Flow Measurement and Instrumentation 88 (2022) 102257

Contents lists available at ScienceDirect

Flow Measurement and Instrumentation

journal homepage: www.elsevier.com/locate/flowmeasinst

Adoption of machine learning in estimating compressibility factor for natural gas mixtures under high temperature and pressure applications

Emmanuel Emeka Okoro [a,*], Ekene Ikeora [b], Samuel E. Sanni [c], Victor J. Aimihke [a], Oscar I. Ogali [a]

[a] Department of Petroleum and Gas Engineering, University of Port Harcourt, Choba, Nigeria
[b] Department of Petroleum Engineering, Covenant University, Ota, Nigeria
[c] Department of Chemical Engineering, Covenant University, Ota, Nigeria

ARTICLE INFO

Keywords:
Z-factor
Pseudo reduced temperature and pressure
HPHT reservoir
Artificial intelligence model

ABSTRACT

One of the essential properties of natural gas is its compressibility factor (z-factor), which is required for the efficient design of natural gas pipelines, storage facilities, gas well testing, gas reserve estimation, etc. Its importance has led to the development of several approaches involving new laboratory methods, equations of state (EOS), empirical correlations, and artificial intelligence for estimating gas compressibility factors. Most of the developed Z factor models have a limited range of applicability. They are unsuitable for predicting Z factors of highly pressurized gas reservoirs and natural gas systems with pseudo-reduced temperatures less than 1. Where such models exist, they are scarce and less accurate. In this study, three machine learning models, including the Gradient Boosted Decision Tree (GBDT), Support Vector Regression (SVR), and Radial Basis Function-Neural Network (RBF-NN), were developed for predicting the z-factor of natural gas mixtures with a range of P_{pr} and T_{pr} of 0–30 and 0.92–3.0, respectively. The results showed that the Gradient Boosted Decision Tree (GBDT) model outperformed other selected machine learning algorithms and published correlations. The proposed model gave a superior coefficient of determination (R^2 score), and root mean square (RMSE) of 0.99962 and 0.01033, respectively. Also, the variation of the Z factors from the GBDT model with pseudo-reduced pressures at different pseudo-reduced temperatures using the isotherm plot was found to be adequate. Hence, the GBDT model in this study is a reliable method for predicting Z factors of natural gas mixtures with P_{pr} and T_{pr} of 0–30 and 0.92–3.0, respectively. The plot revealed that the GBDT model performed extremely well in predicting compressibility factor with an MAPE of about 1%. The findings of this study shows that the proposed intelligent model can be utilized in predicting the gas Z-factor.

1. Introduction

The compressibility factor, also known as the gas deviation factor or Z-factor, is a correction factor that defines how a real gas behaves compared to ideal gas behavior. The gas compressibility factor, also known as Z-factor, plays the determinative role for obtaining thermodynamic properties of gas reservoir. Literature have highlighted that obtaining fluid properties from gas and oil reservoirs has been of great importance to many researchers and petroleum engineers. The significance of this knowledge becomes more brilliant when the oil and gas capacity of reservoirs, dissolved gas, aquifer model and other reservoir properties depends directly or indirectly on fluid properties [1]. It compares the molar volume of gas in real conditions to the molar volume of the same gas at ideal conditions to determine how far a gas deviates from its ideal behavior at a similar temperature and pressure. The Z-factor is extensively used in numerous engineering disciplines when working with gases at low pressure and temperature, ideal gas relationship is a useful and typically satisfactory method. The application of the ideal gas equation causes errors at higher pressures and temperatures because Z-factor is usually defined as a function of reduced temperature and reduced pressure [1]. Therefore, the Z-factor is introduced to correct these errors. The gas compressibility factor (Z-factor) is a critical parameter in upstream and downstream operations, an accurate estimation of the property will be evident in material balance, gas reservoir simulation, gas reserve evaluation, gas processing and well control calculations [2]. The Z-factor values are commonly determined

* Corresponding author.
E-mail address: emeka.okoro@uniport.edu.ng (E.E. Okoro).

https://doi.org/10.1016/j.flowmeasinst.2022.102257
Received 19 April 2022; Received in revised form 26 August 2022; Accepted 17 October 2022
Available online 21 October 2022

Table 1
Some selected Literature relevant to this Study.

Author's Work	Aim of the Study	Method/Approach Adopted	Outcome of the Study	Gaps in the Study
Azizi et al. [13]	To develop an explicit correlation for Z-factor	Mathematical model which contained 20 tuned coefficients.	Accurate correlation which rapidly estimates Z-factor for sweet gases.	Valid for a narrow range of pseudo-reduced Pressure and Temperature: $0.2 \leq P_r \leq 11$ (217 P_r values) and $1.1 \leq T_r \leq 2$ (14 T_r values).
Festus and Ikiensikimama [14]	Evaluation of widely used natural gas Z-factor correlations for reservoirs in the Niger Delta	A statistical approach and also the use of cross plots to evaluate the best correlation.	The best correlation for Niger Delta Field through the ranking method used was Beggs and Brill, with a percentage absolute error of 3.234.	The result is specific to Niger Delta fields and was ranked with a narrow range $0.5796 \leq T_r \leq 1.758$ and $0.410 \leq P_r \leq 8.985$.
Heidaryan et al. [2]	To develop accurate correlation to quickly estimate Z-factor as a function of T_{pr} and P_{pr}	Multiple rational regression equation - a numerical method	The correlation performed better than three common Z-factor correlations when compared by their statistical parameters.	The equation developed was not suitable for estimating Z-factor of $T_{pr} < 1.2$.
Kamyab et al. [15]	To obtain a method for predicting Z-factor	Back Propagation-Artificial Neural Network with data gotten directly from Standing and Katz (S&K) chart	The results showed that the ANN model performed better, faster and covered a wider range of reduced pressure.	Although the model has shown good results for the Average Absolute Error (0.1060), it can be improved. Also, it is not for HPHT reservoirs.
Al-Anazi et al. [16]	Prediction of compressibility factor for sour and natural gases	A Multi-layer feed forward neural network based on experimental data obtained from seven studies.	From the statistical parameter criteria AAD, RMSE, and R^2 gave 0.965, 0.024 and 0.991 respectively.	From the analysis of the proposed and investigated Z-factor models, it is observed that the new model did not perform better than most of the other models in terms of R^2.
Baniasadi et al. [17]	Prediction of Natural gas compressibility factor	Artificial neural network to develop the model based on the input M-factor, T_r and P_r.	The developed model had an incredible accuracy (R^2 value of 0.99992) which was better than the other equations compared.	Low ranges of P_r ($0.02 \leq P_r \leq 8$) and T_r ($1 \leq T_r \leq 2$) which means low range of applicability.
Sanjari and Lay [6]	A simple empirical correlation that rapidly predicts Z-factor of natural gas that outperforms other empirical correlations.	Use of experimentally derived data to develop correlation.	The new technique outperformed the other methods with an average absolute relative deviation (AARD) of 0.6535.	The proposed correlation contains many tuned (8) coefficients and their values depend on the range of P_r values.
Sanjari and Lay [18]	The use of Experimental data in designing an ANN to estimate the Z-factor of natural gas.	Artificial neural network (ANN) based on back-propagation method	The neural network predicted natural gas compressibility factors using T_{pr} and P_{pr} with AARD of 0.593.	According to the study, a single statistical measurement was used to validate the accuracy of the ANN model. Only a single hidden layer was used. An increase in the layers can achieve better results.
Shokir et al. [19]	Development of a simplified model that predicts Z-factor in sweet, sour, rich and lean gas condensate reservoirs	The use of Genetic Programming based P_{pc} and T_{pc} models	The new method gave better results compared to other EOS and correlations considered in the study.	A short range of Z-factor can be predicted ($0.456 \leq z \leq 1.361$)
Chamkalani et al. [20]	To develop a smart, but precise model to predict the Z-factor within wide ranges of T_{pr} and P_{pr} conditions.	Coupled simulated annealing (CSA) algorithm with the conventional Least Square Support Vector Machine, known as CSA-LSSVM	The CSA-LSSVM outperformed the ANN it was compared to.	The model comes with a noisy pattern which is observed when the P_r is lower than 2.
Kamari et al. [21]	Prediction of Z-factor for sour gases using an intelligent approach.	CSA-LSSVM	Based on the results, CSA-LSSVM outperformed all the equations of state and the empirical correlations.	The data used to build the model has narrow ranges of values and the accuracy outside the range of values for training is questionable.
Fatoorehchi et al. [22]	Development of an explicit formula to predict natural gas Z-factor.	The Adomian decomposition method was used.	The formula developed was found to be good in converging to an accurate Z-factor.	Although accurate, a lot of variables and computation is involved.
Fayazi et al. [23]	To estimate gas Z-factor	Machine learning algorithm – LSSVM	The LSSVM model outperformed the other existing predictive models % AARE of 0.19 and correlation coefficient of 0.999.	The range of pressures and temperatures where the model thrives is unclear. The model may not be general since the composition, which is particular to a single fluid, is considered in developing the model.
Ghiasi et al. [24]	To estimate gas Z-factor for retrograde gas systems	LSSVM based on Constant Volume Depletion	The proposed model outperformed the 120 other models it was compared with.	From the results, the new model had an R^2 value of 0.970. This accuracy can be improved.
Li et al. [25]	To predict gas Z-factor of gas condensate for wide range of pressures.	A model based on EPT EOS combined with Elliott and Daubert binary interaction coefficient correlation, Hosseinifar and Jamshidi characterization methods, and Ahmed et al. (1985) splitting.	The proposed model outperformed ten other empirical correlations with R^2, AARE and MSE of 0.989, 1.45% and 0.000577, respectively.	The model developed was not based on dimensionless parameters (P_r, T_r) and therefore will not be applicable to every gas condensate.
Mahmoud [26]	Determination of gas compressibility factor for High pressure gas reservoirs	Linear Regression function to fit data and develop a new correlation	The correlation developed predicts Z-factor for very high-pressure gas reservoirs: 20,000 psi for mixture or P_r of 30.	Applicable temperature range for the model was not considered.
Mohamadi-Baghmolaei et al. [7]	To predict accurately the compressibility factor	ANN, Fuzzy Interface System (FIS), Adaptive Neuro-Fuzzy System (ANFIS) and Equation of State optimization with Genetic Algorithm	The accuracy of intelligent models proved to be better than the empirical models. Also, an improvement was observed	The data used to develop the model had Z-factor values between 0.66 and 2.04. Expanding the application range is therefore necessary.

