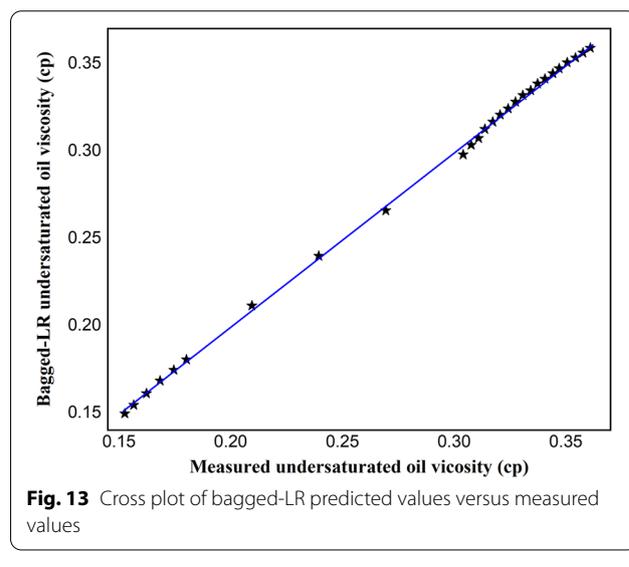
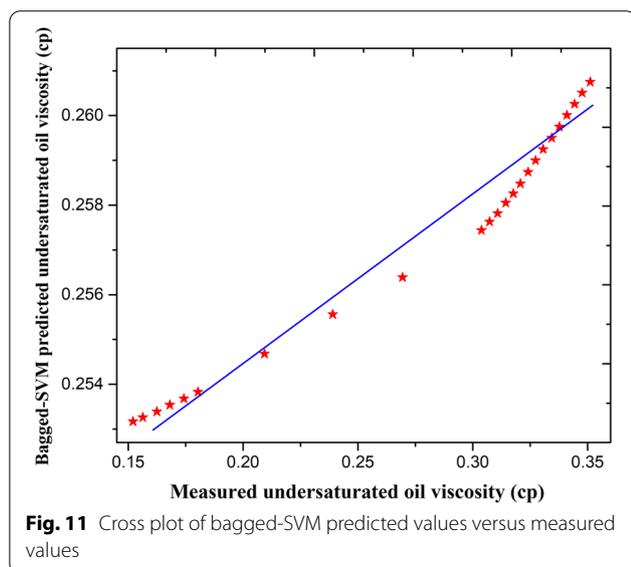
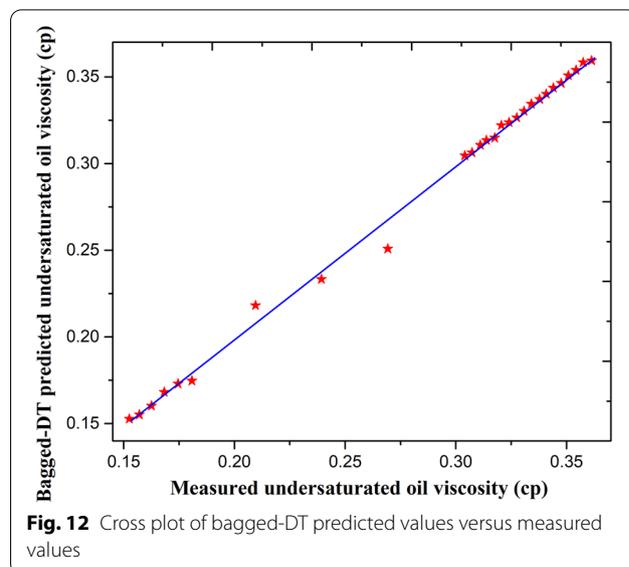
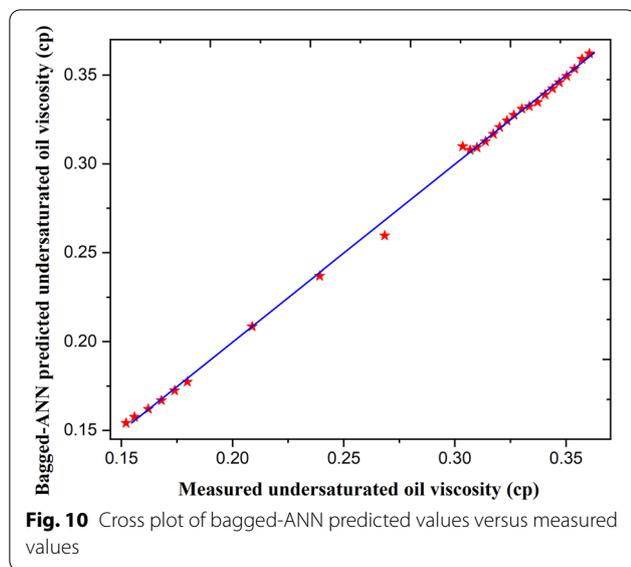
SVM, bagged-SVM and AdaBoost-SVM, respectively. The non-fitness of the points on the 450 line reveals a poor agreement between the measured values and predicted values. Hence, for this dataset, we can conclude that those respective models are weak and unreliable. However, this situation was the difference in the voting approach of the ensemble method as all the combinations involving SVM turned out a good fit as can be found in Figs. 20, 23, 24, 25 and 27. The fitting of the data points on the 450 lines reveals the accuracy of the predicted results using these approaches. This finding indicates that ensemble methods with voting in this particular case have the potential of boosting weak single-based machine learning algorithms. This is evident the way the previous weak single-based SVM could

produce an accurate prediction of the undersaturated oil viscosity having combined with other techniques using the voting ensemble method.

