Compositional Modeling of the Oil Formation Volume Factor of Crude Oil Systems: Application of Intelligent Models and Equations of State

Aydin Larestani,* Abdolhossein Hemmati-Sarapardeh,* Zahra Samari, and Mehdi Ostadhassan*

ABSTRACT: This communication primarily concentrates on developing reliable and accurate compositional oil formation volume factor (\(B_o\)) models using several advanced and powerful machine learning (ML) models, namely, extra trees (ETs), random forest (RF), decision trees (DTs), generalized regression neural networks, and cascade-forward back-propagation network, alongside radial basis function and multilayer perceptron neural networks. Along with these models, seven equations of state (EoSs) were employed to estimate \(B_o\). The performance of the developed ML models and employed EoSs was assessed through various statistical and graphical evaluations. Overall, the ML models could provide much more accurate predictions in comparison to EoSs. However, the results indicated that tree-based models, specifically ET models, could outperform the other models and can be reliably applied for estimating \(B_o\). The most reliable ET model could predict \(B_o\) with a total average error of 1.17%. Lastly, the outlier detection approach verified the dataset’s consistency detecting only 17 (out of 1224) data points as outliers for the proposed \(B_o\) models.

1. INTRODUCTION

It is undeniable that petroleum and fossil fuels have dramatically transformed humans’ life. Since the advent of oil, many scholars and people who are involved in oil exploration and production have aimed at the determination of its properties, most of which are controlled by pressure, volume, and temperature and accordingly named PVT properties. The oil formation volume factor (\(B_o\)) that refers to the ratio of the volume of oil under reservoir conditions to its corresponding value under standard conditions is one of the most significant PVT properties. The reliability of many reservoir engineering computations, such as reservoir simulations, well completion, inflow performance, fluid flow in porous media, reserve and oil recovery estimation, material balance, well test analysis, and economic analysis, highly depends on having accurate knowledge of PVT properties.

Indisputably, precise determination of PVT properties can aid both academia and industry. The most ideal and reliable way toward the determination of \(B_o\) is to conduct experiments, specifically, constant composition expansion (CCE), differential liberation (DL), and separator tests, on bottom-hole reservoir fluid samples or recombined oil and gas samples at the surface. However, experimental determination of this property involves costly and time-consuming procedures, and low-quality sampling may lead to significant errors. In the absence of experimental measurements, various predictive approaches are proposed on the basis of equations of state (EoSs), empirical correlations, and artificial intelligence (AI) techniques. EoSs are not so accurate approaches and involve complex computational procedures. Therefore, numerous scholars have proposed empirical correlations and AI-based models to provide reliable estimations.

Table 1. Summary of Previous Attempts toward Estimating \(B_o\) Using ML

| author(s) | algorithm(s) | error (AAPRE, %) |
|-----------|--------------|-----------------|
| Gharbi and Elsharkawy | ANNs | 2.79 |
| Elsharkawy | RBF | 0.53 |
| Mahdiani and Norouzi | SA | 1.25 |
| Fattah and Lashin | GP | 0.3252 |
| Elkatatny and Mahmoud | ANN | 0.99 |
| Saghafi et al | ANFIS | 1.8 |
| Seyyedattar et al | ET | 0.099 |
| Rashidi et al | PSO-MELM | |
| Tariq et al | PSO-FN | |

Received: March 10, 2022
Accepted: June 8, 2022
Published: July 6, 2022

https://doi.org/10.1021/acsomega.2c01466
Table 2. Details of the Lumped Data

| group | components |
|-------|------------|
| $i_1$ | $N_2$, $C_1$ |
| $i_2$ | $H_2$, $CO_2$, $C_2$, $C_5$ |
| $i_3$ | $C_6$, $C_{11}$ |

Table 3. Statistical Description of Normal Datasets

| statistical parameters | mean | standard deviation | min. | max. |
|------------------------|------|--------------------|------|------|
| temperature, °F | 218.20 | 39.73 | 135 | 290 |
| $H_2S$, mole % | 0.61 | 0.01 | 0 | 3.3997 |
| $N_2$, mole % | 0.32 | 0.73 | 0.0071 | 5.8559 |
| $CO_2$, mole % | 1.91 | 1.91 | 0.0286 | 8.2855 |
| $C_1$, mole % | 34.27 | 12.62 | 5.0321 | 57.992 |
| $C_2$, mole % | 7.66 | 1.78 | 3.0490 | 13.699 |
| $C_3$, mole % | 5.79 | 1.51 | 1.7510 | 14.039 |
| $C_4$, mole % | 4.43 | 1.19 | 1.6308 | 11.124 |
| $C_5$, mole % | 2.52 | 0.71 | 0.6500 | 5.2203 |
| $C_6$, mole % | 4.82 | 1.40 | 2.1037 | 10.248 |
| $C_7$, mole % | 3.80 | 1.09 | 1.5222 | 7.4999 |
| $C_8$, mole % | 3.11 | 0.95 | 1.6444 | 7.3577 |
| $C_9$, mole % | 2.95 | 0.93 | 1.3481 | 5.4485 |
| $C_{10}$, mole % | 2.54 | 0.70 | 1.1149 | 4.3286 |
| $C_{11}$, mole % | 2.25 | 0.62 | 0.9146 | 4.0375 |
| specific gravity of $C_{12}$ | 0.93 | 0.04 | 0.7910 | 1.0014 |
| molecular weight of $C_{12}$ | 384.48 | 71.28 | 218 | 545 |
| pressure, psi | 2687.4 | 1808.8 | 14.7 | 10072 |
| $B_o$, bbl/STB | 1.46 | 0.36 | 1.0309 | 3.9454 |

Table 4. Statistical Description of Lumped Datasets

| statistical parameters | mean | standard deviation | min. | max. |
|------------------------|------|--------------------|------|------|
| Temperature, °F | 218.20 | 39.73 | 135 | 290 |
| $I_1$, mole % | 34.59 | 12.76 | 5.1668 | 63.719 |
| $I_2$, mole % | 22.92 | 4.48 | 8.82 | 44.309 |
| $I_3$, mole % | 19.47 | 4.87 | 9.8602 | 31.373 |
| specific gravity of $C_{12}$ | 0.93 | 0.04 | 0.7910 | 1.0014 |
| molecular weight of $C_{12}$ | 384.48 | 71.28 | 218 | 545 |
| pressure, psi | 2687.4 | 1808.8 | 14.7 | 10072 |
| $B_o$, bbl/STB | 1.46 | 0.36 | 1.0309 | 3.9454 |

for $B_o$ during the recent few decades. What follows provides a brief review of some of the most important studies in this field.

One of the very first predictive approaches toward the estimation of oil PVT properties dates back to the early 40s, when Katz attempted to provide estimations for $B_o$ based on the solution gas–oil ratio ($R_g$), oil and gas gravity, and reservoir conditions. The following attempt was made by Standing in 1947, where two distinct correlations were proposed based on $R_g$, oil and gas gravity, and reservoir temperature for the estimation of $B_o$ and saturation pressure ($P_s$) of Californian crude oils. Later, in 1980, Vazquez and Beggs aimed at providing accurate correlations for $B_o$ based on 600 datasets gathered from various geographical locations. In this year, Glaso proposed new correlations for $B_o$ based on 41 datasets collected from the North Sea. More recently, in 2014, Arabloo et al. aimed at providing accurate $B_o$ correlations through the constrained multivariable search method. Although empirical correlations are easy-to-use and can be employed in the absence of costly experiments, they are sometimes associated with considerable errors as they are not so powerful in expressing the exact relationship between the input and output parameters.

With the emergence of AI techniques and the proof of their power in solving classification and regression problems, they have been extensively used in numerous fields of study, including the $B_o$ determination field. In 1997, Gharbi and Elsharkawy proposed one of the earliest approaches for predicting $B_o$ of Middle Eastern crude oils using artificial neural networks (ANNs) considering the specific gravity (SG) of oil, relative density of gas, temperature ($T$), and $R_g$ as input variables. They employed 520 data points to train and test their model and reported the value of 2.79% as the total average absolute percent relative error (AAPRE) of the developed $B_o$ model. In a subsequent approach, Elsharkawy attempted to predict various properties of crude oils including $B_o$ using a radial basis function (RBF) neural network. The developed $B_o$ model led to a total AAPRE of 0.53%.

More recently, in 2018, three studies focused on representing novel $B_o$ correlations based on machine learning (ML) techniques. Mahdiani and Norouzi utilized the simulated annealing (SA), which is basically an optimization approach, as the modeling approach to develop their correlation. They gathered 160 data points for $T$, $R_g$, and the gravity of oil and gas to train and test the correlation. The results showed a total AAPRE of 1.25%. In the study of Fattah and Lashin, genetic programming (GP) was applied to correlate $B_o$ with $T$, American Petroleum Institute (API), $R_g$, and gas SG. The developed correlation could predict more than 1200 measured points with a total average error of 0.3252%. The third correlation in this year was developed by Elkatatny and Mahmoud using 760 data points. They first compared the efficiency of the adaptive neuro-fuzzy inference system (ANFIS), support vector regression (SVR), and ANN in predicting the $B_o$ and concluded that the ANN is the most efficient tool. Then, a correlation was developed based on the ANN model, which led to a total AAPRE of 0.99%.