(*continued on next page*)

Table 1 (*continued*)

Author's Work	Aim of the Study	Method/Approach Adopted	Outcome of the Study	Gaps in the Study
			for the optimized EOS. Overall, the ANN was the most accurate with an R^2 of 0.9999.	
Sarrafi et al. [27]	Accurate determination of compressibility factor	Adaptive Network-based Fuzzy Interface System with Neural Network	From the statistical analysis of ANFIS and other older models, in particular the DAK correlation, the ANFIS yielded the highest accuracy.	The study did not consider $T_r < 1$. This range is necessary in order to meet more areas across downstream and upstream.
Shateri et al. [28]	Prediction of natural gas Z-factor	Wilcoxon generalized radial basis function network (WGRBFN) was applied using 978 data point obtained from literature.	The data was split 50-50 for training and test data and has an R^2 for the training set and test set of 0.9241 and 0.9195 respectively.	Although the robustness of the model was shown through the results, there is room for improved accuracy of the proposed model.
Azubuike et al. [1]	To estimate Z-factor using 513 data point from Niger Delta region.	Back Propagation Neural Network with Levenberg-Marquardt procedure for the optimization	The model developed had a good accuracy in terms of rank (lowest rank – 1.37) and performance plot.	Small range of T_{pr} and P_{pr} for Z-factor prediction
Azubuike et al. [29]	Estimation of Z-factor for HPHT natural gas systems and the evaluation of selected Z-factor correlation	Laboratory measurement of natural gas Z-factor	Selected correlation for study was seen to perform well within low pressure ranges and show higher deviation at elevated pressure region.	The evaluation was done based on narrow range of pressures and temperatures.
Kamari et al. [30]	Estimation of gas Z-factor for natural gas	A gene expression programming (GEP) algorithm was employed to create a corresponding of states model.	The superiority of the model was shown from statistical and graphical analysis. It had an R^2, E_a, and RMSE of 0.898, 3.45 and 0.04, respectively	The model had a short range of P_r, T_r and z, therefore having a short range of applicability.
Azizi et al. [31]	To accurately predict gas Z-factor	Artificial Neural Network with a structure of 2:5:5:1	From the statistical analysis of Z-factor, the ANN was found to be more effective than other correlations.	Low range of application for the model developed due to the low range of data used.
Okoro et al. [32]	Accurate prediction of Z-factor	The use of gas well inflow performance data in Visual Basic.net	It was concluded that the Hall and Yarborough deviation model is the best for Niger Delta.	Although accurate when compared to the other models, it will be less accurate for wider ranges of temperature and pressure.
Salem et al. [33]	Predict and compare z- factor values from different AI techniques.	Radial Basis Function (RBF), ANN, Fuzzy Logic (FL), Functional Network (FN) and Support Vector Machine (SVM).	The results showed that Neural Network could predict Z-factor better than other AI techniques.	The accuracy can be further improved with further optimizing or modeling techniques.
Ekechukwu and Orodu [34]	An explicit correlation for the accurate determination of compressibility factor.	Hybrid optimization technique - Levenberg-Marquart Algorithm -orthogonal distance regression (LMA-ODR) model	The model developed performed better than the other correlations according to the statistical performance metrics	The range of pseudo-reduced temperature is shorter than the original range for S&K chart. This should be improved.
Gaganis et al. [4]	To develop an efficient method that predicts gas Z-factor for natural gas streams	A hybrid modeling technique was employed which combined Truncated Regularized Kernel Ridge Regression (TR-KRR) algorithm, with a simple linear-quadratic interpolation scheme	The model performed extremely well when compared with the S–K chart and other correlations.	The model was built by combining three models for each specific region of P_r values but, achieving similar or greater results with wider ranges of P_r and T_r should be considered.
Maalouf et al. [35]	To accurately predict gas Z-factor	Truncated Regularized – Kernel Ridge Regression algorithm	The TR-KRR model is more computationally efficient than the traditional SVM.	Comparison was done between two models. Improved models can be implemented to achieve higher accuracy.
Sidrouhou et al. [36]	To match Z-factor from existing correlation that best corresponds to the Algerian natural gas.	Experimental study on Algerian natural gas in a PVT laboratory.	The study showed reduction in error for the modification of model by updating the coefficients of the model from experimental results.	The study is specifically for a single reservoir in Algeria.
Tariq and Mahmoud [37]	To predict gas Z-factor for HPHT gas reservoirs	Artificial Neural Network model trained with Levenberg−Marquardt algorithm	A model was created to predict Z-factor for $0.1 < P_r < 40$.	Although the accuracy was good, the proposed ANN did not outperform some of the older correlations using the R^2 metric.
Lin et al. [38]	To develop an efficient model for gas Z-factor prediction	Group Method Data Handling (GMDH) Network	Comparison between the estimated value and expected value of z was given, and it was observed that the R^2 was close to 0.999. The result of the proposed model gave the lowest RMSE of 0.0066.	According to the study, the model is accurate but fails to meet a wide range of application. The calculation range is set to $0 \leq P_{pr} \leq 12$, $1.1 \leq T_{pr} \leq 2.1$.
Ogbunike and Adeyemi [39]	Predicting compressibility factor for high pressure and high temperature.	Stochastic and robust gradient-based optimization algorithm	The error analysis showed that the new model was accurate.	More data can be used to improve accuracy and range of applicability.
Wang et al. [40]	To develop a novel empirical formula of natural gas compressibility factors is obtained, which is suitable for the pseudo-reduced pressure range of 0.2–30.	Multivariate nonlinear regression is used to fit the 6988 data of the Standing–Katz chart	The verification result shows that the mean absolute error, mean relative error and root mean square error between the calculated values and the measured values are 0.01962, 0.01626 and 0.02511 respectively. The proposed correlation is superior to the other five methods because of its higher calculation accuracy.	The model is suitable for calculating natural gas compressibility factors in the range of $0.2 \leq P_{pr} \leq 30$ and $1.05 \leq T_{pr} \leq 3.0$.

through equations of state, experimental measurements, and empirical correlations.

An accurate experimental measurement is the most reliable method of obtaining compressibility factor. However, these experiments are time-intensive and costly. It is also virtually impossible to measure the properties of every possible composition of gases. Consequently, this approach is seldom used [3]. Several "simple" empirical correlations have been developed, with standard Z-factor charts used to evaluate the accuracy of these correlations. Sometimes, these correlations yield poor results, due to their calculation convergence issues or low precision [4]. With the emergence of new technology and techniques, many authors have sought out new ways to improve accuracy in the prediction of Z-factor under various conditions. Some have taken the approach of digitizing the Standing and Katz chart for ease of application [5]. Some of the currently available methods in literature are applicable only to a limited pseudo reduced pressure range, usually $P_{pr} < 15$; some of these models did not consider low ranges of pseudo reduced temperature ($T_{pr} < 1$). The application of two or more hybrid models has led to discontinuity at certain boundaries. Recently, several studies [6–9], sought to accurately predict the Z-factor but, the wide range which includes low pressures and temperatures as well as high pressures and temperatures were seldom considered. This is because, the focus of these studies was on improving prediction accuracy.

