KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

KUMASI, GHANA

COLLEGE OF ENGINEERING

DEPARTMENT OF PETROLEUM ENGINEERING

PROJECT REPORT

TOPIC: PREDICTING OIL FORMATION VOLUME FACTOR WITH MACHINE LEARNING

PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE BSC (ENG.) DEGREE

By:

DABLAH URIEL SELIKEM

EWUSI EMMANUEL

AMOH SARPONG SAMUEL

ABELIMWINI SALOMEY

INCOOM ALBERT

EKORBOR AFIBA BADWO STEPHANIE

BAABA - YAKUUB ABDUL - MUUMIN YAKUUB YEMBONI

SUPERVISOR: DR. CASPER DANIEL ADENUTSI, PHD

Table of Contents

CHAPTER ONE

INTRODUCTION

1.1 Background of study

In the oil and gas industry, the Oil formation volume factor (Bo), is the ratio of the volume of oil (including the gas in solution) at the reservoir temperature and pressure to the volume of oil at standard conditions at surface. As oil is produced from the reservoir, shrinkage in volume occurs due to the release of dissolved gas.

It is usually greater than or equal to unity. This property plays a significant part in computing prominent petroleum engineering parameters, such as oil in place, reservoir simulation, material balance equation, well testing, and reservoir production calculation.

Although most PVT data are determined in the laboratory, petroleum engineers usually rely on empirically developed correlations. This is because of the cost of obtaining PVT analysis, and unreliable and unavailable samples.

Faster and cheaper methods are important for real-time decision making and empirically developed correlations are used in the prediction of this property. One of the key challenges in predicting oil formation volume factor is the complex nature of reservoir fluids and their behavior under various pressures, temperatures, and compositional conditions.

Traditionally, correlations and empirical models have been used to estimate oil formation volume factors based on available data and reservoir characteristics.

Despite the existence of empirical correlations that serve as traditional tools for estimating the oil formation volume factor, their limitations have prompted the exploration of advanced predictive methodologies.

This paper delves into the historical evolution of these empirical correlations, their constraints, and the transformative potential of machine learning (ML) techniques as innovative solutions for enhancing the prediction of oil.

However, with the advancements in machine learning, there is an opportunity to improve the accuracy of these predictions by utilizing algorithms that can capture complex patterns and relationships in the data. Machine learning techniques, such as regression analysis and artificial neural networks, have shown promising results in predicting oil formation volume factors. These techniques can take into account a wide range of variables, including reservoir pressure, temperature, fluid composition, and other relevant parameters.

1.2 Problem statement

Current methods for predicting Oil Formation Volume Factor (Bo), a vital property in oil production, rely on correlations derived from experiments. These correlations often have limitations. They might not capture the intricacies of real reservoirs, are restricted by specific data ranges, and don't account

for unique reservoir characteristics. Additionally, obtaining the necessary data for these correlations can be difficult, especially in mature or unconventional oil fields. Because of these shortcomings, this study explores the use of machine learning techniques to improve the accuracy and reliability of Bo predictions.

1.3 Justification

Compared to the present empirical methods, machine learning methods combine multiple models, reducing bias and variance, and leading to improved overall predictive accuracy.

The machine learning methods can be more robust to outliers and noise in data, as the combined predictions can mitigate the impact of individual model errors. The machine learning methods capture a wider range of patterns and relationships in the data, enhancing their generalization capabilities and also incorporating diverse models, leveraging different perspectives, and reducing the risk of model limitations.

1.4 Aim and Objectives

Aims

This study aims to develop a novel and robust machine learning model for predicting Oil Formation Volume Factor (Bo), offering an improvement over existing widely used conventional methods.

Objectives

To achieve this aim, the following specific objectives will be addressed:

- Identify and extract key features (independent variables) from readily available reservoir fluid properties that significantly influence Bo. Dimensionality reduction techniques may be employed to optimize model performance and interpretability.
- Develop a robust Bo prediction model using machine learning algorithms. This involves training and evaluating various algorithms to select the one that achieves the highest accuracy and generalizability for predicting Bo.
- Conduct a comprehensive evaluation of the developed machine learning model's performance using established metrics. This evaluation will compare its predictive accuracy and efficiency against existing widely used conventional methods, such as empirical correlations.
- Analyze and compare the performance of the best-performing machine learning model with established empirical correlations for Bo prediction. This analysis will highlight the strengths and limitations of each approach, providing valuable insights for reservoir fluid characterization**.**

1.5 LIMITATIONS

Machine Learning (ML) models can enhance the accuracy of Formation Volume Factor (Bo) predictions significantly, but they come with their own set of limitations. A major issue in ML is overfitting, where a model learns too much from the noise and unnecessary details in the training data, affecting its performance on new data. This is especially relevant to Bo prediction due to the diverse and complex geological and reservoir characteristics. An overfitted model might perform

well on its training data but fail to generalize to different, unseen datasets, leading to unreliable predictions when applied to various reservoirs.

Furthermore, the effectiveness of ML models heavily depends on the quantity and quality of the data used in training. For Bo prediction, obtaining a large and accurate dataset that reflects subsurface conditions accurately is challenging. Data that is incomplete or biased can result in models that do not fully capture the complexity of reservoir behaviours, thus reducing prediction accuracy.

Selecting the right input features is crucial for ML model performance. Features that are not precise or relevant can lead to poor predictions. In the context of Bo prediction, it is essential to identify the key geological and fluid properties that affect the Bo. Additionally, feature engineering, which involves creating new features from existing data, can be a complicated process.

CHAPTER TWO

LITERATURE REVIEW

2.1 EMPIRICAL CORRELATIONS FOR OIL FORMATION VOLUME FACTOR (Bo) PREDICTION.

In the field of petroleum engineering, predicting oil formation volume factor is crucial for various applications such as reservoir characterization, production optimization, and determining the recoverable reserves of an oil field (Honarpour et al., 2006).

Several empirical correlations have been developed over the years to estimate the oil formation volume factor. These correlations are based on the analysis of experimental data and empirical relationships between formation volume factor and various reservoir fluid properties (Salim Basaleh, n.d.) such as oil composition, temperature, pressure, and solution gas-oil ratio.

One of the earliest correlations for oil formation volume factor estimation was developed by Katz in 1942(Ahmed, 2018; Standing & Katz, 1942a).Katz's study was based on 16 saturated hydrocarbon samples at certain pressure and temperature ranges. The graphical correlation used reservoir temperature, pressure, solution gas oil ratio, oil and gas gravity. The correlation was published only in graphical form was difficult to use due to the combination of graphs and calculations.

Standing further refined Katz's correlation by incorporating additional variables such as oil gravity and reservoir temperature into the equation. The correlation was based on experiments carried out on samples from 105 experimental data set collected from different oil fields located in California to develop his graph (Standing, 1947). Standing introduced an empirical correlation that included oil

gravity and reservoir temperature as variables in the estimation of the oil formation volume factor. Standing's correlation, developed in 1947, became widely used and is still utilized today.

Another notable correlation was developed by Glaso in 1980 based on Standing's work (Glaso, 1980a). Glaso correlation (1980) published a reservoir fluid properties correlation for North Sea hydrocarbon mixtures. It was based on Standing's model with some modifications. An important feature of Glaso's model is that it accounts for paraffinicity effect by correcting the flash stock tank oil gravity to an equivalent corrected value with reservoir temperature and oil viscosity. The correlation also considers the presence of nonhydrocarbons on saturation pressure by using correction factors for the presence of N_2, CO_2, and H_2S in the total surface gases. He used a total of 45 oil samples, most of which came from the North Sea oil fields ranging from heavy oils and tars to volatile oils.

Additionally, Vasquez and Beggs in 1977 developed correlations based on data obtained from fields worldwide by applying the concept of regression analysis on 6000 data sets collected from fields worldwide, generally applicable for all oil types and covering a wide range of pressures, temperatures, and oil properties (Vazquez & Beggs, 1977a). In their correlation, Vazquez and Beggs added reservoir pressure and isothermal oil compressibility to the input parameters

Al-Marhoun (Al-Marhoun, 1988a) published a new correlation Bo at the bubble point pressure as a function of gas specific gravity, gas solubility, oil specific gravity, and reservoir temperature using nonlinear multiple regression analysis for 160 experimental sets of data collected from 69 Middle Eastern reservoirs. He developed another correlation to predict the total Bo below the bubble point pressure by adding the reservoir pressure to the input parameters. The Al-Marhoun correlation could not offer reliable predictions beyond the conditions for which it was established. The correlation could not

accurately predict oil formation volume factors under extreme conditions, such as very high or very low temperatures and pressures and in reservoirs with highly viscous oils or gas condensates.

In 1992, 51 data sets from the UAE crudes had been used by (Dokla & Osman, 1992a) to develop another correlation to predict BO. Again, they used the (AlMarhoun, 1988a) correlation as a base to build his correlation.

Omar & Todd (Omar & Todd, 1993a) used 93 data sets collected from various reservoirs in Malaysia to develop a PVT correlation for the Malaysian crude oil using linear and non-linear regression analysis. Their correlation is based on Standing (1947) work; thus, the same input parameters were used. The correlation developed to predict BO at the bubble point pressure is reported to have an average absolute percentage error of 1.44% and a correlation coefficient of 99%, which was a clear improvement in the accuracy

Petrosky and Farshad proposed correlations using samples from Gulf of Mexico oils in 1993(Petrosky Jr. & Farshad, 1998a). The new correlation coefficients were based on Standing's correlations for bubble point pressure, GOR, and Bo. This correlation is based on the Standing Bo correlation. To achieve the best results, nonlinear multiple regression analysis was used.