In another study, Saghafi et al. gathered 1200 data points to propose an ANFIS model and two GP-based correlations for $B_o$. In their study, $R_g$, $T$, API, and gas SG were employed as input parameters. The results indicated that the ANFIS model could provide predictions with slightly better accuracy (total AAPRE of 1.8%). In another study this year, Seyyedattar et al. proposed the first application of extra trees (ETs) in predicting $B_o$. The performance of the developed ET model was compared to that of ANFIS and least square support vector machine-coupled simulated annealing (LSSVM-CSA) models. Their models were trained using 561 data points, which contained values for $R_g$, $T$, API, and gas SG. They claimed that the ET approach provided the most accurate results for $B_o$ in comparison to two other models with a total AAPRE of 0.099%. However, a large difference is observed between training and testing performance of the model, and hence, their model is suspected of overfitting.

In 2021, two studies were focused on $B_o$ prediction using ML approaches. Rashidi et al. employed LSSVM and multi-layer extreme learning machine (MELM) approaches to model $B_o$ and $P_s$, considering $T$, API, $R_g$, and gas SG. They developed their $P_s$ and $B_o$ models based on 591 and 599 data points, respectively.
The models were optimized using particle swarm optimization (PSO) and genetic algorithm (GA) techniques. The results revealed the higher performance of PSO-MELM in predicting both $P_b$ and $B_o$. In another study, Tariq et al.\textsuperscript{20} coupled the functional network with PSO to predict $P_b$, $B_o$, and oil viscosity at $P_b$ based on 760 gathered data points. It is claimed that the developed PSO-FN models could provide better predictions in comparison to other ML techniques, that is, ANN, SVR, and ANFIS. A summary of the previous studies that tried to use ML approaches to predict $B_o$ has been provided in Table 1.

To the best of our knowledge, all the existing empirical correlations and ML models (herein, ML models are referred to as intelligent models) that were proposed to estimate $B_o$ are based on black oil properties, and no effort has been made to develop compositional models toward predicting this important parameter. This is mainly because of the lack of comprehensive compositional data for this parameter. Since compositional models are usually more robust in comparison to black oil models, this communication principally concentrates on developing reliable and accurate compositional $B_o$ models for Middle East crude oil systems using various advanced and powerful intelligent approaches, namely, ETs, random forest (RF), decision trees (DTs), generalized regression neural networks (GRNNs), and cascade-forward back-propagation network (CFBPN) and multilayer perceptron (MLP) and RBF neural networks. Moreover, the efficiency of three different training algorithms, that is, Bayesian regularization (BR), scaled conjugate gradient (SCG), and Levenberg–Marquardt (LM), is assessed in training the CFBPN and MLP models. The models accept oil composition ($H_2S$, $CO_2$, $N_2$, and $C_1$ to $C_{11}$), temperature, pressure, and the specifications of $C_{12+}$ as inputs. This is the first time that ML models are employed for compositional modeling of $B_o$. The performance of the offered models is assessed with respect to many statistical and graphical error assessments. Furthermore, the efficiency of the developed models at various temperatures was evaluated. Also, a trend analysis was performed to evaluate the power of the proposed models in detecting the physical trend of the target parameters. Finally, an outlier detection is performed using the Leverage approach to assess the models’ applicability domain and detect outliers.

In the following, section 2 presents details about experimental procedures and the gathered databank. Then, section 3 provides an exhaustive explanation of AI modeling approaches. Section 4 represents the modeling outcomes and evaluates the proficiency of the developed models in contrast to various EoSs. In the end, section 5 highlights the points that can be concluded from this research.
2. MATERIALS AND METHODS

A set of PVT experiments, including CCE, DL, and the separator test, is carried out to determine the oil formation $B_o$ for various Iranian oil samples. The parameter is obtained through a specific combination of these tests. The data obtained from DL and separator tests leads to the determination of $B_o$ at pressures above the saturation pressure, while $B_o$ at pressures lower than bubble point pressure is determined using the combination of the data attained from the CCE and separator tests. The following equations express how these parameters were calculated:

$$B_o = \left( \frac{V_t}{V_b} \right)_\text{CCE} B_{osb}(P \geq P_b) \quad (1)$$

$$B_o = B_{osb} B_{odb}(P_b > P) \quad (2)$$

where $(V_t/V_b)_\text{CCE}$ = relative total volume by the CCE test, $B_{osb}$ = formation volume factor at bubble point pressure from the separator test, $B_{odb}$ = relative volume at bubble point pressure from the DL test, $B_{odb}$ = relative volume from the DL test.

Then, various ML approaches are applied to precisely estimate the oil formation $B_o$. The models are developed and tested based on 1224 data points. The input vectors of all models comprise values for oil composition, reservoir temperature and pressure, and $C_{12+}$ specifications [molecular weight (MW) and SG]. Prior to the models’ training phase, the datasets are prepared based on two different schemes. In the first type of datasets, the mole percent of each fraction is used as a distinct input parameter, while the fractions’ mole percent is regrouped into three clusters in the other type of datasets. Table 2 presents details about the performed grouping. Accordingly, the first kind of model is called the “Normal Model”, and the other ones are named “Lumped Models”. The aim of considering two different approaches is to reduce the computational cost through decreasing the number of input (without excluding any parameter) so that in case of integrating the developed models with commercial simulators, they can respond, consuming less time and memory.
predictive models are to learn from 80 percent of the data and then be tested against the remaining 20 percent. The subsets were chosen arbitrarily. A brief statistical description of the employed databank for the normal and lumped models is presented in Tables 3 and 4. The employed data was gathered from the Research Institute of Petroleum Industry (RIPI) that is under the surveillance of the National Iranian Oil Company (NIOC), and the Supporting Information contains the gathered data for representative 10 oil samples.

3. INTELLIGENT METHODS

3.1. Decision Tree. In 1984, Breiman et al. introduced a new supervised non-parametric learning algorithm, namely, the classification and regression tree (CART), which is applicable to both classification and regression problems. With respect to the problem they are being applied to, a decision tree is termed “The classification tree (CT)” or “The regression tree (RT)”.

Variable selection, data manipulation, handling of missing values, and prediction were among the most popular fields in which DTs have been employed owing to their simplicity, interpretability, ability to provide graphical representation, and low computational cost. The algorithm of DTs involves root, internal, and leaf nodes, which are joined through branches. The root node that carries the input data is located at the top of the tree, while the leaves deliver the output. The data starts moving from the root node to internal nodes and then leaf nodes. Thus, the shape of the model is similar to that of an upside-down tree.

The DT tries to provide an easy-to-interpret solution by splitting a complicated problem. Splitting, stopping, and pruning are the main actions in the development phase of a DT. The development phase starts from the root node by dividing the training data. The splitting moves forward to internal nodes. The splitting process continues until predefined stopping criteria are satisfied. Then, the complexity of the tree is reduced by the pruning approach, which prevents overfitting. A typical DT model is illustrated in Figure 1.

3.2. Random Forest. Seventeen years after the introduction of DTs, Breiman developed an RF as a more robust model. The RF involves a collection of several independent DTs (weak learners). The algorithm is developed based on the notion of random feature selection and bagging simultaneously. In order to minimize the generalization error, the RF randomly selects input variables instead of considering the best variable.

Once the training data is loaded, a number (N) of unpruned regression/classification trees are generated using N bootstrapped datasets produced by bagging the original databank. In other words, the whole databank is divided into several subsets, and different trees will be trained based on each subset. This process guarantees the trees’ diversity and lessens the model’s total error. The term “out-of-bag” (OOB) refers to the samples that are not included in the training data. The OOB subset is unique for each tree as just a sample of the original data is used to generate each tree. Thus, the OOB subsets are used to validate the model. In the RF algorithm, the Gini index is employed to evaluate the impurity (error) of a particular component compared to other classes and to select the best split. Finally, the aggregation of the outputs of generated trees represents the output of the RF model. Figure 2 illustrates a schematic of an RF model.

3.3. Extra Tree. The ET, that is, another variation of tree-based ensemble models, was proposed by Geurts et al. for the first time. Like the DT and RF, it solves classification and regression problems. This relatively novel ML approach is also called the extremely randomized tree (ERT). The algorithm is based on the same notions as those of the RF with two fundamental differences: In the ET, no bootstrap or bagging procedure is performed, and trees are generated based on all of the training datapoints. Another difference is that the ET splits nodes randomly, while the RF employs the Gini index to select the best split.