The use of Artificial Intelligence (AI) approaches to solve problems in the oil and gas industry has become very popular, due to the non-linear relationship between input parameters and output parameter. Many published literatures have used powerful artificial intelligence/machine learning techniques to achieve incredible results due to their response speed and capability to pick hard-to-spot trends in the data provided [10–12]. Recently, explicit correlations and AI models have been developed or modified to consider some of the challenges in estimation of gas compressibility factor. A critical issue is accurately estimating the gas Z-factor for high pressure and high temperature (HPHT) reservoirs. This would require higher values for Z-factor dependent properties which, for most cases, is just the T_{pr} and P_{pr}. The advantage of AI (such as Artificial Neural Networks, ANNs) over traditional correlations is that, when it comes to fitting parameters, neural networks contain several degrees of freedom. Therefore, they can capture non-linearity better than regression methods. The ANN technique can be used to model various scientific problems in engineering domains and they are also superior to regression models in that they can be trained and improved as new data becomes available, thereby improving prediction accuracy [1].

Table 1 shows some relevant literatures to this study where the aims, approaches or methods adopted, and results of the study are summarized.

One of the essential properties of natural gas is its compressibility factor (z-factor), which is required for the efficient design of natural gas pipelines, storage facilities, gas well testing, gas reserve estimation, etc. Its importance has led to the development of several approaches involving new laboratory methods, equations of state (EOS), empirical correlations, and artificial intelligence for estimating gas compressibility factors. Most of the developed Z factor models have a limited range of applicability. They are unsuitable for predicting Z factors of highly pressurized gas reservoirs and natural gas systems with pseudo-reduced temperatures less than 1. Where such models exist, they are scarce and less accurate. The estimation of gas compressibility factor is important because it is the relative change in the volume of the gas with respect to the change in pressure at constant temperature.

In dealing with gases at a very low pressure, the ideal gas relationship is a convenient and generally satisfactory tool. At higher pressures, the use of the ideal gas equation-of-state may lead to errors. Thus, the advantage of this study; basically, the magnitude of deviations of real gases from the conditions of the ideal gas law increases with increasing pressure and temperature and varies widely with the composition of the gas. Studies of the gas compressibility factors for natural gases of various

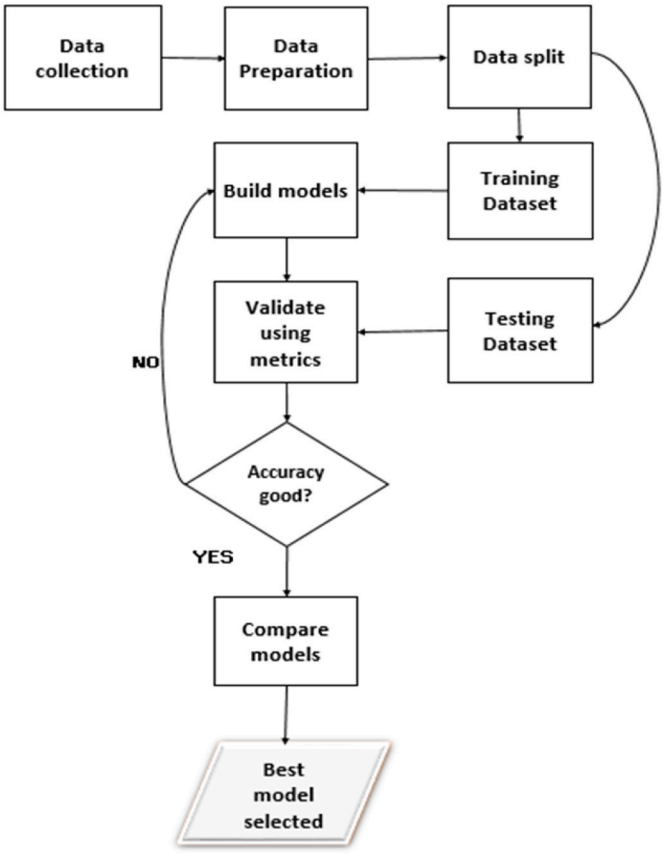

Fig. 1. Flowchart showing the Model Workflow.

compositions have shown that compressibility factors can be generalized with sufficient accuracies for most engineering purposes when they are expressed in terms of the following two dimensionless properties: Pseudo-reduced pressure and Pseudo-reduced temperature.In the gas industry, it is an important tool for computing reservoir fluid properties either directly or indirectly. Accurate estimation of compressibility factor (z) is very essential, most especially when it comes to quick estimation of initial gas in place. It is also an important factor to rely on when dealing with gas metering, where the volume flow of gas obtained from the orifice meter depends on the accuracy of the z-factor. Natural gas compressibility factor (z) is also a key factor in the gas industry for natural gas production and transportation.

Tariq and Mahmoud [37], attempted to increase the range of application of Z-factor by developing model for pseudo-reduced pressure ranging from 0.1 to 40 and pseudo-reduced temperature ranging from 1.05 to 3.05 using ANN algorithm. While considering high-pressure high-temperature (HPHT) conditions, they did not account for T_{pr} less than 1. Orodu et al. [41], considered T_{pr} less than 1 but, did not extend the P_{pr} to 30. Liu et al. [42] through experimental results of five groups of HPHT natural gas samples indicated that the Z-factors of these natural gases under reservoir conditions are considerably higher than those of conventional natural gases. Wang et al. [40] proposed a correlation for calculating gas Z-factor for a wide range of pressure conditions using the data from Standing-Katz chart; they did not cover the HPHT conditions. Having a simple and robust correlation to identify Z-factor values for HPHT reservoirs has become a requirement in the oil and gas industry. This study will delve into AI and machine learning techniques that avoids using more than one hybrid model which implies the combination of models in predicting results. Single models will be analyzed and compared through statistical criteria to achieve accurate prediction of compressibility factor for wide ranges of P_{pr} and T_{pr} less than 1. The objective of this study is to predict accurately the

Table 2
Statistical analysis of the data points.

Parameters	P_{pr}	T_{pr}	Z-factor
Count	1954	1954	1954
mean	14.1145	1.8576	1.3603
std	9.3403	0.5974	0.5111
min	0	0.92	−0.1164
25%	4.9572	1.4	0.9536
50%	13.9828	1.8	1.4231
75%	22.375	2.4	1.7429
max	30	3	2.66

compressibility factors for wide ranges of pseudo-reduced pressure and temperature using single algorithm for the full range ($0 \leq P_{pr} \leq 30$; $0.92 < T_{pr} \leq 3.0$).

2. Methodology

The main goal of intelligent software is to connect sets of input and output variables while keeping the system specifications in mind [7]. Intelligent models are more effective in time-consuming situations involving non-linear mathematical modeling, adaptive learning, and no significant relationship between a system's input and output. Amongst the various machine learning classification, supervised learning fits the purpose of this study. Supervised learning can either be a regression task or a classification task. If the model's expected outputs are continuous values, a supervised learning technique will be used for the regression task, whereas a classification task will have outputs in predetermined classes. For the prediction of compressibility factor, supervised learning technique was employed. Two inputs (T_{pr} and P_{pr}) were used to predict output (Z-factor). Three different models were built and evaluated based on some selected metrics to select the best performing model for a wide range of T_{pr} and P_{pr}.

2.1. Workflow for the model

Fig. 1 shows a simple workflow for the study starting from the data collection phase down to the selection of the best model. The following tools and libraries were used in this study: Python 3 (Jupyter Lab), Numpy, Pandas, Scikit-learn, Matplotlib, Seaborn and Pickle.

2.2. Data gathering and processing

The initial gas compressibility data was obtained from an HPHT gas field in Gulf of Guinea, and more data set were simulated using Beggs and Brill's correlation inputted into a Python program to generate a wide range of T_{pr}. Beggs and Brill's correlation was selected because it works accurately at low temperatures but begins to deviate at higher temperatures. The source data can be found online (https://github.com/f0nzie/zFactor/blob/master/notebooks/SK%20data.xlsx). The combined source data contained 57,060 data points for $0 < P_{pr} < 15$ and $0 < T_{pr} < 3.904$ data points were within the range of $16 < P_{pr} < 30$ and $1.4 < T_{pr} < 2.8$, and finally 150 data points were within the range of $0 < P_{p\,r} < 14.85$ and $0.92 < T_{pr} < 1.0$. Down sampling was done to the first batch of data points to reduce the chances of bias towards the range of values in the batch of data points, and also to avoid over fitting of the proposed models. Down sampling is basically lowering a larger set of data points of a particular class so as to match the set of data points. This is to allow for fairly equal representation of classes in the dataset. The limitation of this study is that the data composition did not account for CO_2 and H_2S impurities.