In 1997, Almehaideb (Almehaideb, 1997a) proposed a new correlation to predict the PVT properties of crude oil in the UAE. A data set comprised of 62 data samples was collected from several UAE oil fields in his study. Then, regression analysis was performed to develop correlations that can predict PVT properties of crude oil in the UAE. The input parameters used to develop Bo correlation are oil-specific gravity, gas-oil ratio, and reservoir temperature. The reported average absolute percentage error for estimating Bo at the bubble point pressure is 1.35%, and the correlation coefficient is 0.9985.

Al-Shammasi (Al-Shammasi, 2001a) presented a new correlation. This work was based on 1709 global data sets gathered from 13 published papers. His first correlation made used of four parameters: GOR, reservoir temperature, gas-specific gravity, gas solubility. It was an improvement over previously published correlation. In his second correlation, the gas-specific gravity. The reported average absolute percentage error for the first and second equations is 1.806% and 3.033%, respectively

These correlations have been continuously refined and expanded to accommodate diverse reservoir characteristics and fluid properties, contributing to advancements in petroleum engineering practices. However, it is essential to acknowledge the ongoing development of correlations and the potential of emerging technologies, such as machine learning, to further improve predictions of oil formation volume factor.

While empirical correlations have been widely used in the prediction of oil formation volume factor, they have some limitations.

One limitation is that these correlations are based on regional data and may not accurately represent the characteristics of different oil reservoirs. For example, the correlations developed by Katz and Standing were based on specific regions such as California, and may not apply to oil reservoirs in other areas. Likewise, Glaso's correlation was developed for North Sea crudes and may not be suitable for reservoirs in other regions.

Another limitation of empirical correlations is their reliance on simplified assumptions and linear relationships between variables. These assumptions may not hold for all oil reservoirs, particularly those with complex fluid systems or non-linear relationships between variables. In addition, empirical

correlations are based on a limited range of experimental data and may not capture the full complexity of all oil reservoir behaviors.

Table 1:Published empirical correlations used to predict oil formation volume factor (BO)

Author	Origin	Dataset	Applicable Range	Limitations

Standing & Katz (1942)	California	105	Bo: 1.0240 - 2.150 (bbl/stb) T: 100 - 258 (F˚) Rs:201-425(scf/stb) API: 16.5 - 63.8 γg: 0.9 - 0.955	- Ignores non-hydrocarBon components
Vazquez & Beggs (1977b)	Worldwide	5008	Bo: 1.028 - 2.226 (bbl/stb) T: 75 - 294 (F˚) Rs: 0 - 2199 (scf/stb) API: 15.3 - 59.3 γg : 0.65 - 1.28 Pb: 15 - 6055 (psia)	- May not represent certain oil types well
Glaso (1980b)	North Sea	45	Bo: 1.032 - 2.588 (bbl/stb) T: 80 - 280 (F˚) Rs: 90 - 2637 (scf/stb) API: 22.3 - 48.1 γg : 0.65 - 1.276 Pb:165 - 7142 (psia)	- Limited applicability, struggles with generalization

Al-Marhoun (1988)	Middle East	160	Bo: 1.032 - 1.997 (bbl/stb) T: 74 - 240 (F˚) Rs: 26 - 1602 (scf/stb) API: 19.4 - 44.6 γg:0.75 - 1.673 Pb:130 - 357 (psia)	- Inaccurate for oils unlike Middle Eastern crudes
Kartoatmodjo & Schmidt (1994)	Worldwide	5392	Bo: 1.007 - 2.747 (bbl/stb) T:75 - 320 (F˚) Rs: 0 - 2890 (scf/stb) API: 14.4 - 59 γg: 0.4824 - 1.1.668 Pb:24.7 - 4746.7(psia)	- Potential bias towards specific regions/oil types
Dokla & Osman (1992b)	UAE	51	Bo: 1.216 - 2.493 (bbl/stb) T: 190 - 275 (F˚) Rs: 181 - 2266 (scf/stb) API:28.2 - 40.3 γg: 0.798 - 1.29 Pb: 590 - 4640 (psia)	- Limited capture of crude oil property diversity (even within UAE)

Macary & El-Batanoney (1993)	Gulf of Suez	90	Bo: 1.2 - 2 (bbl/stb) T: 130 - 290 (F˚) Rs: 200 - 1200 (scf/stb) API: 25 - 40 γg: 0.7 - 1 Pb: 1200 - 4600 (psia)	- Application to other regions with significantly different compositions could result in less accurate predictions of
Petrosky Jr. & Farshad (1998b)	Gulf of Mexico (Texas, Louisiana)	81	Bo:1.118 - 1.623 (bbl/stb) T: 114 - 288 (F˚) Rs: 217 - 1406 (scf/stb) API: 16.3 - 45.0 γg: 0.58 - 0.85 Pb: 1574 - 6528(psia)	- Significant variations within Gulf of Mexico crudes and other locations might not be fully captured.
Omar & Todd (1993b)	Malaysia	93	Bo:1.085 - 1.954 (bbl/stb) T: 125 - 280 (F˚) Rs: 142 - 1440 (scf/stb) API: 26.6 - 53.2 γg: 0.612 - 1.32 Pb: 790 - 3851 (psia)	- Limited applicability, might not be accurate for crudes with significantly different compositions.

Hanafy et al. (1997)	Egypt	324	Bo: 1.032 - 4.35 (bbl/stb) T: 107 - 327 (F˚) Rs: 7 - 4272 (scf/stb) API: 17.8 - 48.8 γg: 0.623 - 1.627 Pb: 36 - 5003 (psia)	- It might not fully capture the diversity of crude oil systems globally.
Almehaideb (1997b)	UAE	62	Bo: 1.142 - 3.562 (bbl/stb) T: 190 - 306 (F˚) API: 30.9 - 48.6 Rs: 128 - 3871 (scf/stb) γg: 0.746 - 1.116 Pb: 501 - 4822 (psia)	- Restricted applicability to a wider variety of UAE crudes with diversity in oil compositions.
Hemmati & Kharrat (2007)	Iran	287	Bo: 1.091 - 2.54 (bbl/stb) T: 77.5 - 290 (F˚) Rs: 125 - 2189.25 (scf/stb) API: 18.8 - 48.34 γg : 0.523 - 1.415 Pb: 348 - 5156 (psia)	- It might be more accurate for Iranian oil than others because of compositional similarities.

Al-Shammasi (2001)	Worldwide	1243	Bo: 1.02 - 2.916(bbl/stb) T: 74 - 341.6 (F˚) Rs: 6 - 32.98.6 (scf/stb) API: 6 - 63.7 γg: 0.51 - 3.44 Pb: 31.70 - 7127 (psia)	Not all encompassing, for instance it does not consider the exact compositions of the oil.
Elsharkawy & Alikhan (1997)	Middle East	254	Bo: 1.057 - 1.770 (bbl/stb) T:130 - 250 (°F) Rs: 34 - 1400 (scf/stb) API: 20 - 45 γg: 0.663 - 1.064 Pb: 317 - 4375 (psia)	When employed for oils from other geographical areas with significantly different geochemical characteristics, the predictions might hold less precision

2.2 MACHINE LEARNING (ML) MODELS BEING USED FOR OIL FORMATION VOLUME FACTOR (Bo) PREDICTION.

To overcome these limitations, researchers have turned to machine learning techniques for predicting oil formation volume factor. Machine learning techniques offer the potential to overcome the limitations of empirical correlations in predicting oil formation volume factor. These machine learning models can capture non-linear relationships and handle complex data sets, allowing for more accurate predictions of oil formation volume factor.

Moreover, machine learning models can be trained on a larger and more diverse dataset, encompassing various reservoir characteristics and fluid properties.
This allows for a more comprehensive understanding of the factors influencing oil formation volume factor and improves the accuracy of predictions.
Overall, while empirical correlations have been widely used in the prediction of oil formation volume factor, they have their limitations.

In recent years, the use of machine learning techniques for predicting formation volume factor in the oil and gas industry has gained significant attention. Various studies have explored the application of machine learning models such as the transparent open Box (TOB), artificial neural networks, support vector machines, genetic programming and random forest in predicting the formation volume factor of oil.(Rashidi et al., 2021; Saghafi et al., 2019; Wood & Choubineh, 2019)

(Gouda & Attia, 2024). conducted a comprehensive study on the use of artificial neural networks for predicting formation volume factor based on a combination of input variables such as reservoir pressure, temperature, oil gravity, and gas-oil ratio. Their research demonstrated the effectiveness of ANNs in capturing complex non-linear relationships between reservoir conditions and oil volume, leading to accurate predictions.

Similarly, (Rashidi et al., 2021) investigated the application of support vector machines for formation volume factor prediction. Their study showcased the capability of SVM in handling high-dimensional data and identifying patterns that contribute to accurate estimation of oil volume factors(El-Sebakhy, 2009).

(Karimnezhad et al., 2014) in his paper, proposed a new correlation to predict the Bo for Middle East crudes. Genetic Algorithm (GA) was used as the dominant tool for development of this correlation. A total of 429 data sets of different crude oils from Middle East reservoirs were used. Among those, 286 data sets as training data and 143 data sets as test data were randomly selected for constructing the correlation and for correlation validation, respectively. These results show a very good agreement with experimental data and are more accurate for Middle East crudes than those of all existing empirical correlations at the time.

These studies collectively underscore the potential of machine learning approaches in predicting formation volume factor, offering promising avenues for improving the efficiency and accuracy of oil extraction processes.

Table 2:Bo published prediction studies that have used machine-learning methods.