3.4. Multilayer Perceptron Neural Network. As a basic form of ANNs, they were developed and introduced in the 1960s. The structure of MLP, analogous to other forms of ANNs, mimics the human brain’s nervous system’s process. The algorithm involves three classes of layers. The input layer is the first layer that carries the input data, where the number of input variables determines the number of neurons. The output layer, located at the end of the model, represents the output(s). The hidden layer(s) connects the input and output layers and is responsible for detecting the relationship between the input and output variables via transfer (activation) functions. The mathematical expressions of some of the transfer functions are presented below

\[ \text{sinusid} : f(x) = \sin(x) \]  
\[ \text{tansig} : f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{2}{1 - e^{-2x}} - 1 \]  
\[ \text{logsig} : f(x) = \frac{1}{1 + e^{-x}} \]  
\[ \text{purelin} : f(x) = x \]

Table 5. Optimum Values for Hyperparameters of the Developed \( B_S \) Models

| Hyperparameters                                      | normal model (s) | lumped model (s) |
|------------------------------------------------------|------------------|------------------|
| DT                                                   |                  |                  |
| - minimum leaf size                                 | 5                | 1 12             |
| - minimum parent size                                | 10               |                  |
| - predictor selection                                | curve all        |                  |
| - number of variables to samples                     |                  |                  |
| RF                                                   |                  |                  |
| - number of learners                                 | 10               | 71 all           |
| - minimum leaf size                                  | 2                |                  |
| - number of variables to samples                     | All              |                  |
| ET                                                   |                  |                  |
| - number of learners                                 | 30               | 30               |
| - minimum leaf size                                  | 5                | 1                |
| - number of variables to samples                     | all              | all              |
| GRNN                                                 |                  |                  |
| - spread coefficient                                 | 0.3000           | 0.3300           |
| CFBPN                                                |                  |                  |
| - training algorithm                                 | BR, LM, SCG      | BR, LM, SCG      |
| - number of hidden layers                            | 1                | 1                |
| - number of hidden neurons                           | 7                | 12               |
| - transfer function                                  | tansig           | tansig           |
| RBF                                                  |                  |                  |
| - maximum number of neurons                          | 150              | 150              |
| - spread coefficient                                 | 1.5              | 1                |
| MLP                                                  |                  |                  |
| - training algorithm                                 | BR, LM, SCG      | BR, LM, SCG      |
| - number of hidden layers                            | 1                | 1                |
| - number of hidden neurons                           | 8                | 15               |
| - transfer function                                  | tansig           | tansig           |

https://doi.org/10.1021/acsomega.2c01466
ACS Omega 2022, 7, 24256–24273
Table 6. Statistical Evaluation of the Proposed $B_o$ Models

| Model   | Normal Models | Lumped Models |
|---------|---------------|---------------|
|         | Train         | Test          | Total          |
|         | $APRE,\%$ | $AAPRE,\%$ | $rmse, SCF/STB$ | $R^2$ | $SD$ | $APRE,\%$ | $AAPRE,\%$ | $rmse, SCF/STB$ | $R^2$ | $SD$ |
|RF      | -0.1705      | 0.9390       | 0.0425         | 0.9866 | 0.0226 | -0.2476    | 1.0424       | 0.0541         | 0.9745 | 0.0304 |
|        | -0.0778      | 1.2312       | 0.0645         | 0.9703 | 0.0285 | -0.0279    | 1.4002       | 0.0430         | 0.9797 | 0.0250 |
|        | -0.0406      | 1.2132       | 0.0261         | 0.9944 | 0.0174 | -0.0438    | 1.3511       | 0.0342         | 0.9929 | 0.0187 |
|ET      | -0.5869      | 1.4146       | 0.0457         | 0.9846 | 0.0312 | -0.3706    | 1.5790       | 0.0572         | 0.9709 | 0.0375 |
|        | -0.0250      | 1.3243       | 0.0345         | 0.9908 | 0.0207 | -0.0376    | 1.5109       | 0.0370         | 0.9898 | 0.0237 |
|        | -0.0249      | 1.4292       | 0.0331         | 0.9918 | 0.0208 | 0.0179     | 1.7540       | 0.0399         | 0.9860 | 0.0274 |
|GRNN    | -0.0745      | 1.8705       | 0.0452         | 0.9835 | 0.0301 | -0.0437    | 1.9824       | 0.0672         | 0.9711 | 0.0331 |
|        | -0.0439      | 1.6833       | 0.0409         | 0.9882 | 0.0237 | -0.3377    | 1.8155       | 0.0565         | 0.9611 | 0.0329 |
|        | -0.1946      | 1.5354       | 0.0353         | 0.9904 | 0.0221 | 0.2221     | 1.6220       | 0.0475         | 0.9830 | 0.0269 |
|        | 0.1056       | 1.7019       | 0.0433         | 0.9853 | 0.0261 | -0.0174    | 1.8199       | 0.0467         | 0.9846 | 0.0275 |
|MLP-BR  | -0.0692      | 1.6991       | 0.0408         | 0.9878 | 0.0272 | -0.0188    | 1.8507       | 0.0435         | 0.9824 | 0.0312 |
|        | -0.0779      | 0.9600       | 0.0395         | 0.9894 | 0.0197 | -0.1257    | 1.0426       | 0.0250         | 0.9898 | 0.0178 |
|        | -0.1409      | 1.3422       | 0.0966         | 0.9293 | 0.0430 | 0.0768     | 1.4343       | 0.0510         | 0.9793 | 0.0253 |
|        | -0.0419      | 1.1404       | 0.0248         | 0.9954 | 0.0161 | -0.0912    | 1.2785       | 0.0320         | 0.9915 | 0.0190 |
|        | -0.5748      | 1.5091       | 0.0399         | 0.9852 | 0.0311 | -0.3443    | 1.5619       | 0.0710         | 0.9615 | 0.0398 |
|        | -0.0295      | 1.3155       | 0.0554         | 0.9900 | 0.0212 | -0.0955    | 1.6516       | 0.0393         | 0.9602 | 0.0230 |
|        | 0.0017       | 1.5235       | 0.0343         | 0.9915 | 0.0203 | 0.0815     | 1.6556       | 0.0319         | 0.9891 | 0.0235 |
|        | -0.0643      | 1.8623       | 0.0343         | 0.9842 | 0.0253 | -0.0507    | 2.0618       | 0.0583         | 0.9793 | 0.0297 |
|        | -0.0362      | 1.5207       | 0.0441         | 0.9862 | 0.0223 | -0.0138    | 1.7192       | 0.0527         | 0.9688 | 0.0303 |
|        | -0.0334      | 1.5675       | 0.0423         | 0.9871 | 0.0226 | -0.1708    | 1.6711       | 0.0352         | 0.9873 | 0.0260 |
|        | -0.0127      | 1.7776       | 0.0449         | 0.9852 | 0.0258 | 0.1701     | 1.9605       | 0.0392         | 0.9856 | 0.0275 |
|MLP-LM  | -0.0774      | 1.8730       | 0.0464         | 0.9837 | 0.0259 | 0.2079     | 2.0637       | 0.0485         | 0.9812 | 0.0274 |
|        | -0.1859      | 0.9597       | 0.0451         | 0.9844 | 0.0243 | -0.0678    | 1.2650       | 0.0608         | 0.9717 | 0.0278 |
|        | -0.0412      | 1.2408       | 0.0279         | 0.9940 | 0.0176 | -0.5436    | 1.4475       | 0.0482         | 0.9822 | 0.0326 |
|        | -0.0436      | 1.3615       | 0.0350         | 0.9906 | 0.0213 |
|        | -0.0086      | 1.5005       | 0.0346         | 0.9908 | 0.0229 |
|        | -0.0683      | 1.8928       | 0.0503         | 0.9806 | 0.0307 |
|        | -0.1027      | 1.7098       | 0.0445         | 0.9848 | 0.0258 |
|        | -0.1115      | 1.5527       | 0.0381         | 0.9889 | 0.0231 |
|        | 0.0878       | 1.7090       | 0.0440         | 0.9849 | 0.0252 |
|        | -0.0592      | 1.7294       | 0.0413         | 0.9869 | 0.0280 |
|        | 0.0473       | 0.9150       | 0.0418         | 0.9838 | 0.0241 |
|        | 0.1701       | 1.9605       | 0.0410         | 0.9873 | 0.0262 |
|        | 0.0238       | 1.8141       | 0.0439         | 0.9853 | 0.0261 | 0.0205     | 1.9110       | 0.0468         | 0.9832 | 0.0262 |
Finding the optimum network structure is necessary to make the MLP network reliable.\(^{35}\) Generally, a one-hidden-layer MLP is appropriate for solving most problems. Nevertheless, for more complex tasks, more hidden layers can be employed.\(^{36}\) It is noteworthy that although using a large number of hidden layers/neurons can improve the model’s precision, in the case of too many hidden layers/neurons, overfitting may occur, and the model fails to provide reliable predictions for the testing dataset. The other elements in an MLP network are the links or weights, which connect the neurons of a layer to the neurons of a subsequent layer. A transfer function is employed to pass the summation value, which is calculated by adding a bias term to multiplication of the summation of the values of preceding neurons and a particular weight.\(^{34,35}\) To make it clear, the output value of a one-hidden-layer network with tansig and purelin transfer functions in hidden and output layers can be expressed as

$$\text{Net. Output} = \text{purelin}(w_2 \times (\text{tansig}(w_1 \times x) + b_1) + b_2)$$  \hspace{1cm} (7)

where \(x\) is the value of an input vector and \(w_1, w_2, b_1, \) and \(b_2\) denote the weight matrices and bias vectors for hidden and output layers, respectively. Various training algorithms are developed for adjustment of the weights and bias, among which the BR, LM, and SCG are the most popular ones.\(^{37}\) The flowchart of an MLP network development process is presented in Figure 3.