Quality assurance and quality control (QAQC) checks were performed on the pseduoreduced input parameters to ensure the reliability of the data. First, the data sets were checked for any duplicate values,

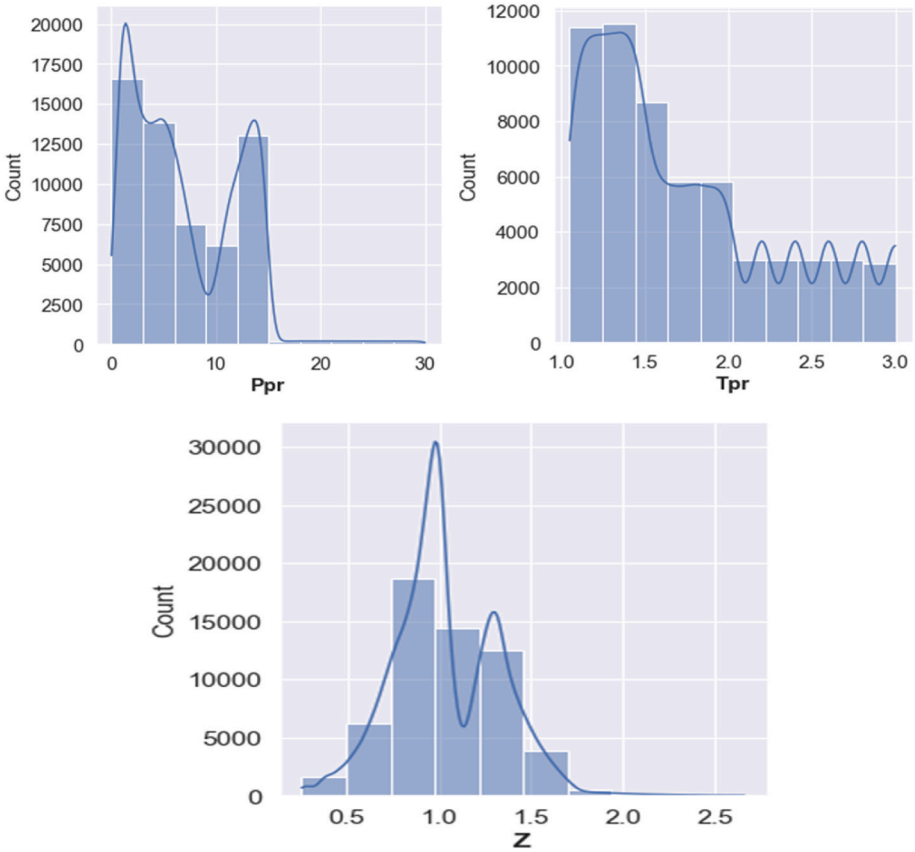

Fig. 2. Histogram plot of data points before processing.

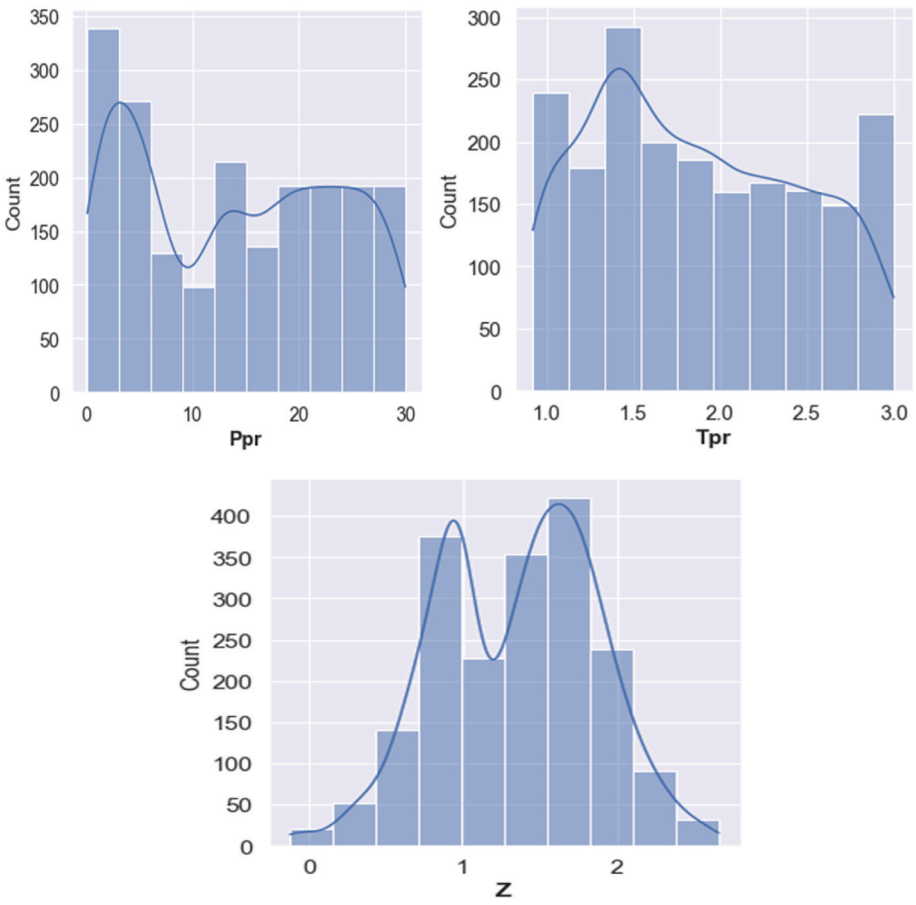

Fig. 3. Histogram plot of the processed data points.

and it was further analyzed for missing data within the range under consideration ($0 \leq P_{pr} \leq 30$; $0.92 < T_{pr} \leq 3.0$). After an initial analysis, some of the data point were discarded due to their inconsistency or existence as outliers and some were not relevant in developing the desired model due to their range. It is worth mentioning that the missing values can be imputed by a mid-point (arithmetic average, median, or mode). The mid-point is used as it can act as the best representative for the whole data.The processed data points were imported into the python interactive development environment. Using the software tools such as Pandas, NumPy and MatplotLib, tables and plots displaying statistical data were generated for the sum of 1954 data points used for the purpose of model building (Table 2).

Fig. 2 shows the histogram representation of the source data for P_{pr}, T_{pr} and Z-factor before data processing and this characteristic allows the investigation of the data for its underlying distribution. The shape of the distribution for the histogram shows right-skewed distributions for the P_{pr} and T_{pr} because, the long tail extends to the right while most of the data points cluster on the left, as shown in Fig. 2. Thus, the P_{pr} and T_{pr} distribution are not symmetrical. The Z-factor shape distribution shows a symmetric trend and Bi-modal distribution for the data points. Fig. 3 shows the data distribution for the processed data. Bi-modal distribution was observed for the Z-factor, indicating independent sources of variation. Both data distribution shows a right skewness, the data points naturally have a skewed distribution, because they are bounded, such as the concentricity data. Concentricity has a natural lower bound at zero, since no measurements can be negative. The majority of the data is just above zero.

2.3. Machine learning algorithms

After data processing, a total of 1954 data points were used to develop the model. These data points were randomly divided into three different groups: training, testing, and validation. Two different partitioning ratios were tested (2:1:1, and 3:1:1). However, the 3:1:1 partitioning rule yielded better training and testing results. Traditionally, to build a program that solves a particular problem, a step-by-step procedure or a sequence of lines to achieve the end result is written. This is what is commonly referred to as Algorithm. The program will generally take inputs provided, run it through the lines of code and generate output. Machine learning algorithms, through observing examples is trained to find complex connections between the inputs and output. In the learning phase, a model will take arbitrary input values within its applicable range to generate output.

2.3.1. Gradient Boosted Decision Tree (GBDT)

Gradient boosted decision trees are ensemble methods for classification and regression issues. Ensemble methods aim to increase generalizability or resilience over a single estimator by combining the predictions of numerous base estimators using a given learning algorithm. Gradient tree boosting necessitates the optimization of a loss function, the use of a weak learner to produce predictions, and the use of an additive model to combine poor learners in order to reduce the loss function. The loss function to be selected is dependent on the nature of the problem. There are various kinds of loss functions that can be utilized and it is possible to create custom loss functions. For example, squared error can be used if it is a regression task, while a logarithmic loss can be applied if the problem is a classification task. In gradient boosting, the weak or poor learners are decision trees [43]. The most common learners are trees because, the level of weakness can be altered by adjusting the depth parameter of the tree model (Fig. 4). At each level, a regression tree is fitted based on the negative gradient of the specified loss function. Trees are created greedily, with the best split