6 Discussion

6.1 The predictive models

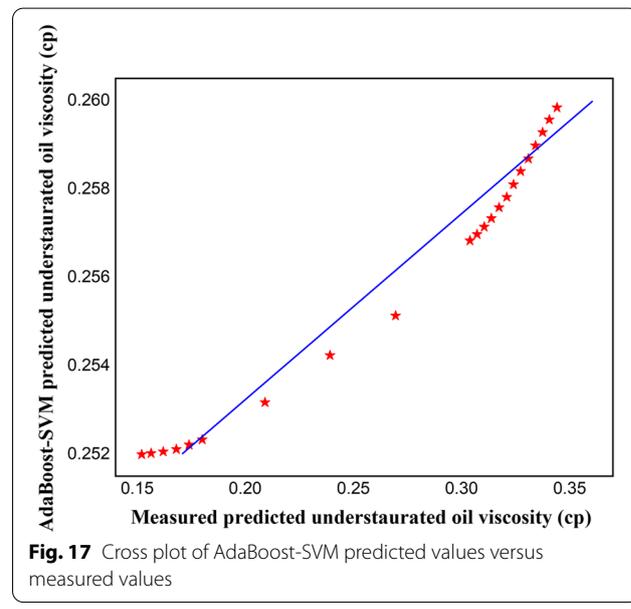
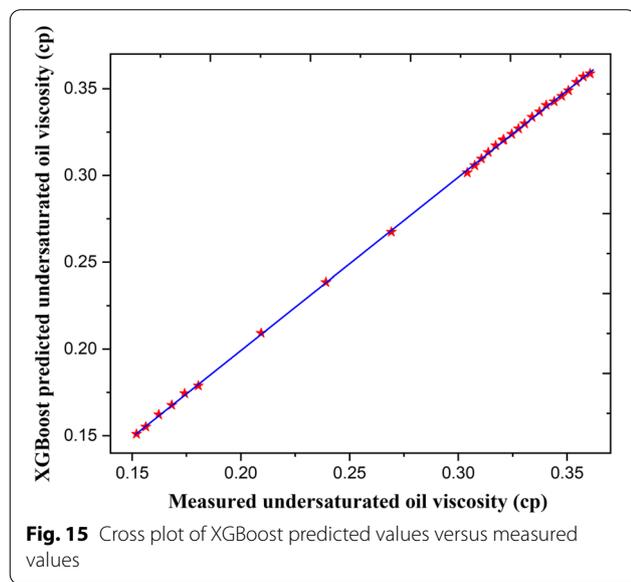
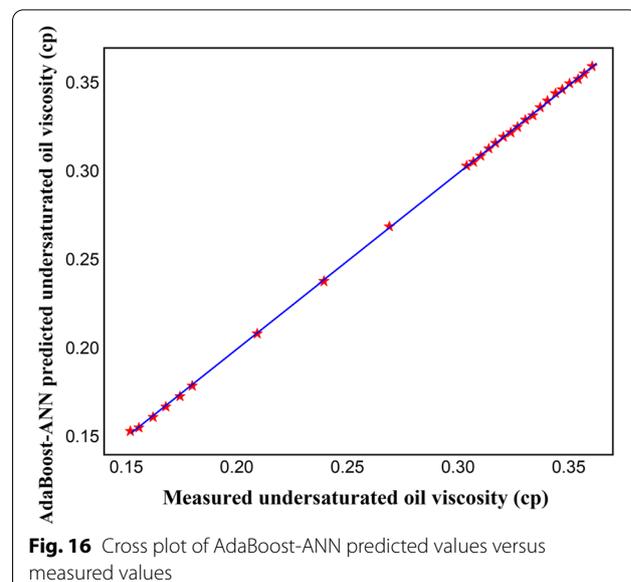
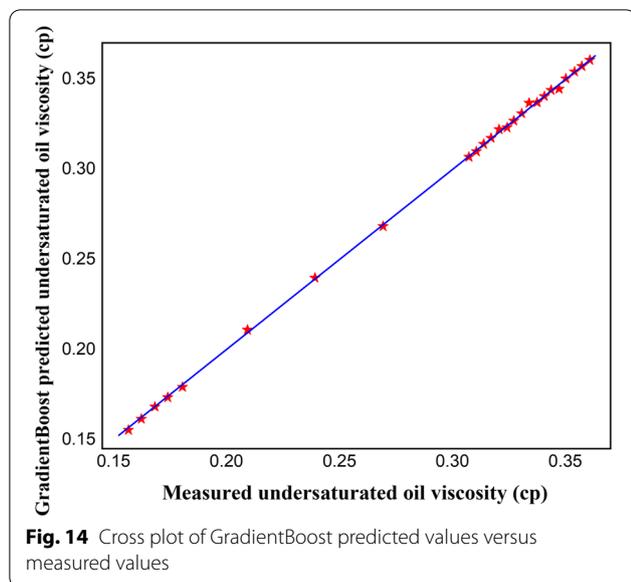
The error evaluations using the five measuring metrics reveal that the ensemble learners had the lowest MAE, MSE, RMSE and RMSLE values and highest R^2 values when compared with the single-based algorithms. Tables 4 and 5 show that the approach described in this work is suitable for predicting undersaturated oil viscosity, with the bagged ensemble method providing the best solution. The Al-Kafaji [89] and Bennison [90] procedures are ineffective for crude oil gravities less than a



particular threshold. The most frequent approach for obtaining oil viscosity data is viscosity correlation, which is extremely helpful and successful in estimating oil viscosity at various temperatures for various oil kinds. The limitation on the parameters from which these correlations have been formed is the essential element in their implementation. As a result, these connections are region-specific and cannot be generalized worldwide.

According to the literature, measuring heavy oil viscosity at low temperatures might be difficult since the predicted values of the viscosity frequently surpass the maximum limit of the equipment. This model is suitable for a wide range of oil viscosities and temperatures.