Author	**Formula Displayed**	**Date**	**Origin**	**Data No.**	**Method**	**Errors**
(Gharbi et al., 1999)	No	1999	World Wide	5200	ANN	R^2 =0.9875 RMSC = - SD = 2.43 AAPD%=1.97 APD% = 1.53
(Boukadi et al., 1999)	No	1999	Oman	92	ANN	R^2 =0.95 RMSC = - SD = - AAPD%=1.6

						APD% = 0.02
(Al-Marhoun & Osman, 2002)	No	2002	Saudi	282	ANN	R^2 =- RMSC = - SD = 0.6626 AAPD%=0.5116 APD% = 0.2173
(Goda et al., 2003)	No	2003	Middle East	160	ANN	R^2 =0.9981 RMSC = - SD = - AAPD%=0.03070 APD% = -0.0078
(Malallah et al., 2006)	No	2006	California	105	ACE	R^2 =0.9854 RMSC = - SD = 0.04 AAPD%=1.94 APD% = 0.002
(E. El-Sebakhy et al., 2007)	No	2007	World Wide (782 dataset)	Dataset1	ANN & SVM	R^2 =0.9769 RMSC = 1.4625 SD = - AAPD%=1.3718 APD% = 0.1808
				Dataset2		R^2 =- RMSC = 0.62

						SD = 0.4743 AAPD%=0.3527 APD% = -0.006
				Dataset3		R^2 =0.9992 RMSC = 0.65 SD = - AAPD%=0.3856 APD% = 0.0651
(E. A. El-Sebakhy, 2009)	No	2009	World Wide	1246	ANN & SVM	R^2 =0.997 RMSC = - SD = 0.4743 AAPD%=0.353 APD% = -0.006
(Dutta & Gupta, 2010)	No	2010	India	1852	ANN	R^2 =0.9890 RMSC = - SD = 1.848 AAPD%=1.779 APD% = -
(Moghadam et al., 2011)	No	2011	Iran	218	ANN	R^2 =0.997 RMSC = - SD = 0.65

						AAPD%=0.53 APD% = -3×10^{-5}
(Khoukhi, 2012)	No	2012	World Wide (1225 dataset)	Dataset1 Dataset2	ANN, GA-ANN and GA-ANFIS	R^2 =0.9998 RMSC = 0.0111 SD = - AAPD%= - APD% = - R^2 =0.9979 RMSC = 43.42 SD = 6.523 AAPD%=4.241 APD% = -0.103
(Rafiee-Taghanaki et al., 2013)	No	2013	World Wide	569	ANN & LSSVM	R^2 =0.95 RMSC = 0.64 SD = - AAPD%=1.45 APD% = -0.07
(Shokrollahi et al., 2015)	No	2015	World Wide	756	MLP, RBF, LSSVM and CMIS	R^2 =0.9760 RMSC = 0.294 SD = 0.0295 AAPD%=1.4611 APD% = -
(Salehinia et al.,	No	2016	Iran	755	NARX-	R^2 =0.999

2016)					HW ANFIS- GP ANFIS- FCM	RMSC = - SD = - AAPD%=0.0137 APD% = $7.1\times`10^{-3}$
(Seyyedattar et al., 2020)	No	2020	World Wide	569	LSSVM- CSA & ANFIS	R^2 =0.9994 RMSC = - SD = - AAPD%=0.099 APD% = 0.012

CHAPTER THREE

METHODOLOGY

Accurate prediction of Oil Formation Volume Factor (OFVF) as an essential fluid property is crucial for optimizing reservoir management and production. Traditional methods, relying on empirical correlations and laboratory measurements, often fall short in terms of efficiency (Standing, 1947). To address these limitations, this study explores the potential of machine learning algorithms to develop OFVF prediction models.

By comparing Support Vector Machines (SVM), Random Forest, Decision Trees, and Artificial Neural Networks (ANN), we aim to identify the most effective model for predicting OFVF.

INTRODUCTION TO THE WORKFLOW

The workflow (**Fig. 1**) for the proposed methods is illustrated in this section. To avoid data range biases in the calculations, the variable values of all data records are standardized to have a mean of 0 and a standard deviation of 1 for their entire dataset distributions, as shown in Equ.(1).

$$Xstd = \frac{X - \mu}{\sigma} \qquad Equ.(1)$$

Where :

- Xstd is the standardized value of X,
- X is the original value of the attribute,
- μ is the mean of the attribute,
- σ is the standard deviation of the attribute.

To achieve correct and reliable results, a k-fold cross-validation technique is applied(**Fig. 2**). Specifically, the data is split into features and target variables with the chosen features. This involves the randomly arranged data records being divided into k categories. The $k-1$ batches are used as the training subset to build and tune the model, and the remaining batch of data records is used to validate the trained models. This process is repeated k number of times, such that all sections of the dataset are used for validation data once. Finally, the model that performs the best in the validation phase is selected as the preferred model. A k value of 10 was found to be appropriate for this study.

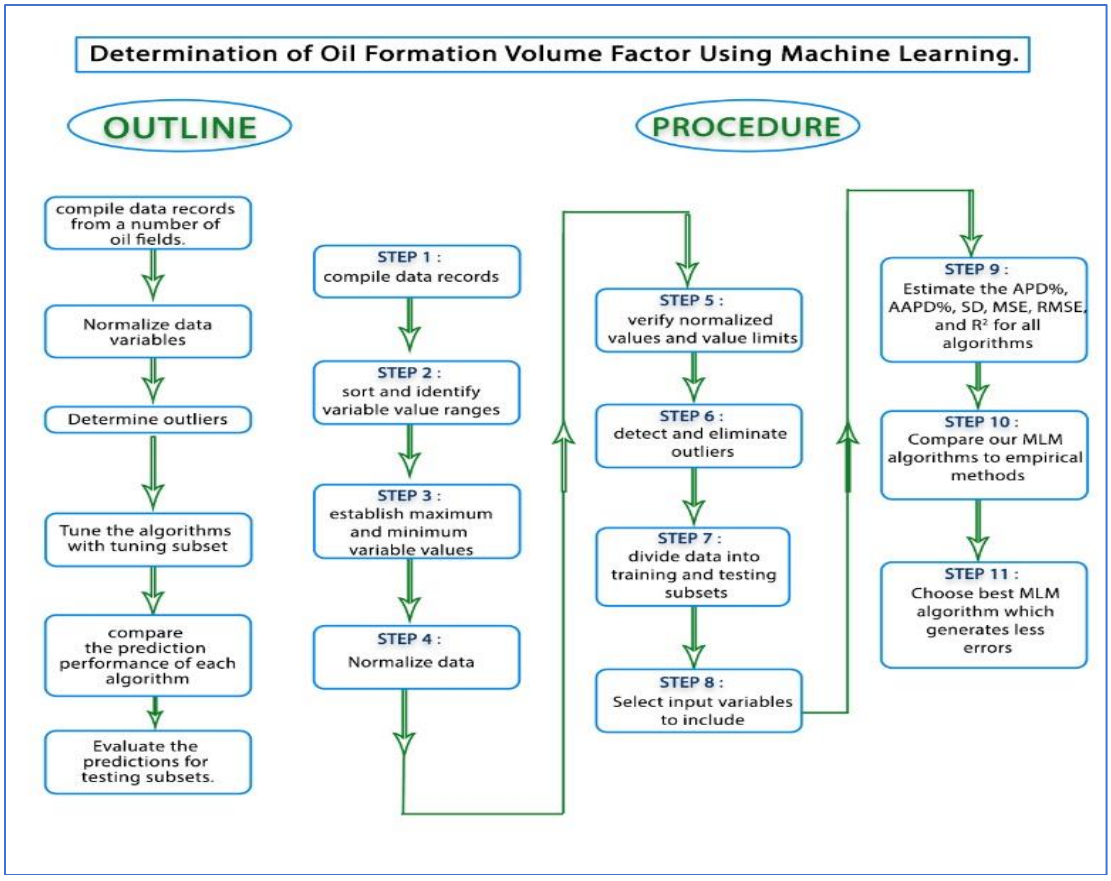


Fig. 1. Schematic diagram of the workflow sequence applied to compare ML models and empirical correlations.

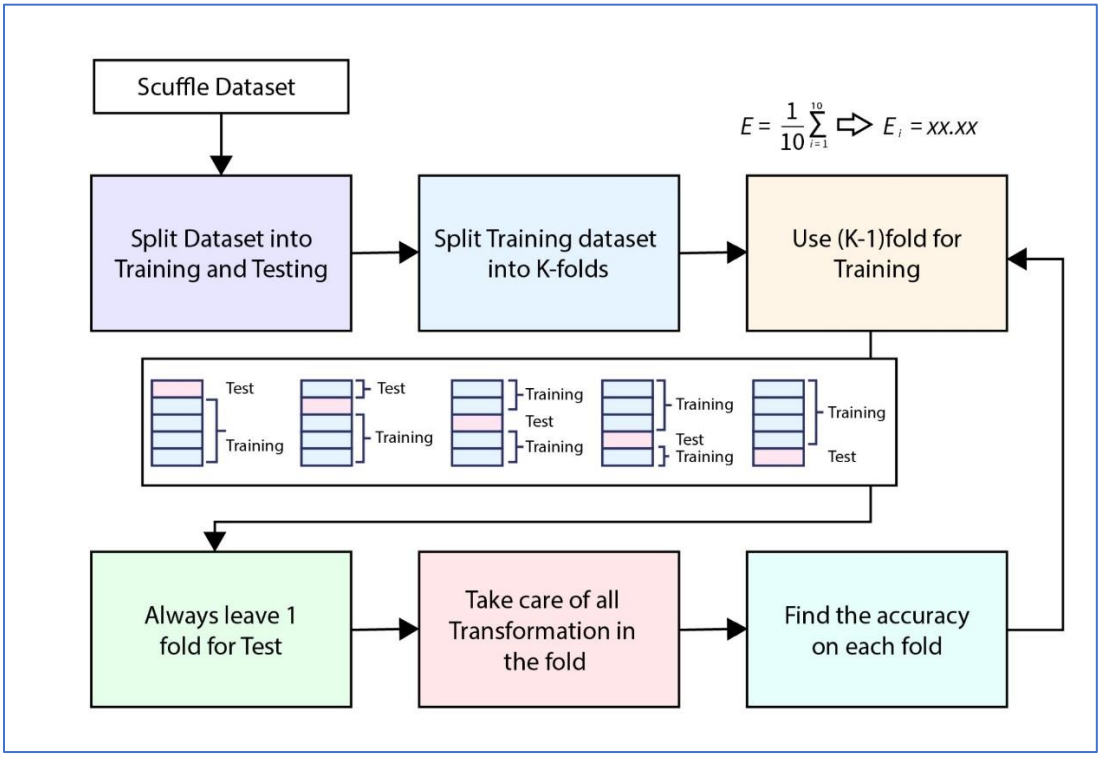


Fig. 2. Diagrammatic illustration of the K-fold cross validation technique applied.