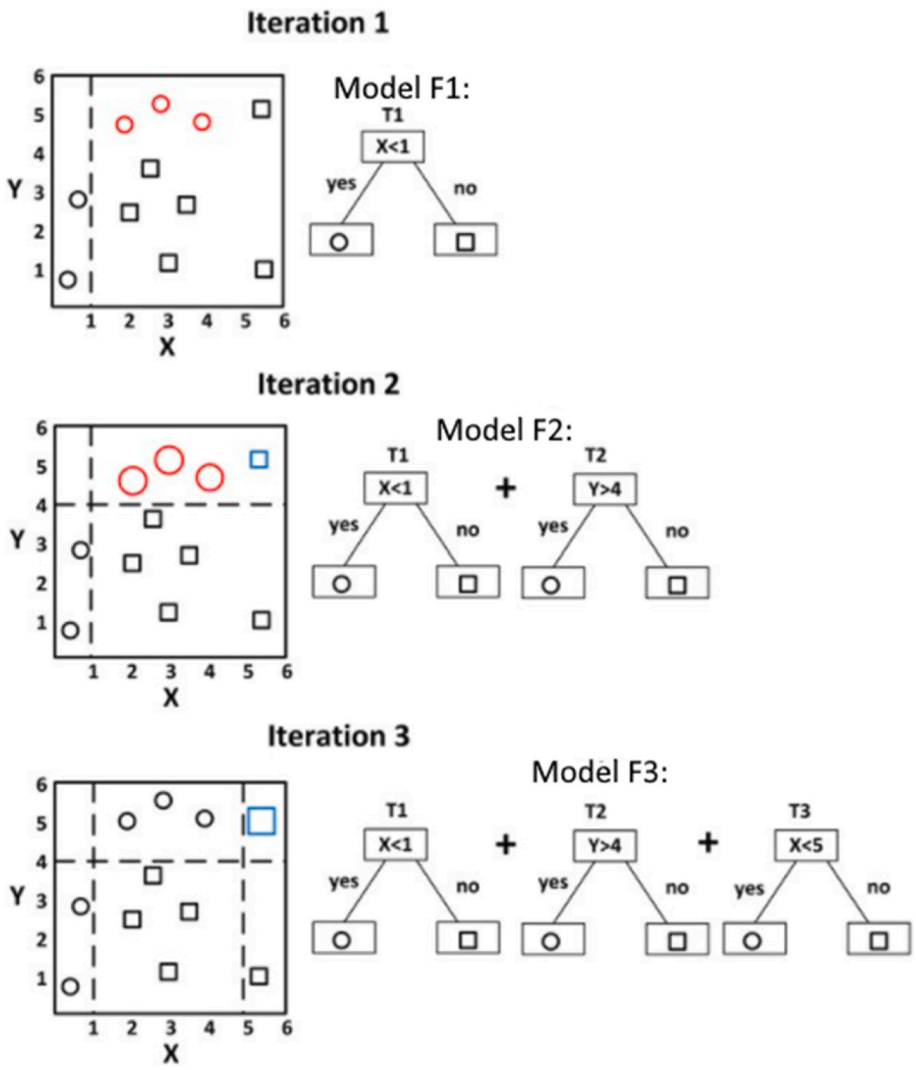

Fig. 4. Basic steps for Gradient Boost Trees [45].

points determined by purity scores like Gini or loss minimization. In gradient boost, the tree depth is generally more than that of its predecessor (Adaboost) but must not be too deep so as to keep the single estimator weak. Finally, Trees are added one at a time, and existing trees in the model are not altered.

Although the gradient boost tree regressor already optimizes and builds upon weak decision tree models, it is important to understand the effects of hyper-parameters on the outcome of the model. This study is limited to the most important parameters in decision trees which are:

i. Number of estimators:

This controls the number of boosting stages to perform. Gradient boosted trees hardly over fit, so a large number usually gives better performance [44]. This is accessed by varying the *n-estimator* keyword argument in the regressor model. By default, this is set to 100. The default was maintained in building the model.

ii. Maximum depth:

The amount of nodes in the tree is limited by the maximum depth. The keyword *max_depth* is used to regulate this parameter. This parameter should be tuned for optimal performance; the best value is determined by the interaction of the input variables. For this study, the maximum depth is set to 10.

2.4. Support vector regression (SVR)

Support vector machine is a machine learning algorithm that is well known for classification problems. They can also be employed for regression tasks as well. The extension of the support vector machine (SVM) to solving linear and non-linear regression problems is what is known as support vector regression.

2.4.1. Hyper-parameters for SVR

i. Kernel function:

This parameter is selected by specifying the keyword *kernel* with the desired kernel function in the SVR model being created. Linear kernel is used for linear problems while rbf, polynomial and sigmoid are used for non-linear problems.

ii. C: This parameter is common to all SVM kernels. It trades off accuracy in classifying training data against simplicity of the decision surface. A high C value will try to fit all points as accurate as possible while a low C value pays less attention to all the points and results in a smooth decision surface.
ii. Gamma: It controls the influence of a single point. The larger the value of gamma is the closer other points need to be to be affected.

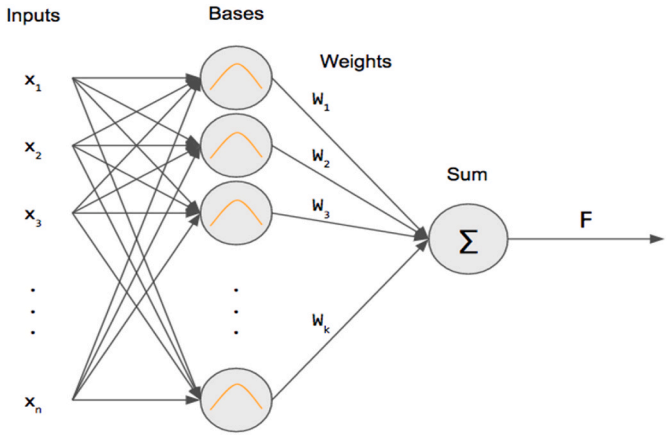

Fig. 5. An rbf-nn structure.

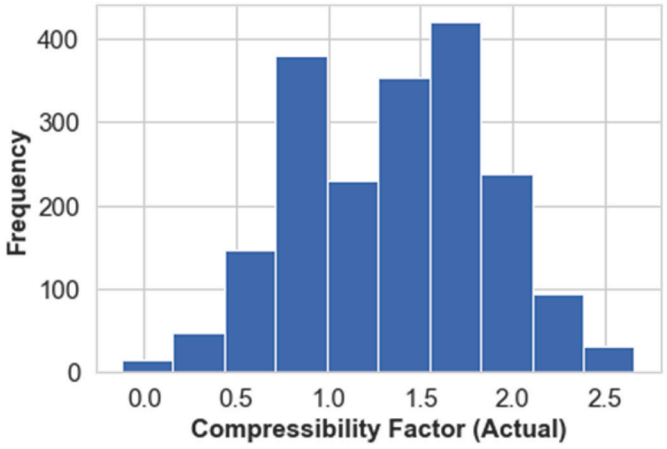

Fig. 6. Frequency distribution of the actual Z-factor.

Parameter tuning was done using grid search. Grid search is an optimization technique that allows us select the best set of hyper-parameters which gives the best score according to some scoring metric. This is done by running combinations for set of specified parameter values for the SVR using *param_grid* argument contained in *sklearn. model_selection. GridSearchCV*.

2.5. Radial Basis Function-Neural Network (RBF-NN)

The RBF-NN is a form of neural network in which the activation function in the hidden layers is the Gaussian function. The activation function in a neural network specifies how a weighted sum of input is converted into an output from nodes in a layer. The base or hidden layer's activation function takes the combined weighted inputs and transforms them using the Gaussian function. Fig. 5 shows a simple RBF-NN structure. The learning process, like that of any other neural network, is critical to RBF-NN performance. The purpose of this step is to fine-tune the network's parameters to reduce some error criteria. The three essential parameters of an RBF-NN with the basic architecture of a single hidden layer are connection weights, widths, and centers. A two-stage training process is the traditional way for training RBF-NN. The centers of the hidden layer and their widths are determined in the first step using an unsupervised clustering technique like k-means [46] or decision trees [47]. The weights between the hidden layer and the output layer are learned in the second stage. The outcomes are often computed linearly with the simple linear least squares (LS), orthogonal least squares (OLS), or gradient descent algorithms.

2.6. Performance metrics

Performance metrics are used to judge or measure how well a model performs on a set of given data. The metrics used will depend on whether it is a classification or regression task. Since this study deals with a regression model it will require metrics for regression. Scikit-learn contains a module called *metrics* for the purpose of evaluating models. The following metrics was used to evaluate and compare the performance of the Z-factor models.

i. Mean Squared Error (MSE):

The metric corresponds to the expected value (mean) of the squared error or loss. It is computed through the equation below.

$$MSE(y, y_{pred}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - y_{pred\ i})^2 \quad (1)$$

where.

$y_{pred\ i}$ is the predicted value of i-th sample,
y_i is the actual value
n_{sample} is the number of samples.

ii. Mean Absolute Error (MAE):

This metric computes the expected value of the absolute error or loss. It is calculated through the equation below.

$$MAE(y, y_{pred}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - y_{pred\ i}| \quad (2)$$

ii. R^2 Score (Coefficient of determination):

It indicates the model's goodness of fit, and hence a measure of how well the model is expected to predict unseen samples. The maximum score is 1.0 and can be negative if the model is very bad.

$$R^2(y, y_{pred}) = 1 - \frac{\sum_{i=1}^{n} (y_i - y_{pred})^2}{\sum_{i=1}^{n} (y_i - y_{ave})^2} \quad (3)$$

Where $y_{ave} = \frac{1}{n}\sum_{i=1}^{n} y_i$

3. Results and discussions

Different modeling methods and optimization algorithms will be used in this study. As there is a wealth of information in the literature about all the used methods and material they are not described in this study. In this section, the performance of new models was compared using different metrics of accuracy. The accuracy of each model is first tested, followed by a comparison between the models and the base case (Standing and Katz chart) and finally comparison of the best model with other correlations. The results obtained from the first two sections help us to understand the predictive power of machine learning models and select the best model which was compared with other relevant correlations.