The suggested oil viscosity correlation does not need any compositional investigation of the oil or asphalt content, which is a significant benefit of the ensemble model. The created model outperforms both a single-learner method and the leading correlation in terms of predictive potential. The model may be used as a quick tool to validate the quality of experimental data and/or the validity and accuracy of various viscosity models, particularly when there are differences and uncertainties in the datasets. The suggested model, which has increased the accuracy and efficiency of crude oil viscosity, may be used in any reservoir simulator software. This demonstrates the ensemble learners algorithm's



stability, dependability and high performance in modeling crude oil viscosity.

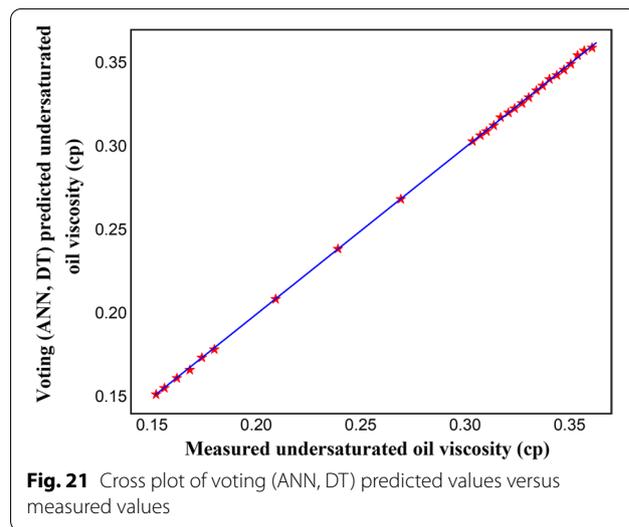
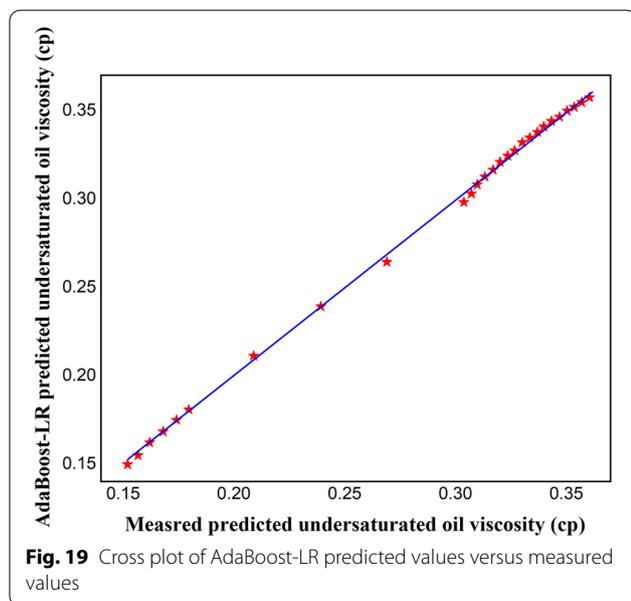
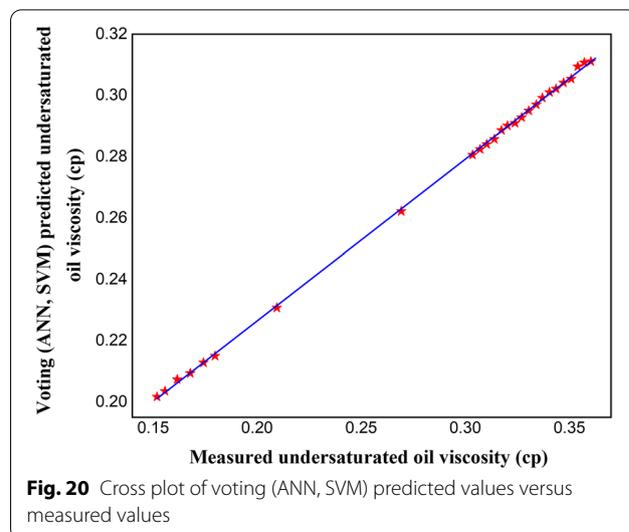
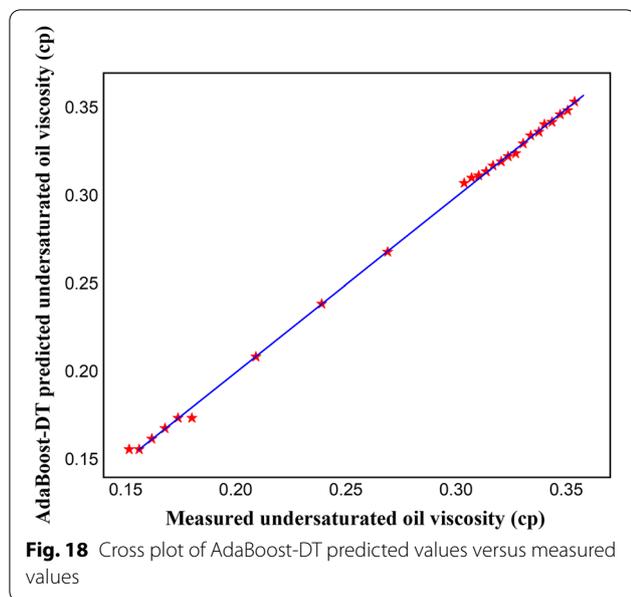
6.2 The sensitivity analysis of the ensemble learners

To show the impact of all input factors, such as GOR, temperature, gas gravity and oil gravity, the effects of the individual independent variables on the ensemble learners were examined. The outcomes of the ensemble learners’ sensitivity analysis are shown in Fig. 28. The rank correlation coefficients between the output variable and the samples for each of the input parameters

are displayed in this figure. In general, the effect of any input in deciding the value of the output grows as the correlation coefficient between that input and the output variable rises. The chart makes it clear that the main factor affecting oil viscosity is GOR.