### 3.5. Cascade-Forward Back-Propagation Network
The CFBPN algorithm was first proposed by Fahlman and Lebiere.\(^{38}\) They hoped to enhance the network’s ability in catching the relation between input and output parameters by creating more connections among layers. The fundamental difference between the structure of a cascade network and conventional ANNs is the number of links between layers. In the cascade algorithm, each layer is linked to all previous layers, while each layer is linked only to its preceding layer in conventional ANNs.\(^ {39}\) Usually, CFBPNs can provide more accurate results with a lower number of hidden neurons in comparison to conventional ANNs.\(^ {40}\) However, the higher number of adjustable weights may result in higher computational costs.\(^ {41}\) Figure 4 depicts a typical CFBPN.

### 3.6. Radial Basis Function Neural Network
An RBF network that can be applied to both classification and regression problems is derived from the notion of function approximation.\(^{35,32}\) This approach was first developed by Broomhead and Lowe\(^{42}\) in 1988. The main difference of this algorithm in comparison to conventional networks is the way they process the data. The RBF maps the input data into a feature space of a higher dimension using activation functions, instead of using transfer functions. Also, there is always one hidden layer in this network. Therefore, hidden neurons’ quantity is the only remaining parameter to be optimized, which is lower than or equal to the quantity of training points. In the hidden layer, neurons have specified radiuses and are distributed in the feature space. Then, the Euclidean distance measures the distance between the center of hidden neurons and input vectors, as shown in eq \(^{42}\)

$$r_j = \sqrt{\sum_{i=1}^{n} (x_i - C_j)^2} j = 1, 2, ..., N$$  \hspace{1cm} (8)

where \(x\) and \(C\) are the input and center vectors, respectively. \(N\) is equivalent to the number of input variables, while \(N\) is lower than or equal to the training points. Then, one of the following radial basis transfer functions is utilized to transfer the calculated Euclidean distance to the output\(^{35}\)

- **Linear function**: \(\phi_i = r\)

- **Cubic function**: \(\phi_i = r^3\)

- **Gaussian function**: \(\phi_i = \exp\left(\frac{r^2}{2\sigma^2}\right)\)

- **Multi-quadratic function**: \(\phi_i = (r^2 + \sigma^2)^{1/2}\)

The Gaussian function shows a smoother behavior and has been extensively employed. As a matter of fact, there are typically two adjustable variables in an RBF network: the Gaussian function’s spread coefficient (\(\sigma\)) and the number of hidden neurons. The radius of hidden neurons depends on the value of the spread coefficient. Considering numerous hidden neurons may result in a complex model with a low error and low execution rate. On the other hand, employing a large spread coefficient results in a small model with a greater error. A model with a small spread coefficient may precisely predict the training data but may fail to provide reliable answers for the testing data.\(^ {43}\) Hence, in order to have a reliable RBF model, the value of these two parameters should be optimized. Figure 5 depicts the RBF network illustration.

### 3.7. Generalized Regression Neural Network
The GRNN uses the notions of both probabilistic and RBF networks. Specht\(^ {44}\) developed the GRNN algorithm in a way that it does not involve an iterative process to identify nonlinear relationships. Similar to probabilistic networks, the GRNN has a one-pass learning algorithm that needs a few training samples to generalize fast.\(^ {45}\) The main structural difference between the GRNN and RBF network is its extra layer, namely, the summation layer, which connects the hidden layer to the output layer (as depicted in Figure 6). Also, the number of hidden neurons is equivalent to the training points in the GRNN. Hence, the spread coefficient of the Gaussian function is the only parameter that remains in the GRNN to be optimized. In the summation layer, there are two kinds of neurons which calculate the weighted (\(S\)) or algebraic (\(D\)) sum of data through the equations mentioned below\(^ {46}\)

$$D = \sum_{i=1}^{n} e^{-\left(\frac{E^2}{2\sigma^2}\right)}$$  \hspace{1cm} (13)

### Table 7. Time and Memory Assessment for Each Modeling Approach

| Model   | Memory (Mb) | Time (s) |
|---------|-------------|----------|
| DT      | 313         | 0.28     |
| ET      | 339         | 0.99     |
| RF      | 295         | 0.62     |
| GRNN    | 214         | 0.19     |
| RBF     | 274         | 0.41     |
| MLP-LM  | 72          | 0.75     |
| MLP-BR  | 76          | 16.3     |
| MLP-SCG | 63          | 2.39     |
| CFBPN-LM| 31          | 1.23     |
| CFBPN-BR| 36          | 1.58     |
| CFBPN-SCG| 11         | 0.82     |
Figure 7. Cross-plots of the proposed normal $B_0$ models; (a) RF, (b) DT, (c) EsT, (d) GRNN, (e) CFBPN-BR, (f) CFBPN-LM, (g) CFBPN-SCG, (h) RBF, (i) MLP-BR, (j) MLP-LM, and (k) MLP-SCG.
Figure 8. Cross-plots of the proposed lumped $B_p$ models; (a) RF, (b) DT, (c) ET, (d) GRNN, (e) CFBPN-BR, (f) CFBPN-LM, (g) CFBPN-SCG, (h) RBF, (i) MLP-BR, (j) MLP-LM, and (k) MLP-SCG.
4. RESULTS AND DISCUSSION

This study represented the very first attempt toward compositional modeling of oil formation \( B_o \) using various advanced and powerful intelligent systems, namely, the ETs, RF, DTs, GRNNs, and CFBPN and MLP and RBF neural networks. The modeling was done based on oil composition, reservoir temperature and pressure, and \( C_{12+} \) characteristics (SG and MW). The models were developed based on a comprehensive databank consisting of 1224 data points. Moreover, two kinds of models were developed. The first kind of model had 18 input variables, namely, reservoir temperature and pressure, oil composition (\( H_2S, N_2, CO_2, \) and \( C_1 \) to \( C_{12} \)), SG and MW of the \( C_{12+} \) fraction, and one target parameter, that is, \( B_o \). On the other hand, the mole fractions were regrouped into three clusters for the second type of model. Therefore, these models had seven input variables. Accordingly, the models developed based on the first scheme were called “Normal Models”, and the other ones were called “Lumped Models”. The proficiency of developed models was then compared with seven different EoSs, namely, 2-parameter Peng–Robinson (PR), Redlich–Kwong (RK), 2-parameter Soave–Redlich–Kwong (SRK), Zudkevitch–Joffe (ZJ), 3-parameter Peng–Robinson (PR3), 3-parameter Soave–Redlich–Kwong (SRK3), and Schmidt–Wenzel (SW). The upcoming sections discuss the development phase of each model and efficiency assessment benchmarks and the modeling outcomes. At the end of this section, the outlier detection approach is discussed.

4.1. Model Development. As aforementioned, the normal and lumped models were, respectively, developed with 18 and 7 inputs. Throughout the development process, the whole dataset was split with a proportion of 4/1. Hence, the models had to be trained using 80 percent of datapoints, and then, the remaining 20 percent was employed to test the models’ efficiency. Each model was developed several times, and the optimum values attained for the hyperparameters of each model are presented in Table 5. Moreover, the efficiency of various training algorithms in training the CFBPN and MLP network was assessed, namely, the SCG, BR, and LM. It is noteworthy that the hyperparameters of the tree-based models were optimized by the Bayesian optimization approach.