ARCHITECTURE OF MODELS USED.

Accurate prediction of Oil Formation Volume Factor (OFVF) as an essential fluid property is crucial for optimizing reservoir management and production. Traditional methods, relying on empirical correlations and laboratory measurements, often fall short in terms of accuracy and efficiency (Standing, 1947). To address these limitations, this study explores the potential of machine learning algorithms to develop robust OFVF prediction models.

By comparing Support Vector Regression (SVR), Random Forest, Gradient Boosting, Decision Trees, and Artificial Neural Networks (ANN), we aim to identify the most effective model for predicting OFVF.

Gradient Boosting

Gradient Boosting is a powerful machine learning algorithm that has gained significant attention in recent years due to its remarkable performance across a wide range of applications, from predictive modelling to decision-making tasks. This algorithm is particularly adept at handling heterogeneous features, noisy data, and complex dependencies, making it a popular choice for challenges such as web search, recommendation systems, and weather forecasting.

At its core, the Gradient Boosting algorithm is a type of ensemble learning method that combines the strengths of multiple weak models to create a strong predictive model. The algorithm works by iteratively training a series of weak models, each of which focuses on correcting the errors made by the previous model. This process of sequential model building, known as boosting, allows the algorithm to gradually improve its performance and achieve state-of-the-art results.

In any machine learning algorithm, errors play a crucial role. There are mainly two types of errors: bias error and variance error. The gradient boosting algorithm helps us minimize the bias error of the model. The main idea behind this algorithm is to build models sequentially, where each new model attempts to correct the errors of the previous one.

In regression problems, where our target variable is continuous, we use the Gradient Boosting Regressor. The objective is to minimize the loss function, such as Mean Squared Error (MSE), by adding weak learners using gradient descent. This process involves:

1. Initial Model: Start with a simple model to make initial predictions.

2. Calculate Residuals: Compute the errors (residuals) between the predicted values and the actual values.

3. Fit New Model: Build a new model to predict these residuals.

4. Update Predictions: Add the predictions from the new model to the previous predictions to get updated predictions.

5. Repeat: Continue this process for a specified number of iterations or until the residuals are minimized.

By iteratively focusing on the errors of the previous models, Gradient Boosting Regressor effectively reduces the overall error, leading to better predictive performance.

This explanation highlights the key concepts of gradient boosting in the context of regression, focusing on minimizing errors and sequentially improving the model's performance.

One of the key advantages of the Gradient Boosting algorithm is its flexibility and customizability. The algorithm can be tailored to specific applications by adjusting the loss function, which determines the optimization objective. This allows the algorithm to be optimized for different types of problems, such as classification, regression, or ranking tasks.

The mathematical framework of the Gradient Boosting algorithm can be broken down into several core components.

Support Vector Regression (SVR)

Support Vector Regression (SVR), a variant of Support Vector Machines (SVMs) specifically designed for regression problems, offers a powerful machine-learning approach for predicting OFVF. This allows reservoir engineers to optimize oil production and reservoir management strategies.

Unlike SVMs used for classification, which identify decision boundaries to separate data points into distinct classes, SVR focuses on finding a function that best fits the training data for continuous variables like OFVF. This function aims to minimize the prediction error between the actual and estimated OFVF values (Vapnik, 1995).

A key strength of SVR for OFVF prediction lies in its ability to handle high-dimensional data. Reservoir properties influencing OFVF can be numerous, leading to datasets with many input variables like pressure, temperature, gas specific gravity, oil specific gravity, gas solubility and API gravity. SVR excels at handling these complex scenarios, similar to SVMs (Vapnik, 1995).

Another significant advantage of SVR is its capability to model non-linear relationships. The relationship between various reservoir properties and OFVF can be intricate and non-linear. Similar to SVMs, SVR leverages kernel functions to map the data into a higher-dimensional feature space. This allows the model to capture these non-linear relationships, potentially leading to more accurate OFVF predictions compared to algorithms limited to linear models (Schölkopf & Smola, 2002).

Furthermore, SVR can perform well even with limited training data, a common challenge when collecting large datasets for specific oil reservoirs. Additionally, SVR shares the memory efficiency of SVMs, which becomes beneficial when dealing with vast amounts of data from multiple wells or formations (Vapnik, 1995).

Beyond these core advantages, SVR offers additional benefits for OFVF prediction. Like SVMs, SVR exhibits robustness to outliers, which can be present in reservoir data due to measurement errors or specific reservoir characteristics (Cortes & Vapnik, 1995). While not as interpretable as simpler models, SVR can still provide some insights into the relationship between reservoir properties and OFVF through analysis of the support vectors (Guyon & Vapnik, 2002). The figure below illustrates the flowchart of the SVR model used in predicting OFVF.

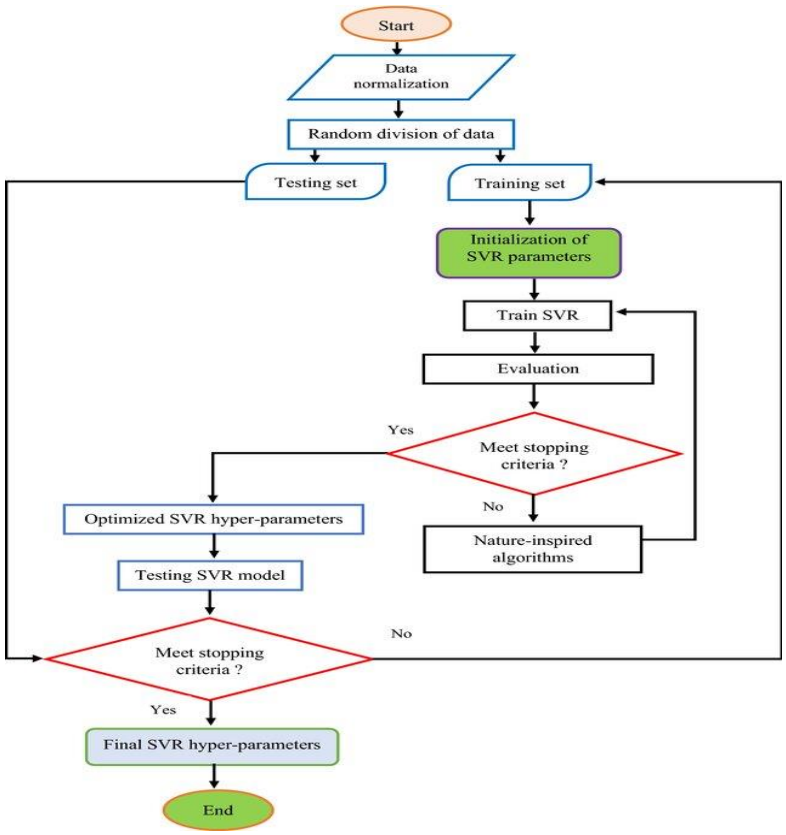


Fig. 3 Flowchart of SVR Model Used.

Random Forest

Random Forest, a robust ensemble learning technique, emerges as a promising tool for predicting Oil Formation Volume Factor (OFVF). This methodology constructs multiple decision trees and aggregates their predictions to enhance predictive accuracy (**Fig. 4**) while mitigating overfitting (Breiman, 2001). Each tree is independently trained on a randomly selected subset of the data (bootstrap sampling) and a subset of features (feature bagging) (Hastie, Tibshirani, & Friedman, 2009). The final prediction is derived by averaging the outputs of all trees in the forest.

The strength of Random Forest lies in its ability to handle complex datasets characterized by high dimensionality, non-linear relationships, and missing values (Liaw and Wiener, 2002). By combining the predictions of numerous diverse trees, the algorithm effectively reduces the risk of overfitting, a common challenge in predictive modelling. Moreover, Random Forest offers a balance between bias and variance, resulting in models that generalize well to unseen data.

In the context of OFVF prediction, Random Forest can effectively address the challenges posed by the heterogeneity and complexity of reservoir data. By incorporating multiple factors such as pressure, temperature, fluid composition, and reservoir properties, the algorithm can build predictive models that offer enhanced accuracy and reliability compared to traditional methods.