7 Conclusion

A rigorous comparison is presented between two different types of regression schemes viz: single-based machine learning techniques (such as ANN, SVM, DT and LR) and ensemble learning techniques (such as

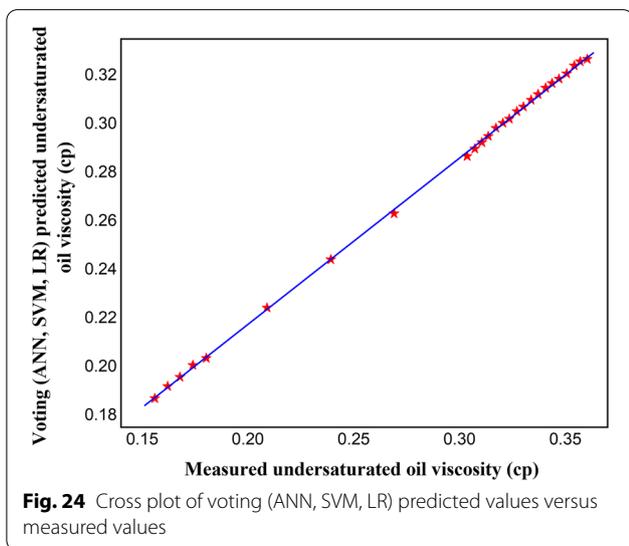
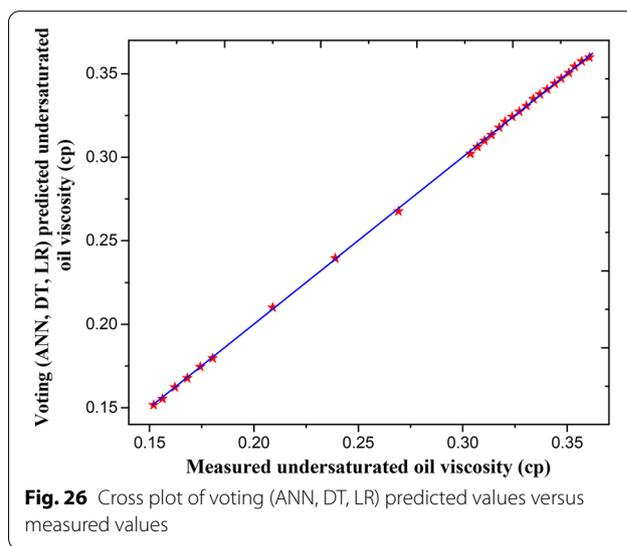
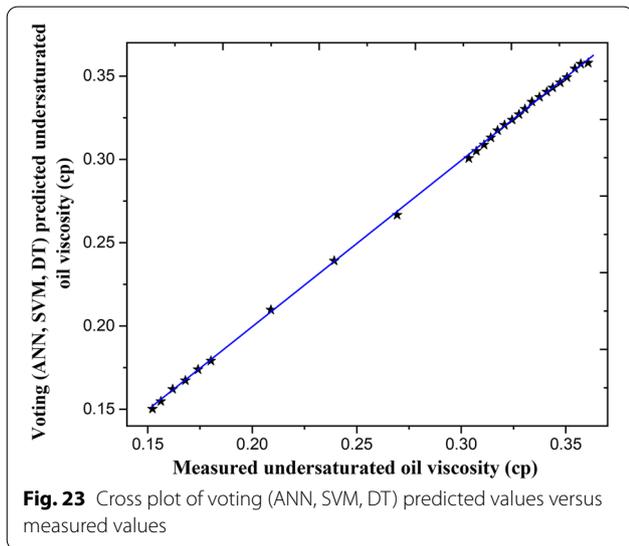
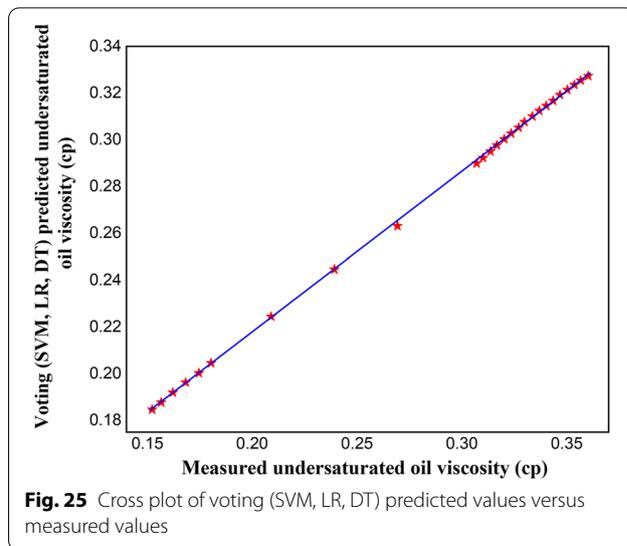
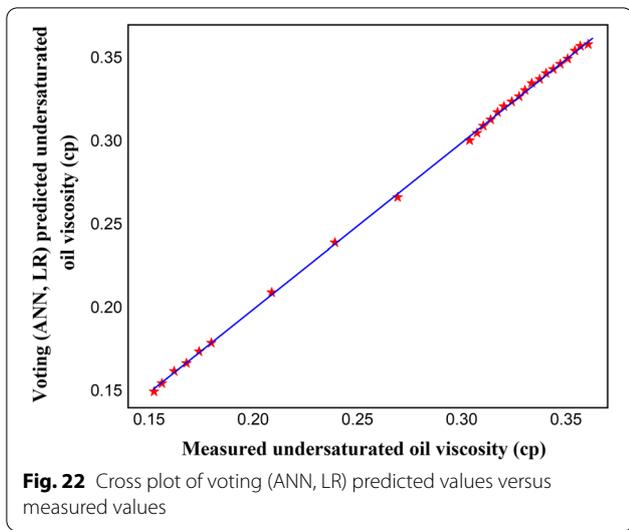


bagging, boosting and voting) to predict undersaturated oil viscosity of Nigeria’s crude oil. Four (4) variables were utilized, namely API gravity, gas/oil ratio, bubble-point pressure and reservoir temperature. The sensitivity analysis reveals that the oil viscosity is mostly affected by the gas/oil ratio, *GOR*.