3.1. Comparison between the actual and predicted values

The word 'actual' was used from time to time to refer to the basis of comparing the models. Fig. 6 shows the frequencies of the actual compressibility values. The histogram distribution shows that the most frequent values of Z-factor are within 1.5–1.75. The second most frequent Z-factor values are within 0.75–1.0. This shape is not specifically defined, but we can note regardless that it is bi-modal, having two separated classes or intervals equally representing the maximum

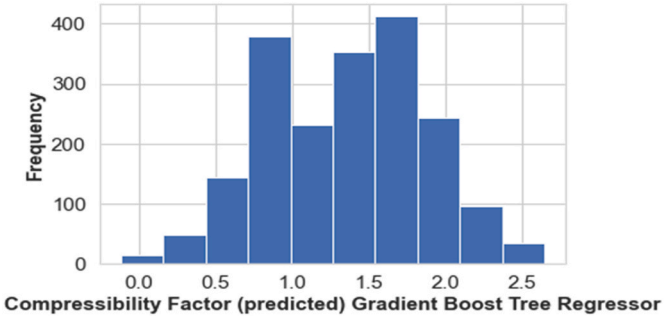

Fig. 7. Frequency Distribution of predicted Z-factor for GBDT.

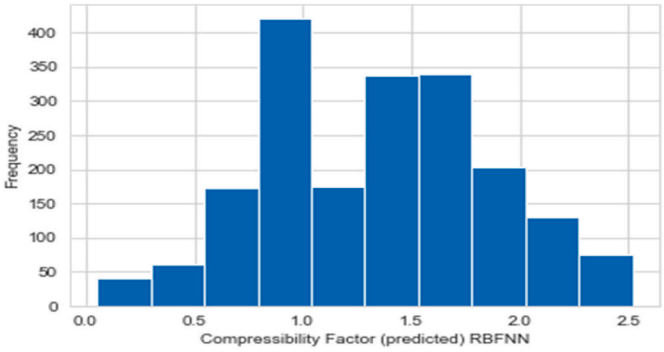

Fig. 8. Frequency Distribution of predicted Z-factor for RBF-NN.

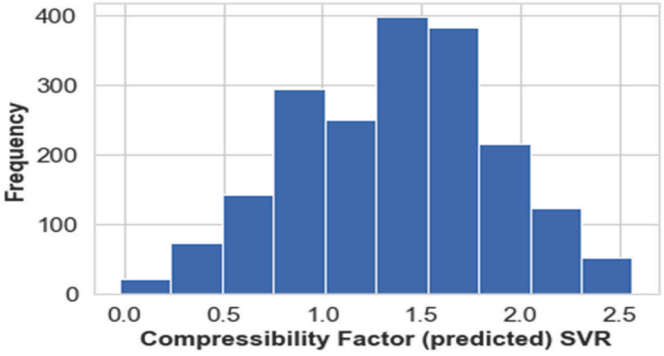

Fig. 9. Frequency Distribution of predicted Z-factor for SVR.

frequency of the distribution for the Z-factor.

Fig. 6 was used as the basis for comparing the three proposed models as shown in Figs. 7–9. The three proposed models predicted a bi-modal distribution of the Z-factor values. The distribution also shows a high variation which means that the Z-factor values are widely spread out about the center of a data set. The distribution of Z-factor predicted by the gradient boosted regression model in Fig. 7 shows a close similarity with that of Fig. 6 (the actual). The two most frequent range of Z-factor for the GBDT model is the same as the actual, and also a general look at the distribution symmetry reveals the closeness of the two distributions. Figs. 8 and 9 show the distribution obtained using the RBF-NN and SVR. Some of the bins in Figs. 8 and 9 were over-estimated when compared to the actual. The inaccuracy in properly estimating the Z-factor can be attributed to the fact that neural networks require much more data than the traditional machine learning algorithm [48].

3.2. Regression plot for the models

This section examines the characteristic departures of Z-factor

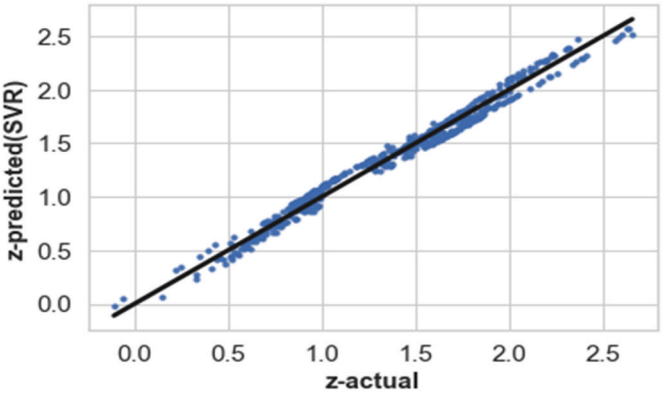

Fig. 10. Regression plot for SVR.

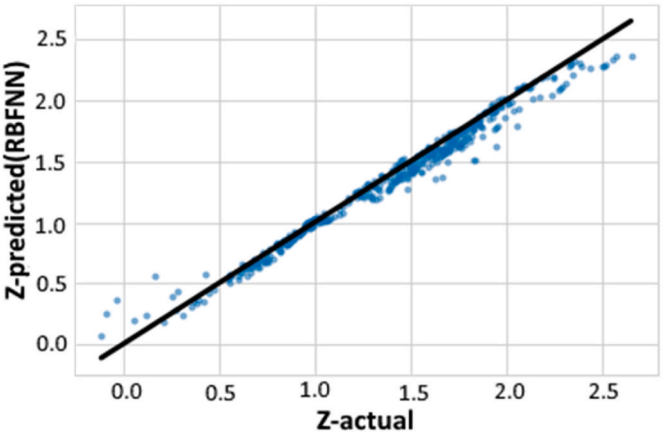

Fig. 11. Regression plot for RBFNN.

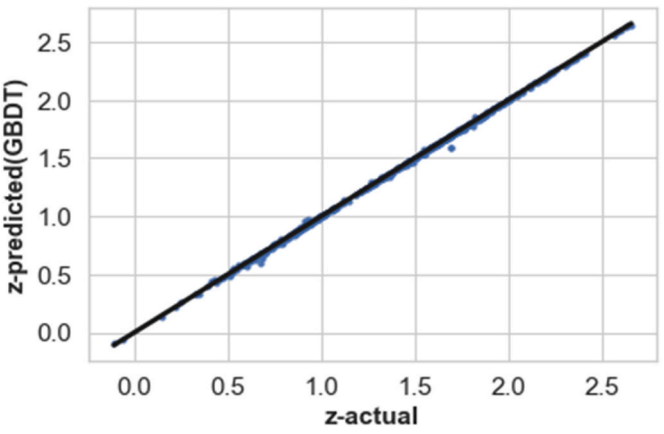

Fig. 12. Regression plot for GBDT.

predicted values from Normality. For the graphical analysis of prediction accuracy, the plots shows the cross-plot in which the vertical and horizontal axes depict the predicted and experimental data nodes. If the Z-factor data distribution approximates a sample from a normal distribution, the scatter plot will fall along a line from the bottom-left to the top-right of the plot. The interpretation is enhanced with a line of "expected Z-factor values" if the sample data points are drawn from a normal distribution. The closer the predicted Z-factor points are to the line, the more closely the points approximate the expectation from a normal distribution. Comparison of experimental data and the proposed models as the best predicting correlation for all data points is shown in

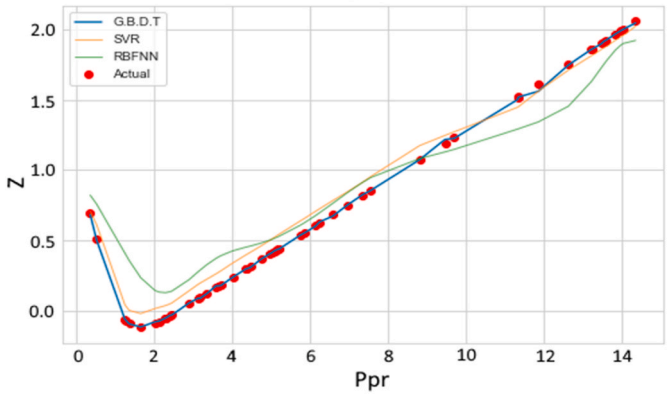

Fig. 13. Isotherm of 0.92 for the three proposed models.