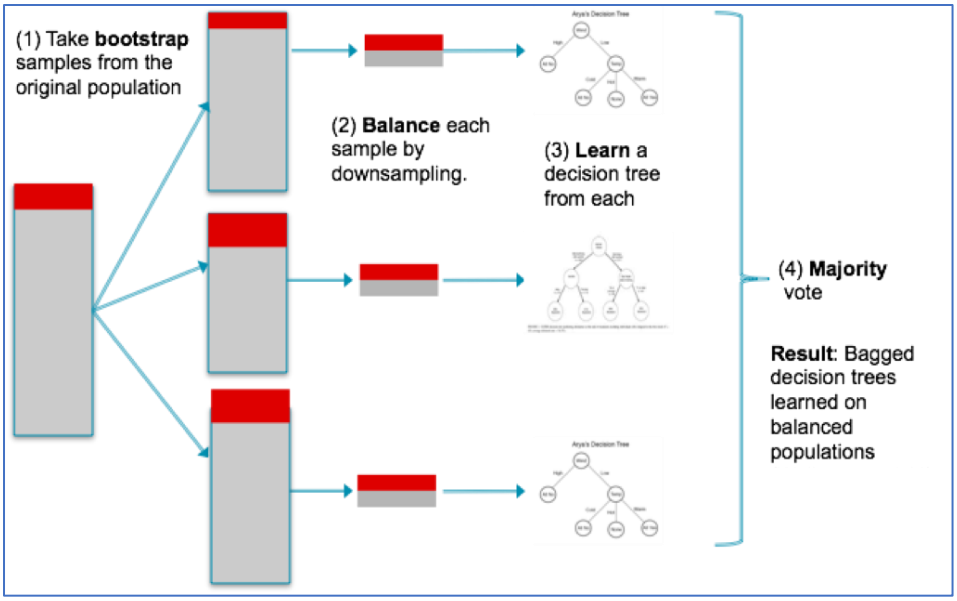


Fig. 4 Flowchart of SVR Model Used.

Artificial Neural Networks (ANN)

ANNs learn through a process called backpropagation, an algorithm that adjusts the weights of connections between neurons (**Fig. 5**) to minimize prediction errors (Rumelhart, Hinton, & Williams, 1986). This iterative process allows the network to gradually improve its predictive accuracy. In the context of OFVF, ANNs can learn to identify intricate relationships between various input parameters

(pressure, temperature, gas specific gravity, oil specific gravity, gas solubility and API gravity) and the corresponding OFVF values.

Network Architecture

The architecture of an ANN is a critical determinant of its performance. It involves specifying the number of hidden layers, neurons per layer, and the type of activation functions. For OFVF prediction, multi-layer perceptrons (MLPs) are commonly employed. These networks consist of an input layer, one or more hidden layers, and an output layer(**Fig. 6**). The hidden layers extract relevant features from the input data, while the output layer produces the predicted OFVF values.

Activation Functions

Activation functions introduce nonlinearity into ANNs, enabling them to learn complex patterns. Common activation functions include Rectified Linear Units (ReLU), sigmoid, and tanh. The choice of activation function can significantly impact the network's performance. ReLU has gained popularity due to its computational efficiency and ability to mitigate the vanishing gradient problem. However, other activation functions may be more suitable depending on the specific characteristics of the OFVF data.

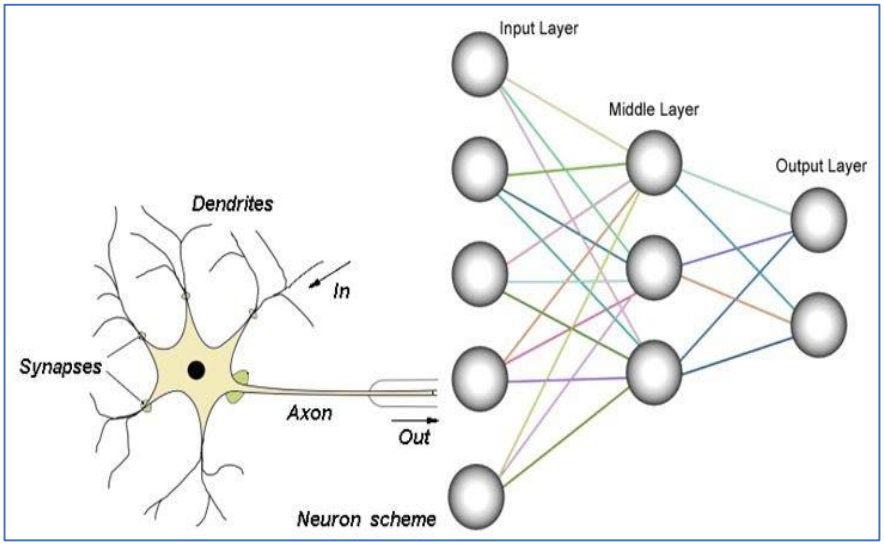


Fig. 5. Major Structure of Biologic Nerve Cell (Neuron)

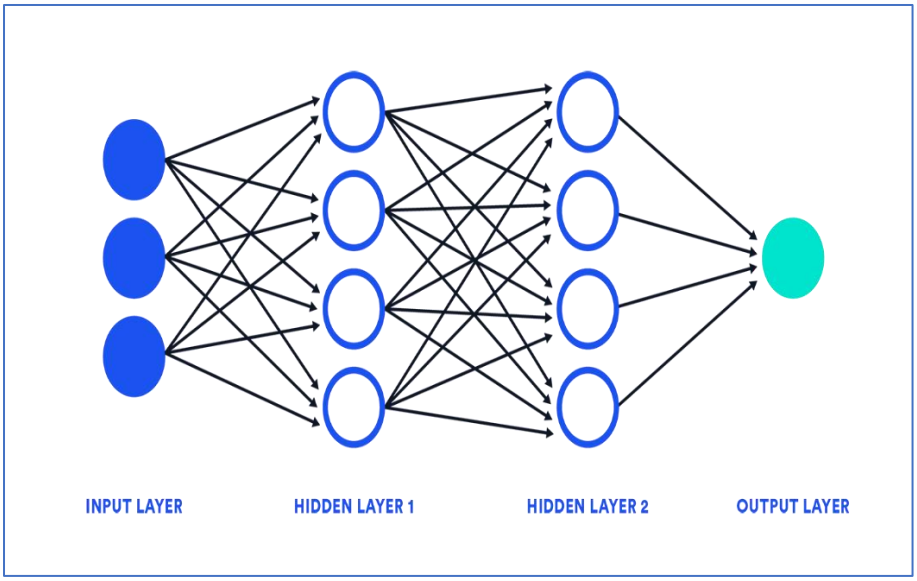


Fig. 6. Fully Connected Network with Two Hidden Layers and Output Layer

Decision Tree

Decision trees employ a hierarchical structure to partition data based on feature attributes. The tree consists of nodes, where each internal node represents a test on an attribute, and branches represent the possible outcomes of the test. At the tree's terminals, leaf nodes signify class labels or predicted values. The tree is constructed through a recursive partitioning process, where an algorithm selects the optimal feature to split the data, aiming to maximize information gain or minimize impurity (Quinlan, 1986). The classification or prediction of a new data instance involves traversing the tree from the root node, following the branches corresponding to the instance's attribute values until reaching a leaf node, which determines the assigned class or predicted value (Stone, 1984).

Data Collection and Processing of Dataset

The dataset provided for this research paper consists of several key petrophysical variables, including the reservoir, temperature (T), gas gravity (γg), oil gravity (γo), reservoir pressure (P), API gravity, and oil formation volume factor. .

After establishing a comprehensive dataset, the next step was to perform detailed data exploration and preprocessing. This included analyzing the statistical properties of each variable, such as the mean, standard deviation, minimum, and maximum values, as well as the 25th, 50th, and 75th percentiles. The data cleaning process involved identifying and addressing any outliers or erroneous data points to ensure the reliability of the subsequent analysis.

For this study, 221 input data related to. These data are used to evaluate novel machine learning models to determine OIL FORMATION VOLUME FACTOR (BO) based on readily available input variables.

Table1 provides statistical characterizations of the dataset as a whole, training subset and the related testing subset, respectively. Details for the variables are shown in the t in Table 1. Note that oil specific gravity and API gravity are related formulaically and are not independent of each other. In the models evaluated API is used as an input variable, oil specific gravity is not, as it makes sense to use only one of these closely related variables. The training and testing subsets are selected randomly but spread across almost the entire range of variable values

	Count	*Mean*	*Standard Deviation*	*Min*	*25%*	*50%*	*70%*	*Max*
Rs	221.0	563.954751	365.420925	24.000	330.000	468.0000	694.0000	2496.0000
T	221.0	225.307692	36.062136	125.000	198.000	232.0000	248.0000	296.0000
γg	221.0	1.366964	0.564725	0.612	0.811	1.3539	1.7167	3.4445
γo	221.0	0.829068	0.024172	0.766	0.815	0.8280	0.8370	0.8950
P	221.0	1452.687783	916.220169	55.000	715.000	1370.0000	2058.0000	4975.0000
API	221.0	39.274661	4.950105	26.600	37.500	39.4000	42.0000	53.2000
Bo	221.0	1.450834	0.252608	1.085	1.297	1.3990	1.5200	2.7130

Feature Selection using Spearman and Pearson Correlation

In this report paper, we investigate the use of Spearman and Pearson correlation methods for feature selection in predicting the oil formation volume factor from various reservoir properties, including Rs, T, γg, γo,P, API, and Bo.

Traditional regression analysis or correlation tests often assume a linear relationship between input features and the target property However, the relationship between field measurements and most reservoir properties are considered nonlinear (The feature selection technique aims to remove redundant or irrelevant features without much loss of information, making the model easier to interpret and improving generalization by reducing variance).

To this end, we employed Spearman and Pearson correlation analyses to identify the most relevant features for predicting the oil formation volume factor. Spearman correlation is a nonparametric measure of rank correlation, which is more suitable for exploring nonlinear relationships (Zhao et al., 2023). Pearson correlation, on the other hand, is a measure of the linear relationship between two variables (Zhao et al., 2023).

The results of the feature selection analysis using Spearman and Pearson correlation are presented in the following sections.