The response variables in absolute terms were measured via the MAE, MSE, RMSE and RMSLE. When compared to the single-based learner, the values of these error evaluation metrics are much lower in all ensemble approaches. In the same way, the R^2 values

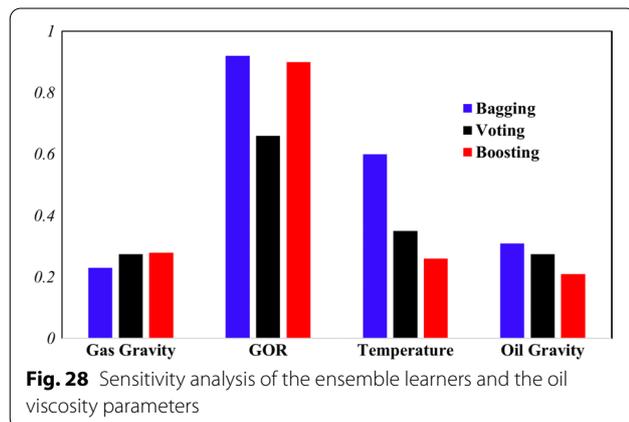
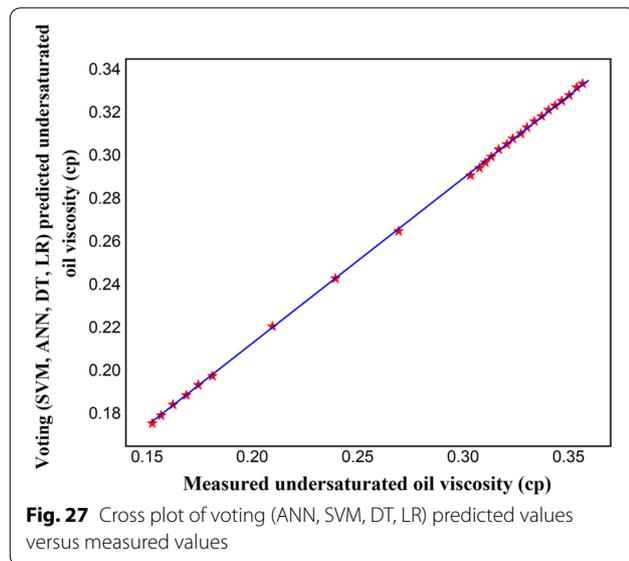
are higher in ensemble methods than in the single-based learner counterpart. This shows that ensemble models are consistent and reliable when compared with their respective individual single-based models. However, the results obtained during the testing and training using the bagging ensemble learners reveal that the developed bagged ensemble method has better capabilities in making known the uncertainties embedded in the data as their correlation coefficient values were significantly improved when compared with the other learners. It also reveals that the created model outperforms both a single-learner method and the leading correlation procedure in terms of predictive potential.

Through the voting ensemble technique, the weak prediction made by single-based SVM had a



momentous improvement from an R^2 value of -0.0195 – 0.88078 . This proves that the application of ensemble methods can transform weak learners into strong ones using the combination of the appropriate algorithms. In general, ensemble methods have great prospects of enhancing the overall predictive accuracy of single-based learners.

Finally, this study looked at a method for estimating viscosity regardless of temperature, oil type or other crucial characteristics that are difficult to collect experimentally. In the quest for more rigorous and reliable tools by petroleum and reservoir engineers, future works in this domain should incorporate ensemble learning techniques in predicting other PVT properties such as GOR, B_o and isothermal compressibility of oil.



Abbreviations

ADAM: Adaptive movement estimation; API: American petroleum institute; ANN: Artificial neural network; Bo: Oil formation volume factor, RB/STB; DS: Dataset; DT: Decision tree; FVF: Formation volume factor; GOR: Gas oil ratio; IG: Information gain; lbfgs: Limited-memory-Broyden Fletcher-Goldfarb-Shanno; LR: Linear regression; MAE: Mean absolute error; MLP: Multilayer perceptron; NN: Neural network; PVT: Pressure volume temperature; RBF: Radial basis function; RF: Rain forest; RLU: Rectified linear unit; R^2 : Relative squared error; RMSE: Root mean squared error; RMSLE: Root mean squared log error; SVM: Support vector machine.

Acknowledgements

Not applicable.

Author contributions

TTA drafted the work and analyzed the results. CCJ did the data acquisition and development of the algorithm together with the simulations. The authors have read and endorsed the manuscript and shared responsibility for the decision to publish. The final document was read and approved by all of the authors.

Funding

This research work has no funding.

Availability of data

Not applicable

Declarations

Ethics approval and consent to participate

There are no ethical issues.

Consent for publication

Authors' consent is given.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Systems Engineering, University of Lagos, Lagos, Nigeria.

²Department of Mechanical Engineering, University of Botswana, Gaborone, Botswana.