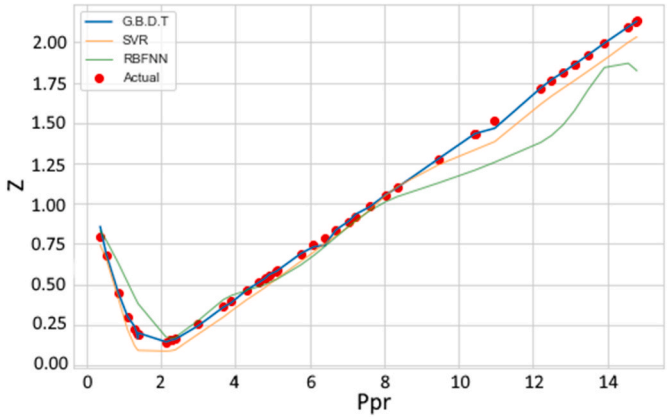

Fig. 14. Isotherm of 0.95 for the three proposed models.

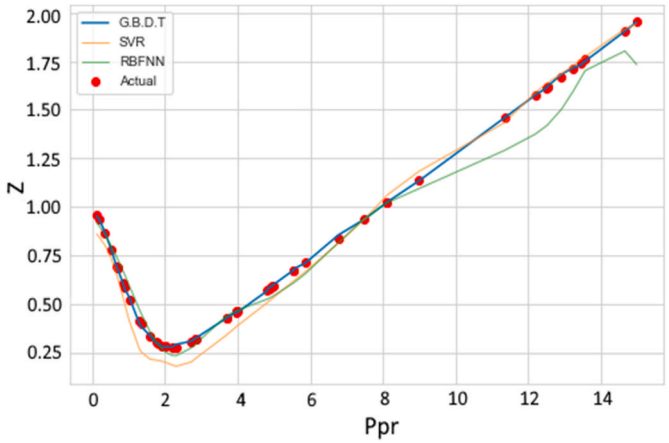

Fig. 15. Isotherm of 1.0 for the three proposed models.

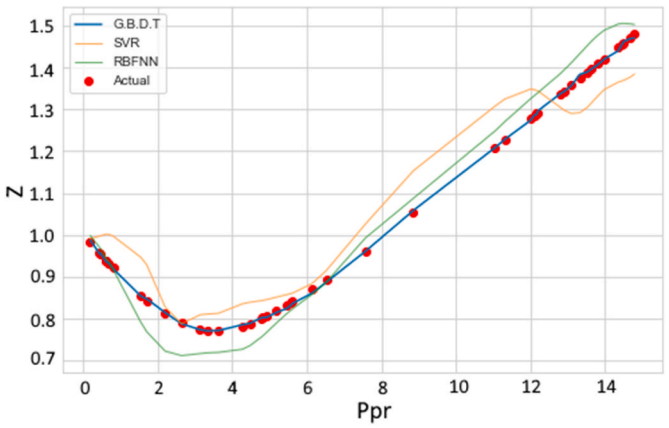

Fig. 16. Isotherm of 1.5 for the three proposed models.

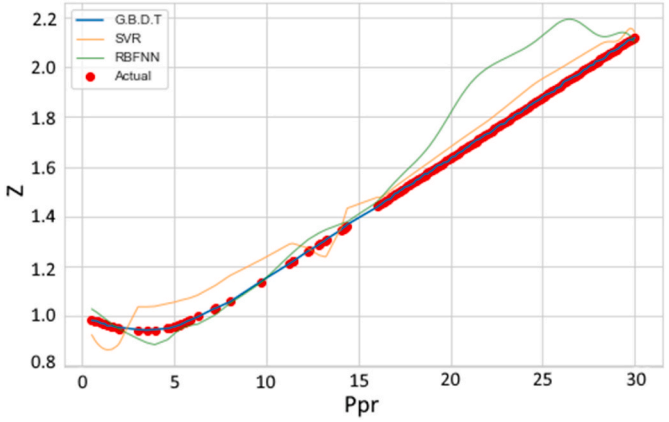

Fig. 17. Isotherm of 2.0 for the three proposed models.

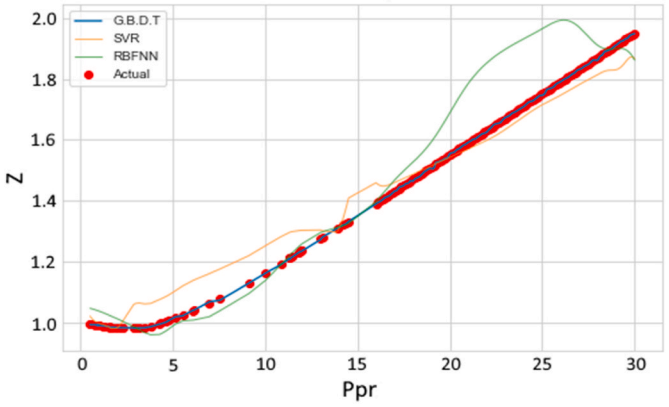

Fig. 18. Isotherm of 2.4 for the three proposed models.

Figs. 10–12, the best-fit line (black line) through the training data indicates the variance explained by each of the proposed models. Each model is a line of best fit minimizing the sum of the squared differences between each Z-factor data point and the line of best fit (45°). If the Z-factor data point falls along the roughly straight line at a 45°, then the Z-factor values are roughly normally distributed. According to the predicted data set, accuracy of models increases as the nodes become more concentrated. The evolutionary trend of improvement is clearly understandable by the comparison of the scattered diagrams model. Also, there are some poorly predicted nodes that approach their corresponding experimental values. The regression plot verifies the accuracy of the gradient boosted regression model over the other models. In Fig. 10, it is observed that the RBF-NN model had predicted values that deviated far from the actual at lower and the upper portion along the line of best fit. The SVR performs better than the RBFNN model in Fig. 11, given that the predicted Z-factor values are relatively closer to the line of best fit. The GBDT model outperformed the RBF-NN and SVR, with most Z-factor data points falling on the line of best fit, as shown in Fig. 12.

This approach has provided the best predictions compared with EoS-

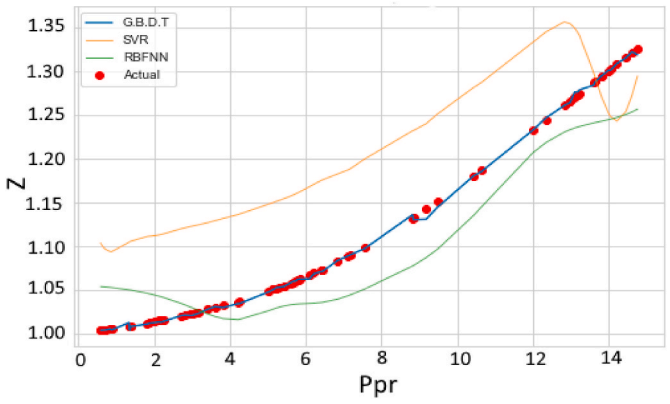

Fig. 19. Isotherm of 3.0 for the three proposed models.

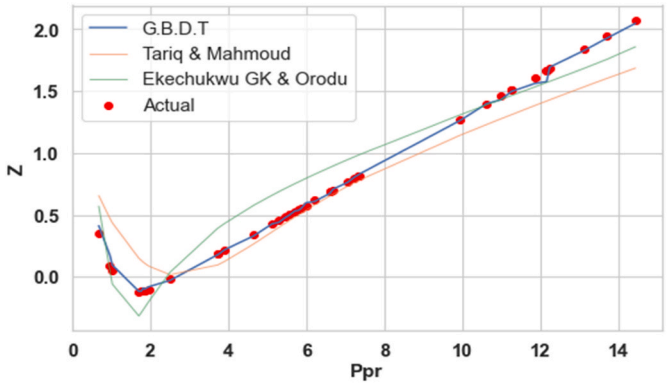

Fig. 20. Z-factor prediction for notable correlation at T_{pr} 0.92.

based methods [25] and correlations (Table 1). Correlations have provided better results than EoS-based models but are not as accurate as AI-based models.

3.3. Variation of Z-factor with pseudo reduced pressure at different pseudo reduced temperatures

The Z-factor variations with pseudo reduced pressures (P_{pr}'s) at different pseudo reduced temperatures (T_{pr}'s) are presented in Figs. 13–19. It is mostly necessary to investigate natural gas Z-factor with respect to P_{pr} and T_{pr} in petroleum engineering applications. In this study, the P_{pr} range was 0.92–3.0. A general trend was observed for the isotherm plots: (i.) for T_{pr}'s of 0.92, 0.95.1.0, and 1.5, the isotherms on the lower portion of the plot significantly deviate from the ideal concave up, increasing relationship; (ii.) for T_{pr}'s greater than 1.5, the typical concave up, increasing curve was observed.

As shown in Figs. 13–19, the GBRT most accurately fits the actual data points for all the Z-factor variation with P_{pr}'s and T_{pr}'s. The SVR follows the trend as well but tends to over-estimate the Z-factor for almost all P_{pr} points at T_{pr} of 0.92, 1.5, 2.0 and 2.4. The RBFNN has the largest deviation from the actual. For the T_{pr} at 1.0 and 1.5, the GBRT model to a good extent matched the actual curve with little errors. At higher T_{pr} values, the SVR model and RBFNN model were unable to make accurate predictions. For all the isotherms considered, the GBDT model had a great fit. The SVR model was better at lower isotherms than higher isotherms. Finally, the RBFNN had the worst fit progressing from the lower isotherms up to the highest isotherm considered.