Influence Analysis of Independent Variables

It is well established from historical empirical relationships and various applications) that Oil Formation Volume Factor (Bo) values of crude oils are primarily a function of the following four parameters:

- Solution Gas Oil Ratio (Rs)
- Gas Specific Gravity (γg)
- API Gravity (API, or oil specific gravity, γo, from which API is calculated)
- Temperature (T)

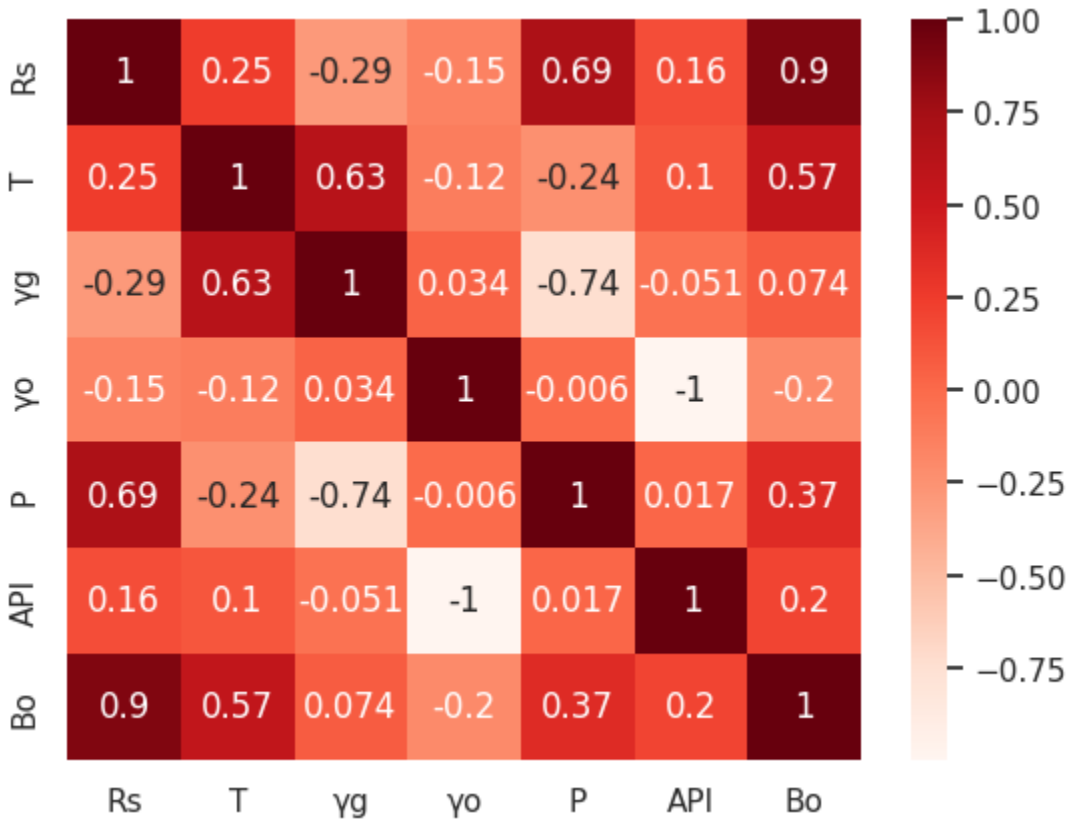


Figure 1 Pearson

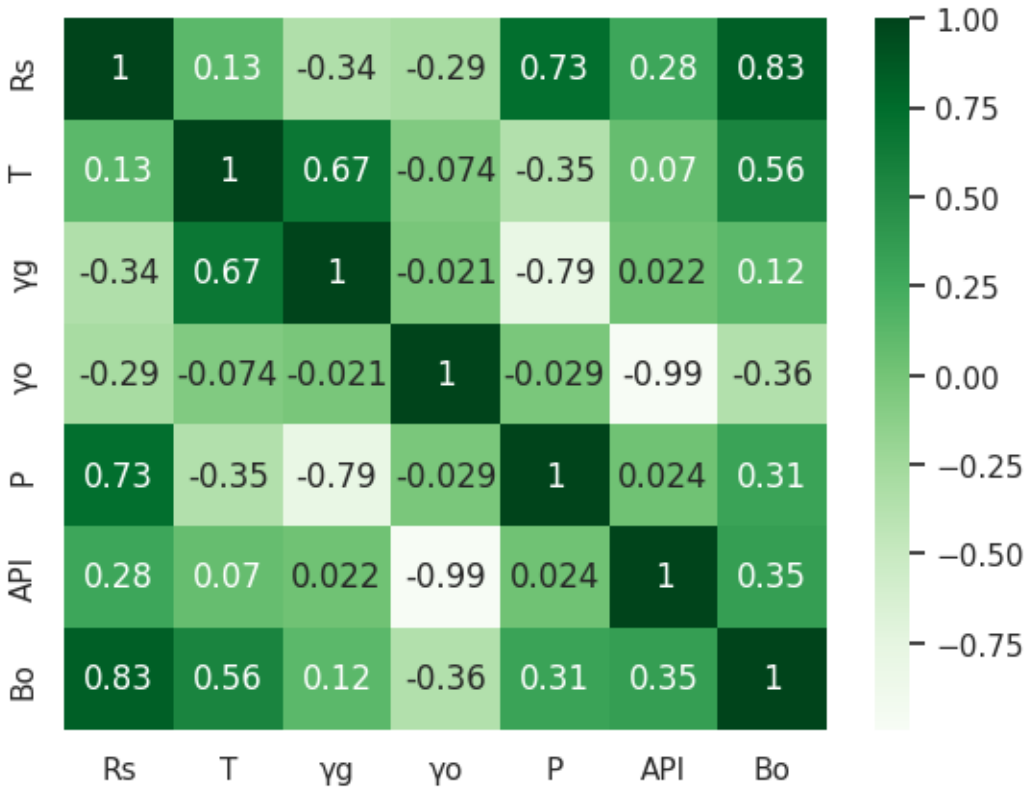


Figure 2 Spearman

It is informative to establish, in a relative sense, how influential these variables are in determining the Oil Formation Volume Factor (Bo). The Pearson correlation coefficient and the coefficient of determination can be used to measure the strength of the linear relationship between normally distributed variables. However, it is not realistic to assume that the influencing variables involved in Equation (26) are linearly related or normally distributed. Therefore, it is more meaningful to use the Spearman rank correlation method or other non-parametric statistical tests to evaluate the non-linear relationships influencing the Oil Formation Volume Factor (Bo) dependent variables (Myers and Sirois, 2004).

As with the Pearson's correlation coefficient, the non-parametric Spearman's correlation coefficient is expressed over the range of -1 (perfect negative correlation) to 1 (perfect positive correlation), with a zero value indicating a total lack of correlation.

Figures 1 and 2 display the p-values for Oil Formation Volume Factor (Bo) with each of the key influencing input variables identified. These results reveal the following:

- Solution Gas Oil Ratio (Rs) : This is the most influential variable with respect to Oil Formation Volume Factor (Bo), as it displays strong positive correlations.
- Temperature (T) : The second most influential variable, showing a strong positive correlation with Bo. API Gravity (API) : Shows a negative correlation with Bo. This is explained by the formulaic relationship between API and oil specific gravity (γo) expressed by Equation (28):

- API = $\frac{141.5}{y_0} - 131.5$
- Oil Specific Gravity (γo): Shows positive correlations with Bo and has a similar magnitude of correlation as API, but in the opposite direction due to their formulaic relationship.
- Gas Specific Gravity (γg : This is the least influential variable with a Spearman's correlation coefficient close to zero.

In summary, the order of influence of the variables on Oil Formation Volume Factor (Bo) for the compiled dataset of 221 records is as follows:

Solution Gas Oil Ratio (Rs)} > Temperature (T)} > API or Oil Specific Gravity; >Gas Specific Gravity (γg)

Figures 1 and 2 illustrate these relationships and highlight the relative influence of each variable on the Oil Formation Volume Factor (Bo).

Results and discussion

In this chapter, we present the results of our machine learning models developed for predicting the Oil Formation Volume Factor (Bo) and discuss the implications of our findings. By leveraging various machine learning algorithms, we were able to capture complex, non-linear relationships between reservoir properties and Bo. We then compare the performance of our machine learning models with each other as well as to traditional empirical correlations and highlight the improvements and potential limitations observed.

4.1. Accuracy metrics used to assess model's prediction performance

Five commonly used statistical measures of prediction accuracy are computed in this study to assess in detail each model's performance in predicting OIL FORMATION VOLUME FACTOR (BO) for the compiled dataset. These measures are: average percentage deviation (APD), average absolute percentage deviation (AAPD%), mean square error (MSE), root mean square error (RMSE), and coefficient of determination (R^2).

$$\text{APD} = \frac{\sum_{i=1}^{n} PD_i}{n}$$

$$\text{AAPD} = \frac{\sum_{i=1}^{n} |PD_i|}{n}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Z_{measured\ i} - Z_{predicted\ i})^2$$

A crucial metric for evaluating the predictive performance of a machine learning model. It measures the average squared difference between the predicted and the actual target values within a dataset. The primary objective of the MSE is to assess the quality of a model's predictions by measuring how closely they align with the ground truth. Good models have values closer to zero

RMSE = $\sqrt{MSE}$

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{N}(Z_{predicted\,i} - Z_{measured\,i})^2}{\sum_{i=1}^{N}(Z_{predicted\,i} - \frac{\sum_{i=1}^{n} Z_{measured\,i}}{n})^2},$$

Also known as the coefficient of determination, is a statistical measure that represents the goodness of fit of a regression model. R^2 ranges from 0 to 1, with 1 indicating a perfect fit and 0 a poor fit. R^2 is not exhaustive for accounting for overfitting. Overfitting occurs when the model performs well on the training set but poorly on the evaluation set. Hence, other metrics were used to evaluate the performances of the models

The focus on OFVF predictions for the overall dataset (221 data records) reveals that the four-hybrid machine-learning-optimization models provide high OFVF prediction accuracy. The four machine-learning-optimizer models substantially outperform the empirical models in their predictions of OFVF for this dataset, in terms of all five prediction accuracy metrics. The ANN model provides the best prediction performance for OFVF in terms of RMSE and R^2. Among the empirical models, the Al-Marhoun pioneering model provides the best prediction performance for OFVF in terms of RMSE and R^2 (Fig. 12). However, the ANN model stands out as providing the most accurate OFVF predictions. The same is true for the other prediction performance metrics.