Received: 9 September 2021 Accepted: 22 November 2022

Published online: 12 December 2022

References

- Ahmed T (2006) Reservoir engineering handbook. Elsevier/Gulf Professional, Oxford OX2 8DP, UK
- Elsharkwy AM, Gharbi RBC (2001) Comparing classical and neural regression techniques in modeling crude oil viscosity. *Adv Eng Softw* 32(3):215–224
- Moharam H, Al-Mehaideb R, Fahim M (1995) New correlation for predicting the viscosity of heavy petroleum fractions. *Fuel* 74(12):1776–1779
- Labedi R (1992) Improved correlations for predicting the viscosity of light crudes. *J Pet Sci Eng* 8(3):221–234
- Salimi H, Sieders B, Rostami J (2022) Non-isothermal compositional simulation study for determining an optimum EOR strategy for a middle-east offshore heavy-oil reservoir with compositional variations with depth. In: SPE. <https://doi.org/10.2118/200274-ms>
- Abedini A, Abedini R (2012) Investigation of splitting and lumping of oil composition on the simulation of asphaltene precipitation. *Pet Sci Technol* 30(1):1–8
- Standing MB (1947) A pressure-volume-temperature correlation for mixtures of California oils and gases. In: *Drilling and production practice*. OnePetro
- Lasater J (1958) Bubble point pressure correlation. *J Petrol Technol* 10(05):65–67
- Chew J-N, Connally CA (1959) A viscosity correlation for gas-saturated crude oils. *Trans AIME* 216(01):23–25
- Beggs HD, Robinson JR (1975) Estimating the viscosity of crude oil systems. *J Petrol Technol* 27(09):1140–1141
- Glaso O (1980) Generalized pressure-volume-temperature correlations. *J Petrol Technol* 32(05):785–795
- Vazquez M, Beggs HD (1977) Correlations for fluid physical property prediction. In: *SPE annual fall technical conference and exhibition*. OnePetro
- Petrosky G, Farshad F (1993) Pressure-volume-temperature correlations for Gulf of Mexico crude oils. In: *SPE annual technical conference and exhibition*. OnePetro
- Dindoruk B, Christman PG (2004) PVT properties and viscosity correlations for Gulf of Mexico oils. *SPE Reserv Eval Eng* 7(06):427–437
- Abd Talib MQ, Al-Jawad MS (2022) Assessment of the common PVT correlations in Iraqi Oil Fields. *J Pet Res Stud* 12(1):68–87
- Hadavimoghaddam F, Ostadhasan M, Heidaryan E, Sadri MA, Chapanova I, Popov E, Cheremisin A, Rafieepour S (2021) Prediction of dead oil viscosity: machine learning vs. classical correlations. *Energies* 14(4):930. <https://doi.org/10.3390/en14040930>
- Ahmed T (2018) Reservoir engineering handbook. Gulf Professional Publishing, Oxford OX2 8DP, UK
- Ahrabi F, Ashcroft S, Shearn R (1987) High pressure volumetric, phase composition and viscosity data for a north sea crude oil and Ngl. *Chem. Eng. Res. Des. (United Kingdom)* 65(11):329–334