4.2. Performance Evaluation. In order to statistically evaluate the performance of the developed models, a variety of statistical indexes, namely, the root mean square error (rmse), average percent relative error (APRE), standard deviation (StD), AAPRE, and correlation coefficient (\( R^2 \)), were employed. The equations of the considered errors are provided here:\textsuperscript{14,49}

\[
S = \sum_{i=1}^{n} \frac{y_i e^{-\frac{e^2}{2\sigma^2}}}{\sum_{i=1}^{n} e^{-\frac{e^2}{2\sigma^2}}} \quad (14)
\]

where Euclidean distance and the spread coefficient are, respectively, represented by \( E \) and \( \sigma \) and \( y_i \) exhibits the output for \( i \)th training data. Then, the division of the results of the summation neurons determines the model’s output as:\textsuperscript{18}

\[
\text{Output}_i = \frac{\sum_{j=1}^{m} y_{ij} e^{-\frac{e^2}{2\sigma^2}}}{\sum_{j=1}^{m} e^{-\frac{e^2}{2\sigma^2}}} \quad (15)
\]

Table 8. Statistical Evaluation of EoSs in Estimating \( B_o \)

| EoSs       | PR  | SRK | RK  | ZJ  | PR3 | SRK3 | SW  |
|------------|-----|-----|-----|-----|-----|------|-----|
| APRE, %    | 8.3609 | 7.5739 | −4.1484 | 2.4926 | −2.2982 | −4.0382 | 3.5853 |
| AAPRE, %   | 8.5179 | 7.9151 | 8.7154 | 4.8136 | 4.2345 | 5.2804 | 4.9487 |
| rmse, SCF/STB | 0.1824 | 0.1708 | 0.3254 | 0.1590 | 0.1095 | 0.1426 | 0.1468 |
| \( R^2 \)  | 0.7452 | 0.7767 | 0.1894 | 0.8064 | 0.9083 | 0.8444 | 0.8350 |
| Std        | 0.0993 | 0.0944 | 0.1810 | 0.0719 | 0.0588 | 0.0771 | 0.0833 |
In addition to statistical evaluations, graphical error analysis was performed to provide visual assessment and comparison. To this end, different kinds of plots including cross-plots, error distribution plots, error bar charts, and cumulative frequency plots were sketched. In cross-plots, the predicted/estimated values are depicted versus experimental ones. In such plots, the

\[ \text{rmse : } E_{\text{rmse}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{i, \text{exp}} - y_{i, \text{pred}})^2} \]  

\[ \text{StD : } E_{\text{StD}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{y_{i, \text{exp}} - y_{i, \text{pred}}}{y_{i, \text{exp}}} \right)^2} \]  

\[ R^2 : E_{R^2} = 1 - \frac{\sum_{i=1}^{n} (y_{i, \text{exp}} - y_{i, \text{pred}})^2}{\sum_{i=1}^{n} (y_{i, \text{exp}} - \bar{y})^2} \]  

Figure 10. Cross-plots of different EoSs in predicting \( B_o \): (a) PR, (b) SRK, (c) RK, (d) ZJ, (e) PR3, (f) SRK3, and (g) SW.
more compact zone around the unit slope line reflects the more accurate model. The error distribution plots show the relative error against the experimental value of the target parameter. If points are distributed in a smaller zone near the zero line, the model is more accurate. The error bar charts provide a visual comparison between the efficiency of the developed models. The cumulative frequency plots show the cumulative frequency against the absolute relative error.

4.2.1. Intelligent Models. The proposed intelligent $B_o$ models were developed based on normal and lumped schemes. The modeling was done using 1224 data points. The statistical evaluations of the developed normal and lumped $B_o$ models are presented in Table 6. Overall, the table expresses that all intelligent models were able to provide reliable estimations for both normal and lumped datasets. However, when the number of inputs decreases, the accuracy of the developed models reduces except for RBF and ET models. As the table reflects, tree-based models showed the lowest errors in terms of AAPRE with the outperformance of the RF model representing total AAPRE values of 0.96 and 0.98% for the normal and lumped datasets, respectively. For both normal and lumped cases, the developed ET models came next with a corresponding value of 1.24 and 1.17%, respectively. The third places in both normal and lumped cases go to DT models with a corresponding total AAPRE of 1.26 and 1.36%, respectively. From another point of view, ET models were the most efficient models in terms of $R^2$, rmse, and StD between both normal and lumped models. Among different kinds of neural networks, CFBPN-BR and GRNN models were the most efficient ones in normal and lumped cases, respectively. The results also indicated that CFBPN models are generally more rigorous than MLP networks. It should be noted that CFBPN models had a simpler structure in comparison to MLP networks. In addition, the BR training algorithm showed more efficiency if training both CFBPN and MLP networks, followed by the LM algorithm.

Alongside the statistical error evaluation, the efficiency of the models was assessed with respect to time and memory occupied by each specific run. The average time consumed by each run was calculated after 100 runs and is reported in Table 7. As the table reflects, despite the fact that kernel-based models and tree-based models showed a higher rate of consumed memory than CFBPN and MLP models, they were remarkably faster and able to be trained in much less time.

In order to graphically assess the performance of the proposed $B_o$ models, the cross-plots of the developed normal and lumped models are illustrated in Figures 7 and 8, respectively. Overall, the figures demonstrate that the performance of each intelligent approach was almost the same for normal and lumped models. Furthermore, the plots show that all models could provide precise and reliable predictions for $B_o$ since the points are well scattered around the unit slope line. As cross-plots are typically affected by correlation coefficient and rmse values, the cross-plots for RF and DT models illustrate a relatively weak agreement between the estimated and measured points since these two approaches did not provide good results in terms of rmse and $R^2$, as was shown in statistical evaluation. On the other hand, the cross-plots of the ET models confirm their higher
performance over the other proposed models. The figures also reveal that the ET models were the only models that could provide accurate predictions for all ranges of $B_o$. Thus, the results of both statistical and graphical error assessments verify that the ET models are the most efficient models for predicting $B_o$ in both normal and lumped cases.

As can be seen in the cross-plots, the test samples vary for different subplots as in the development phase of each predictive model, the whole dataset was randomly divided with a ratio of 80/20. To be more precise, when the models’ error was not satisfying, these processes (models’ training and random splitting of data) would be repeated, and this recursive approach would continue until a satisfying error was reached. Accordingly, based on each ML algorithm, numerous predictive models were developed, and the best model in terms of error was selected and reported as the final model.

The error distribution plots for normal and lumped ET models are portrayed in Figure 9. It is clearly illustrated that the points located far from the zero-error line are scarce, and a high concentration of points can be seen in a short range near the zero-error line, which again verifies the reliability of these models. In Figure 9a, it is shown that only 26 points (2.1% of the total dataset) are estimated with a relative error of greater than 10%. Moreover, the highest absolute relative deviation for PR3 was approximately 30%.

4.2.2. Equations of State. The aforementioned EoSs were employed again to estimate the oil formation $B_o$ using the PVTi module of Eclipse commercial software. Table 8 summarizes the detailed statistical evaluation of the employed EoSs in estimating $B_o$.

According to the table, PR3 EoSs could provide the most accurate predictions for $B_o$ with AAPRE, $R^2$, and rmse values of 4.23%, 0.84, and 0.14 SCF/STB, respectively. After this, ZJ and SW showed the lowest APPRE values of 4.81 and 4.95%, respectively. From another viewpoint, SRK3 could take the second place in terms of rmse and $R^2$. By contrast, PR and RK were the least accurate EoSs in estimating $B_o$ with AAPRE values of 8.52 and 8.71%, respectively. To graphically illustrate the efficiency of EoSs in predicting $B_o$, Figure 10 represents the obtained cross-plot for each EoS. The cross-plots also confirm the higher capability of PR3 over other EoSs in estimating $B_o$ as the points are scattered in a smaller area around the unit slope line in comparison to other plots.

Moreover, the error distribution plot for PR3, as the most accurate EoS, is illustrated in Figure 11. The plots demonstrate that 129 data points (10.5% of the total dataset) estimated using PR3 EoS showed an absolute relative deviation of higher than 10%. Moreover, the highest absolute relative deviation for PR3 was approximately 30%.

4.2.3. Comparison between Intelligent Models and EoSs in Predicting $B_o$. The comparison between the error evaluations of the developed intelligent models and preexisting EoS clearly verifies the higher performance of the intelligent models. Even the least accurate intelligent approach (lumped MLP-SCG) showed about a half value of the AAPRE that was shown by the most accurate EoS (PR3) in predicting $B_o$. It should be mentioned that the most accurate and reliable intelligent approach for both normal and lumped $R$ models was ET models with AAPRE values of 1.24 and 1.17%, respectively, while the most efficient EoS showed an AAPRE of 4.23%. Another equally important point worth mentioning is that the maximum absolute relative error showed by normal and lumped ET models was 13.11 and 12.44%, respectively, while the corresponding value for SW was approximately 30.23%. Moreover, Figure 12 shows the error bar chart of the best-established normal and lumped intelligent models in comparison to different EoSs with respect to AAPRE, rmse, and $R^2$ values. The figure provides a clear graphical comparison, which shows the outperformance of the intelligent models in all error indices.