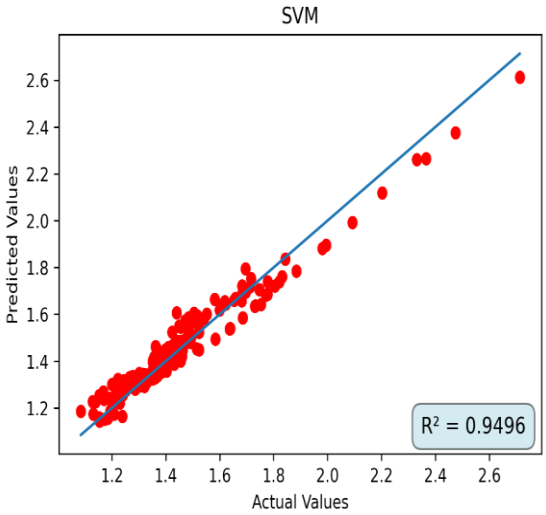
Table 16. Prediction performance for OIL FORMATION VOLUME FACTOR (BO) compared for empirical models and the four machine-learning

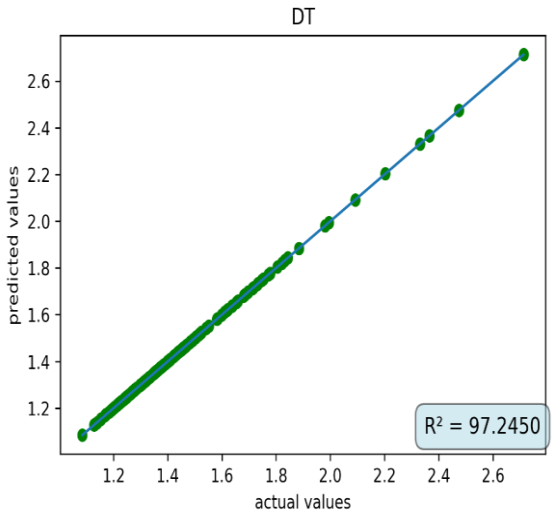
Prediction Performance for OIL FORMATION VOLUME FACTOR (BO) (bbl/STB) for Models Evaluated Applied to the Testing Subset data records)

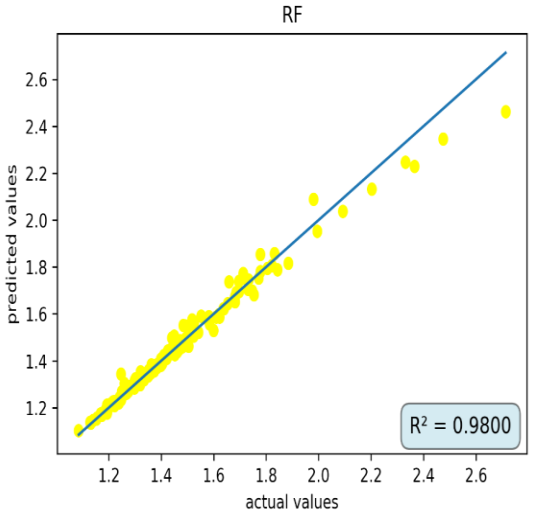
Models	APD%	AAPD%	SD	MSE	RMSE	R^2
Empirical Models						
Standing	4.93	5.80	0.1079	0.0183	0.1351	0.9250
Glaso	12.50	21.91	0.2211	0.1596	0.3995	0.9388
Al-Marhoun	12.50	13.31	0.2034	0.0820	0.2863	0.96116
Dokla & Osman	9.67	11.52	0.1809	0.0575	0.2397	0.9146
Macary & El-Batanony	−7.31	8.01	0.1110	0.0234	0.1531	0.9119
Petrosky & Farshad	13.55	13.59	0.1981	0.0852	0.2920	0.9278

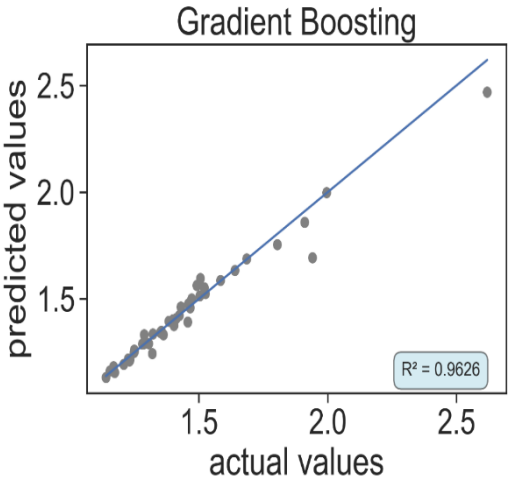
Prediction Performance for OIL FORMATION VOLUME FACTOR (BO) (bbl/STB) for Models Evaluated Applied to the Testing Subset data records)

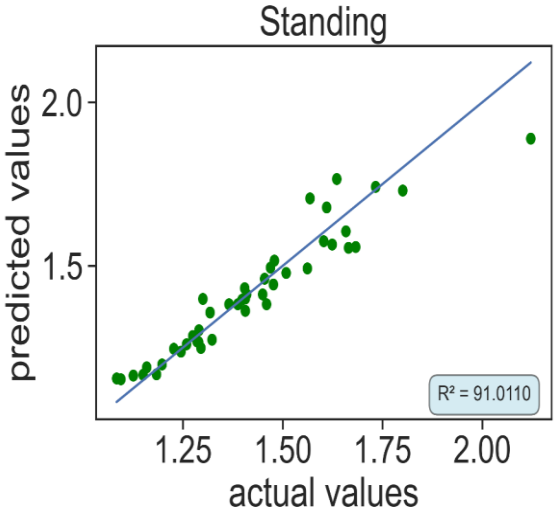
Models	APD%	AAPD%	SD	MSE	RMSE	R^2
Kartoatmodjo & Schmidt	14.34	15.01	0.2159	0.0990	0.3147	0.5158
Almehaideb	13.70	14.64	0.2227	0.0988	0.3142	0.6577
Machine Learning Models						
Decision Tree	-	-	-	0.0027	0.0519	0.9429
Random forest	-	-	-	0.0016	0.0399	0.9800
GB	-	-	-	0.0027	0.0524	0.96259
ANN	-	-	-	0.0011	0.0333	0.9850
SVM				0.0016	0.0399	0.9496

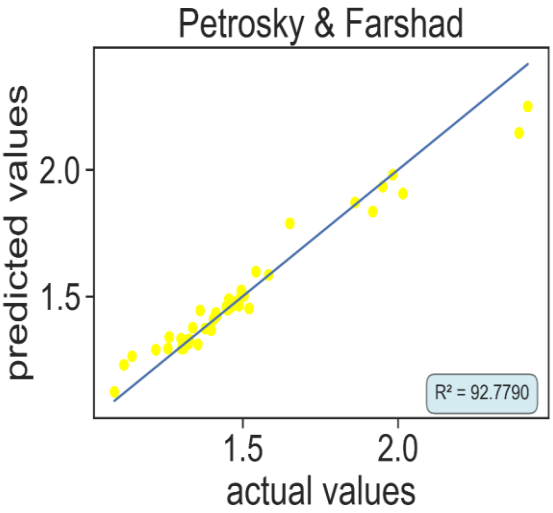


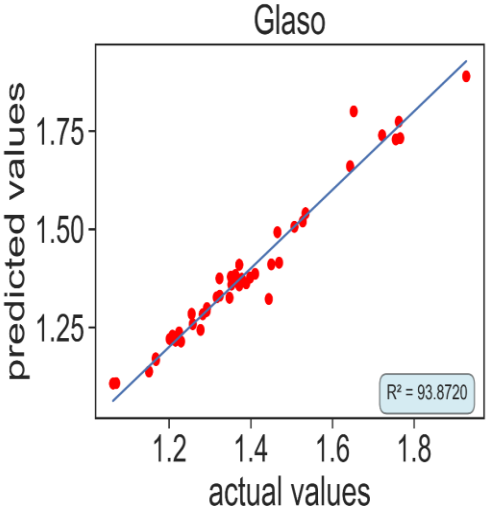


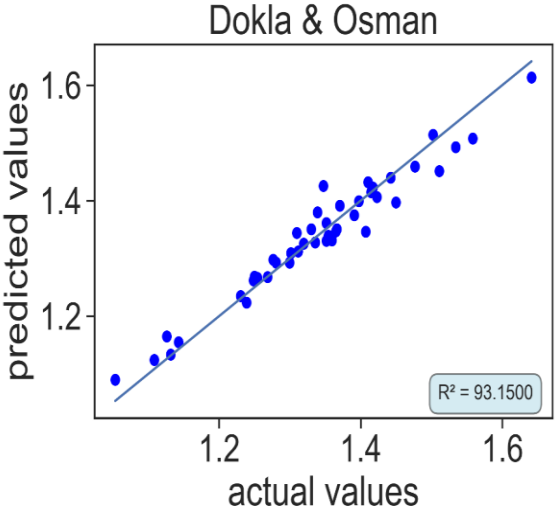

RF
predicted values
actual values
R² = 0.9800

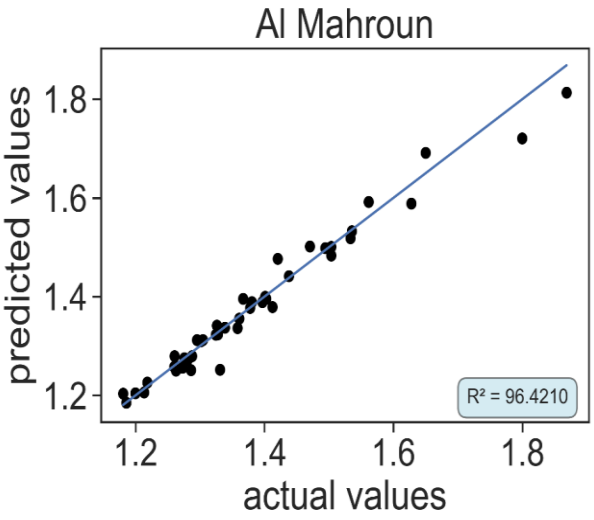

Gradient Boosting
predicted values
actual values
R² = 0.9626


Standing
predicted values
actual values
R² = 91.0110


Petrosky & Farshad
predicted values
actual values
R² = 92.7790


Glaso
predicted values
actual values
R² = 93.8720


Dokla & Osman
predicted values
actual values
R² = 93.1500



The best R^2 value realized (ANN, 0.985) suggests that there remains scope for machine-learning models to further improve upon these OIL FORMATION VOLUME FACTOR (BO) predictions.). This analysis is useful in establishing the OIL FORMATION VOLUME FACTOR (BO) ranges over which each correlation and algorithm perform well and less well.
The OIL FORMATION VOLUME FACTOR (BO) prediction errors models are smaller than the experimental models with a mean close to zero. The error distributions are asymmetric for all models.