19. Beal C (1946) The viscosity of air, water, natural gas, crude oil and its associated gases at oil field temperatures and pressures. *Trans AIME* 165(01):94–115
20. Beggs HD, Robinson JR (1975) Estimating the viscosity of crude oil systems. *J Petrol Technol* 27(09):1140–1141
21. Chew J-N, Connally CA (1959) A viscosity correlation for gas-saturated crude oils. *Trans AIME* 216(01):23–25
22. Egbogah EO, Ng JT (1990) An improved temperature-viscosity correlation for crude oil systems. *J Petrol Sci Eng* 4(3):197–200
23. Elsharkawy A, Alikhan A (1999) Models for predicting the viscosity of middle east crude oils. *Fuel* 78(8):891–903
24. Rice P, Teja AS (1982) A generalized corresponding-states method for the prediction of surface tension of pure liquids and liquid mixtures. *J Colloid Interface Sci* 86(1):158–163
25. Vazquez M, Beggs H (1980) Correlations for fluid physical property prediction. *Ipt* 32(6):968–970. <https://doi.org/10.2118/6719-PA>
26. Little J, Kennedy H (1968) A correlation of the viscosity of hydrocarbon systems with pressure, temperature and composition. *Soc Petrol Eng J* 8(02):157–162
27. Sutton RP, Farshad F (1990) Evaluation of empirically derived PVT properties for Gulf of Mexico crude oils. *SPE Reserv Eng* 5(01):79–86
28. Dexheimer D, Jackson CM, Barrufet MA (2001) A modification of Pedersen's model for saturated crude oil viscosities using standard black oil PVT data. *Fluid Phase Equilib* 183:247–257
29. Taghizadeh M, Eftekhari M (2014) Improved correlations for prediction of viscosity of Iranian crude oils. *Chin J Chem Eng* 22(3):346–354
30. Shokir EME-M, Ibrahim AE-SB (2022) Undersaturated oil viscosity based on multi-gene genetic programming. *J Energy Resour Technol*. <https://doi.org/10.1115/1.4055396>
31. Moghadam EM, Naseri A, Riahi MA (2021) Further model development for prediction of reservoir oil viscosity. *Pet Sci Technol* 40(3):310–321. <https://doi.org/10.1080/10916466.2021.1993914>
32. Sinha U, Dindoruk B, Soliman MY (2022) Physics augmented correlations and machine learning methods to accurately calculate dead oil viscosity based on the available inputs. *SPE J* 27(05):3240–3253. <https://doi.org/10.2118/209610-pa>
33. Kartoatmodjo T, Schmidt Z (1991) New correlations for crude oil physical properties. paper SPE 23556
34. Obanjesu E, Omidiora E (2009) The artificial neural network's prediction of crude oil viscosity for pipeline safety. *Pet Sci Technol* 27(4):412–426
35. Gao X, Dong P, Cui J, Gao Q (2022) Prediction model for the viscosity of heavy oil diluted with light oil using machine learning techniques. *Energies* 15(6):2297. <https://doi.org/10.3390/en15062297>
36. Bhat SS, Selvam V, Ansari GA, Ansari MD, Rahman MH (2022) Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district bandipora. *Comput Intell Neurosci* 2022:1–12. <https://doi.org/10.1155/2022/2789760>
37. Kannan R, Halim HAA, Ramakrishnan K, Ismail S, Wijaya DR (2022) Machine learning approach for predicting production delays: a quarry company case study. *J Big Data*. <https://doi.org/10.1186/s40537-022-00644-w>
38. Zhang Z, Yang L, Han W, Wu Y, Zhang L, Gao C, Jiang K, Liu Y, Wu H (2022) Machine learning prediction models for gestational diabetes mellitus: meta-analysis. *J Med Internet Res* 24(3):26634. <https://doi.org/10.2196/26634>
39. Dhiman P, Ma J, Navarro CLA, Speich B, Bullock G, Damen JAA, Hooft L, Kirtley S, Riley RD, Calster BV, Moons KGM, Collins GS (2022) Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol*. <https://doi.org/10.1186/s12874-022-01577-x>
40. Tercan H, Meisen T (2022) Machine learning and deep learning based predictive quality in manufacturing: a systematic review. *J Intell Manuf* 33(7):1879–1905. <https://doi.org/10.1007/s10845-022-01963-8>
41. Meraihi Y, Gabis AB, Mirjalili S, Ramdane-Cherif A, Alsaadi FE (2022) Machine learning-based research for COVID-19 detection, diagnosis, and prediction: a survey. *SN Comput Sci*. <https://doi.org/10.1007/s42979-022-01184-z>
42. Tao H, Hameed MM, Marhoon HA, Zounemat-Kermani M, Heddad S, Kim S, Sulaiman SO, Tan ML, Sa'adi Z, Mehr AD, Allawi MF, Abba SI, Zain JM, Falah MW, Jamei M, Bokde ND, Bayatvarkeshi M, Al-Mukhtar M, Bhagat SK, Tiyyasha T, Khedher KM, Al-Ansari N, Shahid S, Yaseen ZM (2022) Groundwater level prediction using machine learning models: a comprehensive review. *Neurocomputing* 489:271–308. <https://doi.org/10.1016/j.neucom.2022.03.014>
43. Zhou Y, Han F, Shi X-L, Zhang J-X, Li G-Y, Yuan C-C, Lu G-T, Hu L-H, Pan J-J, Xiao W-M, Yao G-H (2022) Prediction of the severity of acute pancreatitis using machine learning models. *Postgrad Med* 134(7):703–710. <https://doi.org/10.1080/00325481.2022.2099193>
44. Vallim Filho AR, Moraes DF, de Aguiar Vallim MVB, da Silva LS, da Silva LA (2022) A machine learning modeling framework for predictive maintenance based on equipment load cycle: an application in a real world case. *Energies* 15(10):3724. <https://doi.org/10.3390/en15103724>
45. Gulyani BB, Kumar BP, Fathima A (2017) Bagging ensemble model for prediction of dead oil viscosity. *Int J Chem Eng Appl* 8(2):102
46. Zhou ZH (2009) Ensemble learning. In: Li SZ, Jain A (eds) *Encyclopedia of biometrics*. Springer, Boston, MA, pp270–273
47. Zheng Z, Padmanabhan B (2007) Constructing ensembles from data envelopment analysis. *INFORMS J Comput* 19(4):486–496
48. Polikar R (2009) Ensemble learning. *Scholarpedia* 4(1):2776
49. Dietterich TG (2000) Ensemble methods in machine learning. In: *International workshop on multiple classifier systems*. Springer, Berlin, Heidelberg, pp 1–15
50. Santos R, Vellasco MM, Artola F, Da Fontoura S (2003) Neural net ensembles for lithology recognition. In: *International workshop on multiple classifier systems*, pp. 246–255. Springer
51. Gifford CM, Agah A (2010) Collaborative multi-agent rock facies classification from wireline well log data. *Eng Appl Artif Intell* 23(7):1158–1172
52. Masoudi P, Tokhmechi B, Bashari A, Jafari MA (2012) Identifying productive zones of the Sarvak formation by integrating outputs of different classification methods. *J Geophys Eng* 9(3):282–290
53. Davronova R, Adilovab F (2020) A comparative analysis of the ensemble methods for drug design
54. Smirani LK, Yamani HA, Menzli LJ, Boulahia JA (2022) Using ensemble learning algorithms to predict student failure and enabling customized educational paths. *Sci Program* 2022:1–15. <https://doi.org/10.1155/2022/3805235>
55. Whitaker T, Whitley D (2022) Prune and tune ensembles: low-cost ensemble learning with sparse independent subnetworks. <https://doi.org/10.48550/ARXIV.2202.11782>. arXiv [arxiv:2202.11782](https://arxiv.org/abs/2202.11782)
56. Marwah GPK, Jain A (2022) A hybrid optimization with ensemble learning to ensure VANET network stability based on performance analysis. *Sci Rep* 12:1. <https://doi.org/10.1038/s41598-022-14255-1>
57. Banerjee S, Sinclair SR, Tambe M, Xu L, Yu CL (2022) Artificial replay: a meta-algorithm for harnessing historical data in Bandits. <https://doi.org/10.48550/ARXIV.2210.00025>. arXiv [arxiv:2210.00025](https://arxiv.org/abs/2210.00025)
58. Longo L, Riccaboni M, Rungi A (2022) A neural network ensemble approach for GDP forecasting. *J Econ Dyn Control* 134:104278. <https://doi.org/10.1016/j.jedc.2021.104278>
59. Flennerhag S, Schroecker Y, Zahavy T, van Hasselt H, Silver D, Singh S (2021) Bootstrapped meta-learning. arXiv [arxiv:2109.04504](https://arxiv.org/abs/2109.04504)
60. Liu H, Du Y, Wu Z (2022) Generalized ambiguity decomposition for ranking ensemble learning. *J Mach Learn Res* 23(88):1–36
61. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN (2022) Ensemble deep learning: a review. *Eng Appl Artif Intell* 115:105151. <https://doi.org/10.1016/j.engappai.2022.105151>
62. Anifowose F, Labadin J, Abdurraheem A (2015) Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. *Appl Soft Comput* 26:483–496
63. Anifowose FA, Labadin J, Abdurraheem A (2017) Ensemble machine learning: an untapped modeling paradigm for petroleum reservoir characterization. *J Petrol Sci Eng* 151:480–487
64. Bestagini P, Lipari V, Tubaro S (2017) A machine learning approach to facies classification using well logs. In: *Seg technical program expanded abstracts 2017*. Society of Exploration Geophysicists, pp 2137–2142
65. Xie Y, Zhu C, Zhou W, Li Z, Liu X, Tu M (2018) Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances. *J Petrol Sci Eng* 160:182–193
66. Tewari S, Dwivedi U (2019) Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. *Comput Ind Eng* 128:937–947