In addition to bar charts, the cumulative frequency plot is provided for the most efficient normal and lumped intelligent models (ET-based models), and the best EoS (PR3) is sketched in Figure 13. According to this plot, the majority of points were predicted by intelligent models with much less error. Normal and lumped ET models were able to estimate over 95% of points with an error of lower than 4%, whereas the PR3 EoS could forecast only 13% of the points with such errors.

Therefore, through a variety of statistical and graphical error evaluations, the superiority of the developed intelligent models, specifically ET models, in estimating oil formation $B_o$ was confirmed in comparison to seven employed EoSs. It should be mentioned that the data-driven predictive models are to predict the target parameter using the input parameters within the ranges of training samples, and it is not expected that they work well for the data far from the training samples. The aim of using different subsets is to avoid overfitting. If a model overfits, it can precisely estimate the training targets, but it fails to predict the testing targets.
4.2.4. Bo Models’ Performance in Different Temperature Ranges. Figure 14 portrays the group error plot for four ranges of temperature to investigate the efficiency of the proposed ET models in various temperature ranges. Along with ET models, the efficiency of PR3, as the most accurate EoS, is sketched as well. Overall, the ET-based models exhibited a relatively consistent performance. From the statistical point of view, ET models could provide less accurate predictions for temperatures higher than 260 °F, while their performance in other ranges of temperature was nearly the same. In the case of EoSs, the Bo points that had a temperature higher than 260 °F were predicted with higher errors. The results of this figure verify the reliability of the proposed ET models in predicting Bo. The figure also confirms that the ET models can provide precise predictions for Bo in all ranges of temperature. Since the outperformance of ET models was approved, they were kept for further evaluations.

4.3. Trend Analysis. In modeling approaches, it is very important to check if the developed models are able to recognize the physical trend of the output and an input variable.\textsuperscript{50} With the aim of evaluating the capability of the developed ET models in spotting the existing physical trend between the Bo and pressure, Figure 15 illustrates the estimated and measured Bo values against corresponding pressure. The estimations of PR3, the most efficient EoS, are sketched as well. In this figure, three plots are represented which are sketched for constant compositions and temperatures. Therefore, the only remaining parameter that can affect the value of Bo is pressure. Plot (a) in this figure is sketched for an oil sample with a reservoir temperature of 170 °F, while plots (b) and (c) show the trend of Bo at constant temperatures of 207.3 °F and 240 °F, respectively.

The figure evidently demonstrates the capability of the proposed ET models in reproducing the physical trend of the problem at different temperatures. On the other hand, SW was not as precise as intelligent models in recognition of this physical trend. It also can be inferred from the figures that Bo increases at a low rate as the pressure declines above saturation pressure (Pb). This is because no gas has come out from the fluid at pressures higher than Pb, and pressure decrease leads to oil expansion, which results in volume increase. Once the pressure falls below Pb, further pressure reduction results in Bo decrease at a relatively higher rate since the gas is evolved from the fluid, and oil volume decreases. The observed trend is in agreement with the general knowledge.

4.4. Outlier Detection and the Applicability Domain of the Proposed Bo Models. In this research, the Leverage approach, a reliable outlier detection method,\textsuperscript{51,52,53} was applied to assess the reliability of the used data and to appraise the applicability realm of the proposed normal and lumped ET-based Bo models. In this approach, the datapoints located far from the bulk data are detected as outliers.\textsuperscript{52} The main principles of the Leverage approach are the hat matrix (H), Leverage limit (H*), and standardized residuals (SRs). These principles are calculated as\textsuperscript{53}

\[
SR = \frac{z_i}{(MSE(1 - H))^{0.5}}
\]  

Figure 15. Measured and predicted values of Bo for three oil samples at (a) 170 °F, (b) 207.3 °F, and (c) 240 °F.
where $z_i$, MSE, and $H_{ii}$, respectively, denote the error, mean square error, and hat indices of the $i$th data point. $I$ equals the number of inputs plus one, and $N$ expresses the total number of data points. $X$ is a two-dimensional $N \times b$ matrix ($b$ shows the model’s dimension), while $t$ means the transpose matrix.\textsuperscript{53,54}

The Leverage approach provides a visual representation, the so-called Williams plot, which makes it easy to interpret the results. This technique categorizes points into “Good/Bad High Leverage”, “Lower/Upper Suspected Data”, and “Valid Data”, and each category is located in a particular zone of the Williams plot. The points that have a hat value lower than the Leverage limit exist in the applicability domain of the model. The points having SR values higher than 3 or lower than $-3$ are not predicted well. Therefore, the points located before the Leverage limit and in the interval of $-3 < \text{SR} < 3$ are referred to as “Valid Data”.\textsuperscript{53}

In the case of developed normal and lumped ET models for predicting $B_o$, the Leverage limit was obtained as 0.0466 and 0.0196, respectively. The Williams plots for the proposed ET-based $B_o$ models are illustrated in Figure 16. The figure clearly reflects that the majority of points were valid. For the normal model, only 49 points (out of 1224) were not inside the “Valid Data” region. Only 17 points (1.39% of the total dataset) were detected as outliers. Also, 25 points (2.12% of the total dataset) were found to be in the “Good High Leverage” zone, and 7 points (0.57% of the total dataset) could be seen in the “Bad High Leverage” zone. Moreover, in the case of the lumped ET model, 55 points (out of 1224) were outside of the “Valid Data” zone. Among these points, 33 points (2.70% of the whole dataset) were detected as outliers, and 22 remaining points (1.80% of the whole dataset) could be seen in the “Good High Leverage” zone. It is noteworthy that no point was seen inside the “Bad High Leverage” zone. Hence, the consistency of the used databank was confirmed by the Leverage approach.

5. CONCLUSIONS

In this communication, various compositional models for predicting oil formation $B_o$ were developed based on various advanced and powerful intelligent models, namely, the ETs, RF, DTs, GRNN, and CFBPN and MLP and RBF neural networks. Moreover, the efficiency of three different training algorithms, that is, the BR, SCG, and LM, was assessed in training the CFBPN and MLP network. Additionally, the hyperparameters of the proposed tree-based models, that is, ET, RF, and DT
models, were adjusted by the Bayesian optimization method. The input vectors of all models contained values for reservoir temperature and pressure, oil composition (H, C, O, N, and S), and properties (SG and MW). Prior to the models’ training phase, the datasets were prepared based on two different schemes. In the first type of dataset, the mole percent of each fraction was used as a distinct input parameter, while the fractions’ mole percent was regrouped into three clusters in the other type of dataset. To the best of our knowledge, it was the first time that ML models have been employed for compositional modeling of $B_o$. In addition to intelligent models, seven EoSs were employed to estimate $B_o$. The performance of the proposed models and EoSs was evaluated using various graphical and statistical error assessments. Furthermore, the efficiency of the developed models at various temperatures was evaluated. Also, a trend analysis was performed to evaluate the power of the proposed models in detecting the physical trend of the target parameters. Finally, the outlier detection was performed using the Leverage approach to assess the models’ applicability domain and detect outliers. What follows highlights the most important points that were raised by this study:

1. Intelligent modeling techniques are reliable approaches to estimate oil PVT properties, and alongside their individual use to estimate these properties in a fraction of a second, they can be employed in commercial simulators for large-scale simulations to significantly reduce the computational burden according to their computational cost, which is much lower than that of EoSs.
2. The ET could outperform the other models providing the most consistent predictions for both lumped and normal datasets $B_o$. 
3. PR3 was the most efficient EoS in estimating $B_o$.
4. It was shown that the EoSs were defeated by even the weakest intelligent approaches.
5. The Leverage approach clearly confirmed the validity of the used databanks and applicability domain of the proposed models.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c01466.

Sample of the gathered databank containing 180 data-points from 10 oil samples (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Aydin Larestani — Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman 76169-14111, Iran; Email: aydilarestani@gmail.com
Abdolhossein Hemmati-Sarapardeh — Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman 76169-14111, Iran; College of Construction Engineering, Jilin University, Changchun 130012, China; Email: hemmati@uk.ac.ir, aut.hemmati@gmail.com
Mehdi Ostadkhassan — Key Laboratory of Continental Shale Hydrocarbon Accumulation and Efficient Development, Ministry of Education, Northeast Petroleum University, Daqing 163318, China; Institute of Geosciences, Marine and Land Geomechanics and Geotectonics, Christian-Albrechts-University, Kiel 24118, Germany; Department of Geology, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran; Email: mehdi.ostadhassan@nepu.edu.cn, mehdi.ostadhassan@ifg.uni-kiel.de

Author

Zahra Samari — Department of Physical Chemistry, Faculty of Chemistry, Kharazmi University, Tehran 15614, Iran

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acsomega.2c01466

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Ali Naseri and the Research Institute of Petroleum Industry (RIPI) for providing the databank.