The machine learning model demonstrated superior accuracy and reliability in predicting Bo compared to traditional empirical correlations. This improvement can be attributed to the model's ability to capture complex, non-linear relationships between reservoir properties that are often missed by empirical approaches.

Unlike empirical correlations, which are often region-specific, the machine learning model generalized well across different reservoir conditions. This broad applicability makes it a valuable tool for petroleum engineers working in diverse geographical locations.

Despite the promising results, there are limitations and challenges to consider. The performance of the machine learning model is highly dependent on the quality and diversity of the training data. Additionally, the black-box nature of machine learning models can make it difficult to interpret the underlying relationships between variables. Future work could focus on addressing these challenges, potentially by integrating domain knowledge into the model development process.

Conclusion

In conclusion, our machine learning model significantly outperformed traditional empirical correlations in predicting the Oil Formation Volume Factor (Bo). By leveraging advanced algorithms and a robust dataset, we achieved higher accuracy and reliability, demonstrating the potential of machine learning in petroleum engineering applications. Future research should aim to further improve model interpretability and explore the integration of additional reservoir properties to enhance predictive performance.

2.2.1 Bibliography

Ahmed, T. (2018). ***Reservoir engineering handBook*****. Gulf professional publishing.**

Al-Marhoun, M. (1988). PVT Correlations for Middle East Crude Oils. ***Journal of Petroleum Technology*****,** ***40*****, 650– 666. https://doi.org/10.2118/13718-PA**

Al-Marhoun, M. A., & Osman, E. A. (2002). Using Artificial Neural Networks to Develop New PVT Correlations for Saudi Crude Oils. https://doi.org/10.2118/78592-MS

Al-Shammasi, A. A. (2001). A Review of Bubblepoint Pressure and Oil Formation Volume Factor Correlations. ***SPE Reservoir Evaluation & Engineering*****,** ***4*****(02), 146–160. https://doi.org/10.2118/71302-PA**

Boukadi, F., Al-Alawi, S., Al-Bemani, A., & Al-Qassabi, S. (1999). Establishing PVT correlations for Omani oils. Petroleum Science and Technology, 17(5), 637–662. https://doi.org/10.1080/10916469908949738

Dutta, S., & Gupta, J. P. (2010). PVT correlations for Indian crude using artificial neural networks. Journal of Petroleum Science and Engineering, 72(1–2), 93–109. https://doi.org/10.1016/J.PETROL.2010.03.007

El-Sebakhy, E. A. (2009). Forecasting PVT properties of crude oil systems based on support vector machines modeling scheme. Journal of Petroleum Science and Engineering, 64(1–4), 25–34. https://doi.org/10.1016/J.PETROL.2008.12.006

El-Sebakhy, E. A. (2009). Forecasting PVT properties of crude oil systems based on support vector machines modeling scheme. ***Journal of Petroleum Science and Engineering*****,** ***64*****(1–4), 25–34. https://doi.org/10.1016/J.PETROL.2008.12.006**

El-Sebakhy, E., Sheltami, T., Al-Bokhitan, S., Shaaban, Y., Raharja, I., & Khaeruzzaman, Y. (2007). Support vector machines framework for predicting the PVT properties of crude-oil systems. SPE Middle East Oil and Gas Show and Conference, MEOS, Proceedings, 3, 1416–1429. https://doi.org/10.2118/105698-MS

Gharbi, R. B., Elsharkawy, A. M., & Karkoub, M. (1999). Universal neural-network-based model for estimating the PVT properties of crude oil systems. Energy and Fuels, 13(2), 454–458. https://doi.org/10.1021/EF980143V

Goda, H. M., Shokir, E. M. E. M., Fattah, K. A., & Sayyouh, M. H. (2003). Prediction of the PVT data using neural network computing theory. Society of Petroleum Engineers - Nigeria Annual International Conference and Exhibition 2003, NAICE 2003. https://doi.org/10.2118/85650-MS

Glaso, O. (1980). Generalized Pressure-Volume-Temperature Correlations. ***Journal of Petroleum Technology*****,** ***32*****(05), 785–795. https://doi.org/10.2118/8016-PA**

Gouda, A., & Attia, A. M. (2024). Development of a new approach using an artificial neural network for estimating oil formation volume factor at bubble point pressure of Egyptian crude oil. ***Journal of King Saud University - Engineering Sciences*****,** ***36*****(1), 72–80. https://doi.org/10.1016/J.JKSUES.2022.08.001**

Honarpour, M. M., Nagarajan, N. R., & Sampath, K. (2006). Rock/Fluid Characterization and Their Integration— Implications on Reservoir Management. ***Journal of Petroleum Technology*****,** ***58*****(09), 120–130. https://doi.org/10.2118/103358-JPT**

Karimnezhad, M., Heidarian, M., Kamari, M., & Jalalifar, H. (2014). A new empirical correlation for estimating bubble point oil formation volume factor. ***Journal of Natural Gas Science and Engineering*****,** ***18*****, 329–335. https://doi.org/10.1016/J.JNGSE.2014.03.010**

Rashidi, S., Mehrad, M., Ghorbani, H., Wood, D. A., Mohamadian, N., Moghadasi, J., & Davoodi, S. (2021). Determination of bubble point pressure & oil formation volume factor of crude oils applying multiple hidden layers extreme learning machine algorithms. *Journal of Petroleum Science and Engineering*, *202*, 108425. https://doi.org/10.1016/J.PETROL.2021.108425

Saghafi, H. R., Rostami, A., & Arabloo, M. (2019). Evolving new strategies to estimate reservoir oil formation volume factor: Smart modeling and correlation development. *Journal of Petroleum Science and Engineering*, *181*, 106180. https://doi.org/10.1016/J.PETROL.2019.06.044

Salim Basaleh, S. (n.d.). *Bin-Gadeem Salem Mubarak Saleh (1) Bin-Gadeem Ali Salem Mubarak (3)*. Standing, M. B., & Katz, D. L. (1942). Density of Natural Gases. *Transactions of the AIME*, *146*(01), 140–149. https://doi.org/10.2118/942140-G

Khoukhi, A. (2012). Hybrid soft computing systems for reservoir PVT properties prediction. Computers and Geosciences, 44, 109–119. https://doi.org/10.1016/J.CAGEO.2012.03.016

Malallah, A., Gharbi, R., & Algharaib, M. (2006). Accurate estimation of the world crude oil PVT properties using graphical alternating conditional expectation. Energy and Fuels, 20(2), 688–698. https://doi.org/10.1021/EF0501750

Moghadam, J. N., Salahshoor, K., & Kharrat, R. (2011). Introducing a new method for predicting PVT properties of Iranian crude oils by applying artificial neural networks. Petroleum Science and Technology, 29(10), 1066–1079. https://doi.org/10.1080/10916460903551040

Rafiee-Taghanaki, S., Arabloo, M., Chamkalani, A., Amani, M., Zargari, M. H., & Adelzadeh, M. R. (2013). Implementation of SVM framework to estimate PVT properties of reservoir oil. Fluid Phase Equilibria, 346, 25–32. https://doi.org/10.1016/J.FLUID.2013.02.012

Salehinia, S., Salehinia, Y., Alimadadi, F., & Sadati, S. H. (2016). Forecasting density, oil formation volume factor and bubble point pressure of crude oil systems based on nonlinear system identification approach. Journal of Petroleum Science and Engineering, 147, 47–55. https://doi.org/10.1016/J.PETROL.2016.05.008

Seyyedattar, M., Ghiasi, M. M., ZendehBoudi, S., & Butt, S. (2020). Determination of bubble point pressure and oil formation volume factor: Extra trees compared with LSSVM-CSA hybrid and ANFIS models. Fuel, 269. https://doi.org/10.1016/J.FUEL.2019.116834

Shokrollahi, A., Tatar, A., & Safari, H. (2015). On accurate determination of PVT properties in crude oil systems: Committee machine intelligent system modeling approach. Journal of the Taiwan Institute of Chemical Engineers, 55, 17–26. https://doi.org/10.1016/J.JTICE.2015.04.009

Vazquez, M., & Beggs, H. D. (1977). Correlations for Fluid Physical Property Prediction. In *SPE Annual Fall Technical Conference and Exhibition* (p. SPE-6719-MS). https://doi.org/10.2118/6719-MS

Wood, D. A., & Choubineh, A. (2019). Reliable predictions of oil formation volume factor based

on transparent and auditable machine learning approaches. *Advances in Geo-Energy Research,*

***3*(3), 225–241. https://doi.org/10.26804/ager.2019.03.01**