67. Tewari S, Dwivedi U, et al (2018) A novel automatic detection and diagnosis module for quantitative lithofacies modeling. In: Abu Dhabi international petroleum exhibition & conference. Society of Petroleum Engineers
68. Bhattacharya S, Carr TR, Pal M (2016) Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: case studies from the bakken and mahantango-marcellus shale, usa. *J Nat Gas Sci Eng* 33:1119–1133
69. Tewari S, Dwivedi U, Shiblee M et al. (2019) Assessment of big data analytics based ensemble estimator module for the real-time prediction of reservoir recovery factor. In: SPE middle east oil and gas show and conference. Society of Petroleum Engineers
70. Tewari S, Dwivedi UD (2020) A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies. *J Pet Explor Product Technol* 10(5):1849–1868
71. Touati R, Elngar AA (2022) Intelligent system based comparative analysis study of SARS-CoV-2 spike protein and antigenic proteins in different types of vaccines. *Beni-Suef Univ J Basic Appl Sci* 1:11. <https://doi.org/10.1186/s43088-022-00216-0>
72. Mahdy AMS (2022) A numerical method for solving the nonlinear equations of Emden-Fowler models. *J Ocean Eng Sci*. <https://doi.org/10.1016/j.joes.2022.04.019>
73. Ke W, Liu Y, Zhao X, Yu G, Wang J (2022) Study on the effect of threshold pressure gradient on remaining oil distribution in heavy oil reservoirs. *ACS Omega* 7(5):3949–3962. <https://doi.org/10.1021/acsomega.1c04537>
74. Othman K (2022) Prediction of the hot asphalt mix properties using deep neural networks. *Beni-Suef Univ J Basic Appl Sci* 1:11. <https://doi.org/10.1186/s43088-022-00221-3>
75. Ahmad A, Sulaiman M, Aljohani AJ, Alhindi A, Alrabaiah H (2021) Design of an efficient algorithm for solution of Bratu differential equations. *Ain Shams Eng J* 12(2):2211–2225. <https://doi.org/10.1016/j.asej.2020.11.007>
76. Noeiaghdam S, Araghi MAF, Abbasbandy S (2020) Valid implementation of sinc-collocation method to solve the fuzzy Fredholm integral equation. *J Comput Appl Math* 370:112632. <https://doi.org/10.1016/j.cam.2019.11.2632>
77. Khan MM, Sohrab MG, Yousuf MA (2020) Customer gender prediction system on hierarchical e-commerce data. *Beni-Suef Univ J Basic Appl Sci* 9:1. <https://doi.org/10.1186/s43088-020-0035-7>
78. Gumah G, Naser MFM, Al-Smadi M, Al-Omari SKQ, Baleanu D (2020) Numerical solutions of hybrid fuzzy differential equations in a Hilbert space. *Appl Numer Math* 151:402–412. <https://doi.org/10.1016/j.apnum.2020.01.008>
79. Arqub OA (2015) Adaptation of reproducing Kernel algorithm for solving fuzzy Fredholm–Volterra integrodifferential equations. *Neural Comput Appl* 28(7):1591–1610. <https://doi.org/10.1007/s00521-015-2110-x>
80. Arqub OA, Abo-Hammour Z (2014) Numerical solution of systems of second-order boundary value problems using continuous genetic algorithm. *Inf Sci* 279:396–415. <https://doi.org/10.1016/j.ins.2014.03.128>
81. Abo-Hammour Z, Arqub OA, Alsmadi O, Momani S, Alsaedi A (2014) An optimization algorithm for solving systems of singular boundary value problems. *Appl Math Inf Sci* 8(6):2809–2821. <https://doi.org/10.12785/amis/080617>
82. Abo-Hammour Z, Arqub OA, Momani S, Shawagfeh N (2014) Optimization solution of Troesch's and Bratu's problems of ordinary type using novel continuous genetic algorithm. *Discret Dyn Nat Soc* 2014:1–15. <https://doi.org/10.1155/2014/401696>
83. Lopez-Franco C, Hernandez-Barragan J, Alanis AY, Arana-Daniel N (2018) A soft computing approach for inverse kinematics of robot manipulators. *Eng Appl Artif Intell* 74:104–120. <https://doi.org/10.1016/j.engappai.2018.06.001>
84. Dereli S, Köker R (2019) Simulation based calculation of the inverse kinematics solution of 7-DOF robot manipulator using artificial bee colony algorithm. *SN Appl Sci*. <https://doi.org/10.1007/s42452-019-1791-7>
85. Obot NI, Humphrey I, Chendo MAC, Udo SO (2019) Deep learning and regression modelling of cloudless downward longwave radiation. *Beni-Suef Univ J Basic Appl Sci*. <https://doi.org/10.1186/s43088-019-0018-8>
86. Pintelas P, Livieris IE (2020) Special issue on ensemble learning and applications. *Algorithms* 13(6):140
87. Liu Y (2016) Error awareness by lower and upper bounds in ensemble learning. *Int J Pattern Recognit Artif Intell* 30(09):1660003
88. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
89. Al-Khafaji AH, Abdul-Majeed GH, Hassoon SF et al (1987) Viscosity correlation for dead, live and undersaturated crude oils. *J Pet Res* 6(2):1–16
90. Bennison T (1998) Prediction of heavy oil viscosity. In: IBC Heavy oil field development conference, vol 2, p 4

Publisher's Note

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

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