REFERENCES

(1) Mahdiani, M. R.; Kooti, G. The most accurate heuristic-based algorithms for estimating the oil formation volume factor. *Petroleum 2016*, *2*, 40–48.
(2) Arabloo, M.; Amoioe, M.-A.; Hemmati-Sarapardeh, A.; Ghazanfari, M.-H.; Mohammadi, A. H. Application of constrained multi-variable search methods for prediction of PVT properties of crude oil systems. *Fluid Phase Equilib.* 2014, *363*, 121–130.
(3) (a) Baniasadi, H.; Kamari, A.; Heidaramarbi, S.; Mohammadi, A. H.; Hemmati-Sarapardeh, A. Rapid method for the determination of solution gas-oil ratios of petroleum reservoir fluids. *J. Nat. Gas Sci. Eng.* 2015, *24*, 500–509. (b) Dey, P.; Deb, P. K.; Akhter, S.; Dey, D. Reserve estimation of sandalani gas field. *Int. J. Innovation Appl. Stud.* 2016, *16*, 166. (c) Dindoruk, B.; Christman, P. G. PVT properties and viscosity correlations for Gulf of Mexico oils. *SPE annual technical conference and exhibition*; Society of Petroleum Engineers, 2001. (d) Dutta, S.; Gupta, J. P. PVT correlations for Indian crude using artificial neural networks. *J. Pet. Sci. Eng.* 2010, *72*, 93–109. (e) Gharbi, R.; Elsharkawy, A. Predicting the bubble-point pressure and formation-volume-factor of worldwide crude oil systems. *Pet. Sci. Technol.* 2003, *21*, 53–79. (f) Mahdiani, M. R.; Khamehchi, E. A novel model for predicting the temperature profile in gas lift wells. *Petroleum 2016*, *2*, 408–414. (g) Malallah, A.; Gharbi, R.; Algharaib, M. Accurate estimation of the world crude oil PVT properties using graphical alternating conditional expectation. *Energy Fuels* 2006, *20*, 688–698. (b) Osman, E.; Abdel-Wahhab, O.; Al-Marhoun, M. Prediction of oil PVT properties using neural networks.SPE Middle East Oil Show; Society of Petroleum Engineers, 2001. (i) Rafiee-Taghanaki, S.; Arabloo, M.; Chamkalani, A.; Amani, M.; Zargari, M. H.; Adelzadeh, M. R. Implementation of SVM framework to estimate PVT properties of reservoir oil. *Fluid Phase Equilib.* 2013, *346*, 25–32. (j) Valkó, P. P.; McCain, W. D., Jr. Reservoir oil bubblepoint pressures revisited; solution gas-oil ratios and surface gas specific gravities. *J. Pet. Sci. Eng.* 2003, *37*, 153–169.
(4) Elsharkawy, A. M. Modeling the properties of crude oil and gas systems using RBF network.SPE Asia Pacific oil and gas conference and exhibition; OnePetro, 1998.
(5) Ghassemi, F. Generalized pressure-volume-temperature correlations. *J. Pet. Technol.* 1980, *32*, 785–795.
(6) Dake, L. P. Fundamentals of reservoir engineering: Elsevier, 1983.
(7) (a) Tohidi-Hosseini, S.-M.; Hajiarezaie, S.; Hashemi-Doulatabadi, M.; Hemmati-Sarapardeh, A.; Mohammadi, A. H. Toward prediction of petroleum reservoir fluids properties: A rigorous model for estimation of solution gas-oil ratio. *J. Nat. Gas Sci. Eng.* 2016, *29*, 506–516. (b) Zamani, H. A.; Rafiee-Taghanaki, S.; Karimi, M.; Arabloo, M.; Dadashi, A. Implementing ANFIS for prediction of reservoir oil solution gas-oil ratio. *J. Nat. Gas Sci. Eng.* 2015, *25*, 325–334. (c) Assoodeh, M.; Bagheripour, P. Estimation of bubble point pressure from PVT data using a power-law committee with intelligent systems. *J.
multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict solution gas-oil ratio of crude oil systems. Petroleum 2020, 6, 80–91.

(9) Katz, D. L. Prediction of the shrinkage of crude oils. Drilling and Production Practice; American Petroleum Institute, 1942.

(10) Standing, M. A pressure-volume-temperature correlation for mixtures of California oils and gases. Drilling and Production Practice; American Petroleum Institute, 1947.

(11) Vasquez, M.; Beggs, H. D. Correlations for fluid physical property prediction. J. Pet. Technol. 1980, 32, 968–970.

(12) (a) Larestani, A.; Mousavi, S. P.; Hadavimoghaddam, F.; Hemmati-Sarapardeh, A. Predicting formation damage of oil fields due to mineral scaling during water-flooding operations: Gradient boosting decision tree and cascade-forward back-propagation network. J. Pet. Sci. Eng. 2022, 208, 109315. (b) Aamar, M. N.; Larestani, A.; Lv, Q.; Zhou, T.; Hemmati-Sarapardeh, A. Modeling of methane adsorption capacity in shale gas formations using white-box supervised machine learning techniques. J. Pet. Sci. Eng. 2022, 208, 109226. (c) Amir-Ramsheh, B.; Safaei-Farouji, M.; Larestani, A.; Zabihi, R.; Hemmati-Sarapardeh, A. Modeling of wax disappearance temperature (WDT) using soft computing approaches: Tree-based models and hybrid models. J. Pet. Sci. Eng. 2022, 208, 109774. (d) Hashemizadeh, A.; Maaref, A.; Shateri, M.; Larestani, A.; Hemmati-Sarapardeh, A. Experimental measurement and modeling of water-based drilling mud density using adaptive boosting decision tree, support vector machine, and K-nearest neighbors: A case study from the South Pars gas field. J. Pet. Sci. Eng. 2021, 207, 109132. (e) Mahdaviara, M.; Larestani, A.; Aamar, M. N.; Hemmati-Sarapardeh, A. On the evaluation of permeability of heterogeneous carbonate reservoirs using rigorous data-driven techniques. J. Pet. Sci. Eng. 2022, 208, 109685. (f) Zanganeh Kamali, M.; Davoodi, S.; Ghorbani, H.; Wood, D. A.; Mohamadian, N.; Lajmorak, S.; Rukavishnikov, V. S.; Taherizade, F.; Band, S. S. Permeability prediction of heterogeneous carbonate-gas-carbonate reservoirs applying group method of data handling. Mar. Pet. Geol. 2022, 139, 105597. (g) Mosavi, A.; Faghan, Y.; Ghamsi, P.; Duan, P.; Ardabili, S. F.; Salwana, E.; Band, S. S. Comprehensive review of deep reinforcement learning methods and applications in economics. Mathematics 2020, 8, 1640. (h) Nabipour, N.; Mosavi, A.; Hajnal, E.; Nadai, L.; Shamshirband, S.; Chau, K.-W. Modeling climate change impact on wind power resources using adaptive neuro-fuzzy inference system. Eng. Appl. Comput. Fluid Mech. 2020, 14, 491–506. (i) Nourani, M.; Alali, N.; Hpartizanlou, B.; Band, S. S.; Chau, K.-w.; Shu, C.-M. Comparison of machine learning techniques for predicting porosity of chalk. J. Pet. Sci. Eng. 2022, 209, 108853. (j) Shamshirband, S.; Mosavi, A.; Rabczuk, T.; Nabipour, N.; Chau, K.-w. Prediction of significant wave height; comparison between nested grid numerical model, and machine learning models of artificial neural networks, extreme learning and support vector machines. Eng. Appl. Comput. Fluid Mech. 2020, 14, 805–817. (k) Torabi, M.; Hashemi, S.; Saybani, M. R.; Shamshirband, S.; Mosavi, A. A Hybrid clustering and classification technique for forecasting short-term energy consumption. Environ. Prog. Sustain. Energy 2019, 38, 66–76.

(13) Gharb, B.; Elsharkawy, A. M. Neural network model for estimating the PVT properties of Middle East crude oils. Middle East Oil Show and Conference; Society of Petroleum Engineers, 1997.

(14) Mahdiani, M. R.; Norouzi, M. A new heuristic model for estimating the oil formation volume factor. Petroleum 2018, 4, 300–308.

(15) Fattah, K. A.; Lashin, A. Improved oil formation volume factor (Bo) correlation for volatile oil reservoirs: An integrated non-linear regression and genetic programming approach. J. Eng. Sci. King Saud Univ. 2018, 30, 398–404.

(16) Elkhatmy, S.; Mahmoud, M. Development of new correlations for the oil formation volume factor in oil reservoirs using artificial intelligent white box technique. Petroleum 2018, 4, 178–186.

(17) Saghaﬁ, H. R.; Rostami, A.; Arabloo, M. Evolving new strategies to estimate reservoir oil formation volume factor: Smart modeling and correlation development. J. Pet. Sci. Eng. 2019, 181, 106180.

(18) Seyyedettar, M.; Ghiasi, M. M.; Zendeheboudi, S.; Butt, S. Determination of bubble point pressure and oil formation volume factor: Extra trees compared with LSSVM-CSA hybrid and ANFIS models. Fuel 2020, 269, 116834.

(19) Rashidi, S.; Mehrad, M.; Ghorbani, H.; Wood, D. A.; Mohamadian, N.; Moghadasi, J.; Davoodi, S. Determination of bubble point pressure & oil formation volume factor of crude oils applying multiple hidden layers extreme learning machine algorithms. J. Pet. Sci. Eng. 2021, 202, 108425.

(20) Tariq, Z.; Mahmoud, M.; Abduraheme, A. Machine Learning-Based Improved Pressure—Volume—Temperature Correlations for Black Oil Reservoirs. J. Energy Resour. Technol. 2021, 143, 113003.

(21) McCain, W. D., Jr Properties of petroleum fluids, 1973.

(22) Breiman, L.; Friedman, J.; Stone, C. J.; Oshlen, R. A. Classification and regression trees; CRC press, 1984.

(23) Nait Amar, M.; Shateri, M.; Hemmati-Sarapardeh, A.; Alamatzz, A. Modeling oil-brine interfacial tension at high pressure and high salinity conditions. J. Pet. Sci. Eng. 2019, 183, 106413.

(24) Hemmati-Sarapardeh, A.; Larestani, A.; Menad, N. A.; Hajirezaie, S. Applications of artificial intelligence techniques in the petroleum industry; Gulf Professional Publishing, 2020.

(25) Song, Y.-Y.; Lu, Y. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 2015, 27, 130.

(26) Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 2015, 71, 804–818.

(27) Sabah, M.; Talebeikhah, M.; Agin, F.; Talebeikhah, F.; Hasheminasab, E. Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: A case study from Marun oil field. J. Pet. Sci. Eng. 2019, 177, 236–249.

(28) Ahmad, M. W.; Reynolds, J.; Rezgui, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J. Cleaner Prod. 2018, 203, 810–821.

(29) Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

(30) Breiman, L. Bagging predictors. Mach. Learn. 1996, 24, 123–140.

(31) Basith, S.; Manavalan, B.; Shin, T. H.; Lee, G. iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput. Struct. Biotechnol. J. 2018, 16, 412–420.

(32) Rodriguez-Galiano, V. F.; Chica-Olmo, M.; Abarca-Hernandez, F.; Atkinson, P. M.; Jeganathan, C. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sens. Environ. 2012, 181, 93–107.

(33) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. Mach. Learn. 2006, 63, 3–42.

(34) Rosenblatt, F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms; Cornell Aeronautical Lab Inc Buffalo NY, 1961.

(35) Hemmati-Sarapardeh, A.; Varamesh, A.; Husein, M. M.; Karan, K. On the evaluation of the viscosity of nanoluid systems: Modeling and data assessment. Renew. Sustain. Energy Rev. 2018, 81, 313–329.

(36) Hemmati-Sarapardeh, A.; Ghazanfari, M.-H.; Ayatollahi, S.; Shamshirband, S.; Masjuki, M. Accurate determination of the CO2-crude oil minimum miscibility pressure of pure and impure CO2 streams: a robust modelling approach. Can. J. Chem. Eng. 2016, 94, 253–261.

(37) Larestani, A.; Hemmati-Sarapardeh, A.; Naseri, A. Experimental measurement and compositional modeling of bubble point pressure in crude oil systems: Soft computing approaches, correlations, and equations of state. J. Pet. Sci. Eng. 2022, 212, 110271.

(38) Fahlman, S. E.; Lebiere, C. The cascade-correlation learning architecture; CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 1990.
(39) Lashkarbolooki, M.; Vaferi, B.; Shariati, A.; Zeinolabedini Hezave, A. Investigating vapor—liquid equilibria of binary mixtures containing supercritical or near-critical carbon dioxide and a cyclic compound using cascade neural network. *Fluid Phase Equilib.* 2013, 343, 24–29.

(40) Larestani, A.; Mousavi, S. P.; Hadavimoghaddam, F.; Ostadhassan, M.; Hemmati-Sarapardeh, A. Predicting the surfactant-polymer flooding performance in chemical enhanced oil recovery: Cascade neural network and gradient boosting decision tree. *Alex. Eng. J.* 2022, 61, 7715–7731.

(41) (a) Filik, U. B.; Kurban, M. A new approach for the short-term load forecasting with autoregressive and artificial neural network models. *Int. J. Comput. Intell. Res.* 2007, 3, 973–1873. (b) Hedayat, A.; Davlu, H.; Barfrosh, A. A.; Sepanloo, K. Estimation of research reactor core parameters using cascade feed forward artificial neural networks. *Prog. Nucl. Energy* 2009, 51, 709–718.

(42) Broomhead, D. S.; Lowe, D. *Radial basis functions, multi-variable functional interpolation and adaptive networks*; Royal Signals and Radar Establishment Malvern: United Kingdom, 1988.

(43) Hemmati-Sarapardeh, A.; Larestani, A.; Nait Amar, M.; Hajirezaie, S.. *Chapter 2 - Intelligent models*; Gulf Professional Publishing, 2020.

(44) (a) Celikoglu, H. B.; Cigizoglu, H. K. Public transportation trip flow modeling with generalized regression neural networks. *Adv. Eng. Software* 2007, 38, 71–79. (b) Cigizoglu, H. K.; Alp, M. Generalized regression neural network in modelling river sediment yield. *Adv. Eng. Software* 2006, 37, 63–68. (c) Kulkarni, S. G.; Chaudhary, A. K.; Nandi, S.; Tambe, S. S.; Kulkarni, B. D. Modeling and monitoring of batch processes using principal component analysis (PCA) assisted generalized regression neural networks (GRNN). *Biochem. Eng. J.* 2004, 18, 193–210. (d) Rooki, R. Application of general regression neural network (GRNN) for indirect measuring pressure loss of Herschel–Bulkley drilling fluids in oil drilling. *Measurement* 2016, 85, 184–191.

(45) Specht, D. F. A general regression neural network.IEEE Trans. Neural Network. 1991, 2(6), 568–576. DOI: 10.1109/72.97934

(46) Firat, M.; Gungor, M. Generalized regression neural networks and feed forward neural networks for prediction of scour depth around bridge piers. *Adv. Eng. Software* 2009, 40, 731–737.

(47) Ding, L.; Rangaraju, P.; Pourasae, A. Application of generalized regression neural network method for corrosion modeling of steel embedded in soil. *Soils Found.* 2019, 59, 474–483.

(48) Asante-Okyere, S.; Xu, Q.; Mensah, R. A.; Jin, C.; Ziggha, Y. Y. Generalized regression and feed forward back propagation neural networks in modelling flammability characteristics of polymethyl methacrylate (PMMA). *Thermochim. Acta* 2018, 667, 79–92.

(49) Naghizadeh, A.; Larestani, A.; Amar, M. N.; Hemmati-Sarapardeh, A. Predicting viscosity of CO2–N2 gaseous mixtures using advanced intelligent schemes. *J. Pet. Sci. Eng.* 2022, 208, 109359.

(50) Ershadnia, R.; Amooie, M. A.; Shams, R.; Hajirezaie, S.; Liu, Y.; Jamshidi, S.; Soltanian, M. R. Non-Newtonian fluid flow dynamics in rotating annular media: Physics-based and data-driven modeling. *J. Pet. Sci. Eng.* 2020, 185, 106641.

(51) (a) Hemmati-Sarapardeh, A.; Aminshahidy, B.; Pajouhandeh, A.; Yousefi, S. H.; Hosseini-Kaldozakh, S. A. 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. (b) Hosseinzadeh, M.; Hemmati-Sarapardeh, A. Toward a predictive model for estimating viscosity of ternary mixtures containing ionic liquids. *J. Mol. Liq.* 2014, 200, 340–348.

(52) Shateri, M.; Ghorbani, S.; Hemmati-Sarapardeh, A.; Mohammadi, A. H. Application of Wilcoxon generalized radial basis function network for prediction of natural gas compressibility factor. *J. Taiwan Inst. Chem. Eng.* 2015, 50, 131–141.

(53) Rousseeuw, P. J.; Leroy, A. M. *Robust regression and outlier detection*; John wiley & sons, 2005.

(54) Atashrouz, S.; Hemmati-Sarapardeh, A.; Mirshekar, H.; Nasernajad, B.; Keshavarz Moravej, M. On the evaluation of thermal conductivity of ionic liquids: modeling and data assessment. *J. Mol. Liq.* 2016, 224, 648–656.