The isotherm plot was compared with existing correlations in published literature. Tariq and Mahmoud [37], and Ekechukwu and Orodu [34] models were selected because they have wide range of applicability, thus, will render the proposed model somewhat valid if it is

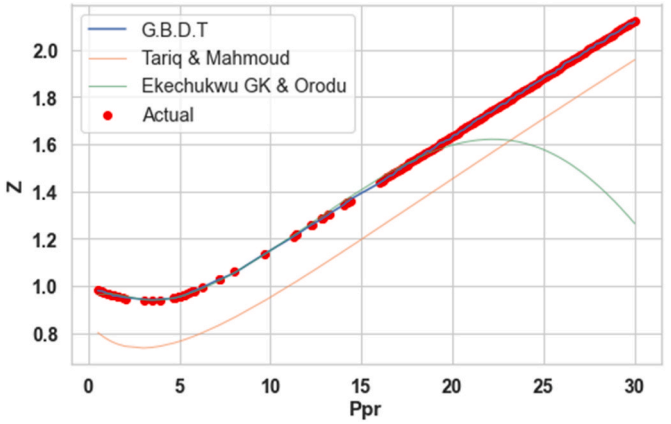

Fig. 21. Z-factor prediction for notable correlation at T_{pr} 2.0.

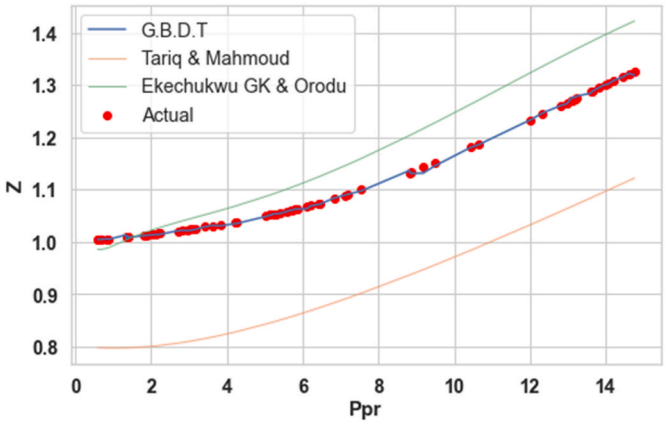

Fig. 22. Z-factor prediction for notable correlation at T_{pr} 3.0.

observed to perform better. Also, a general trend was observed for the isotherm plots: for lower T_{pr}'s, the isotherms on the lower portion of the plot significantly deviate from the ideal "concave up, increasing" relationship; at high T_{pr}'s, the typical "concave up, increasing" curve was observed. We observed that Ekechukwu and Orodu [34] model deviated from the "concave up, increasing" trend to "concave down, decreasing" for T_{pr} of 2.0 (Fig. 21). Ekechukwu and Orodu ([34] model performed well at lower P_{pr} values but, did not fit the actual data points with P_{pr} values beyond 16 (Fig. 21). Tariq and Mahmoud (2019) model matches the actual data points but, under-predicts the values.

At low T_{pr}'s, the GBDT model is able to match the data points of the actual more than the two correlations selected (Fig. 20). It is observed in Fig. 22 that even at high T_{pr}'s, the proposed GBDT model continues to match the data points accurately. However, the Ekechukwu and Orodu (2019) correlation tends to over-predict the Z-factor values, while the Tariq and Mahmoud [37] model still under-estimates the values of Z-factor. All model predictions showed a "concave up, increasing" relationship.

3.4. Model accuracy

3.4.1. SVR accuracy

In building the support vector regression (SVR) model, a grid search was used which allows for the combination of different hyperparameters so as to select the best combination in terms of some metric(s). The parameters of SVR considered were 'C', and 'gamma'. According to VanderPlas [49], they are the two most important parameters to consider when building an SVR model. The resulting model at the end of the training phase had 'C' and 'gamma' values of 20 and 1

Table 3
Table of accuracy for the SVR model.

Parameter	SVR Before	SVR After	Difference
Train R^2 score	0.94895	0.98222	0.03352
Test R^2 score	0.95836	0.98272	0.02406
MAE	0.07359	0.06019	−0.013
MSE	0.01127	0.00476	−0.00651
RMSE	0.10616	0.06903	−0.03718

Table 4
Table of accuracy for the GBDT model.

Parameters	GBDT Before	GBDT After	Difference
Train R^2 score	0.99787	0.99996	0.00212
Test R^2 score	0.99674	0.99962	0.00287
MAE	0.02161	0.00561	−0.01615
MSE	0.00088	0.00010	−0.00077
RMSE	0.02966	0.01033	−0.01933

respectively. SVR (C = 20, gamma = 1). Although, the SVR can be a very good estimator with the default hyper-parameters, tuning becomes important to achieve better results, thereby bringing the best out of the model. Table 3 shows the difference before and after optimization and the values showed an improvement after the optimization process.

3.4.2. GBDT accuracy

This model with default parameters performed better than the SVR model based on all the metrics used for evaluation. This shows how good this ensemble model was in developing the proposed model with the range of data points considered. A grid search was also done in building this model to get the best result. Optimization becomes important here to improve accuracy and avoid overfitting. Table 4 shows the effect of optimizing this model because the model tends to overfit easily and needs to be flexible due to the nature of the isotherms in the S–K chart. It shows that the GBRT gives the best correlation coefficient accuracy and minimum error.

3.4.3. RBF-NN accuracy

The RBF-NN model was built by manually adjusting some of the most influential parameters which are the optimizer and learning rate [50]. The optimizer for a neural network is responsible for reducing the loss function thereby increasing performance. The learning rate for an optimizer controls how fast the error loss is updated. The optimizer used was RMSprop and a learning rate of 0.01 was considered the best. The built RBF-NN model gave a training R^2 score, test R^2 score, MAE, MSE and RMSE of 0.97723, 0.97467, 0.05473, 0.00663 and 0.08144 respectively.

3.5. Comparison of the models prediction with other correlations

The results of Z-factor prediction at different P_{pr} and T_{pr} values, using correlations from previous studies and the best approach in this study (the GBDT model), are presented in Table 5. The selected correlations for comparison were Beggs and Brill [51], Orodu et al. [41], Sanjari and Lay [18], Ekechukwu and Orodu [34]. Orodu et al. [41] presented three correlations for Z-factor prediction. We however selected the two correlations with better performance according to the authors (Model 2 and Model 3). Sanjari and Lay [18] was selected because, it involved a machine learning approach. The benchmark for this comparison was the Standing and Katz chart. Considering that this chart has a minimum T_{pr} of 1.0, selected P_{pr}-T_{pr} pairs with $T_{pr} \geq 1.0$ were used. The Orodu et al. [41] correlations were selected because, they were developed for low T_{pr}'s as well. The GBDT model in general, outperformed the other correlations, for most of the selected pairs of T_{pr} and P_{pr}. Based on the points from Table 5, the mean absolute percentage error was calculated for each selected approach and presented in Fig. 23. As shown, it is clear that the proposed GBDT model outperformed the other correlations and therefore, verifies the authenticity of this model at these ranges of pseudo reduced temperatures and pressures.

4. Conclusion

Accurate determination of gas compressibility factor is critical in many aspects of petroleum engineering. It was observed that there is no single correlation or model that considered the low pseudo reduced temperature (i.e., T_{pr} <1). This study proposed a method of predicting z

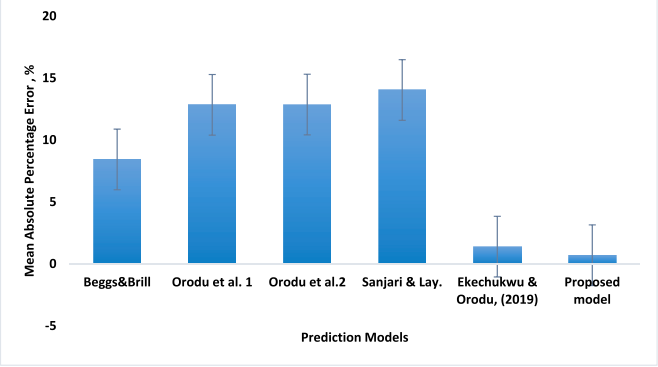

Fig. 23. Mean Absolute Percentage error for correlation.

Table 5
Z-factor correlation values at various P_{pr} and T_{pr}.

T_{pr}	P_{pr}	Standing and Katz	Beggs and Brill	Orodu et al. [41] (Model 2)	Orodu et al. [41] (Model 3)	Sanjari and Lay [18],	Ekechukwu & Orodu [34],	This Study (GBDT)
1.35	0.2	0.97	0.976	1.217	0.735	1.002	0.976	0.974
1	1	–	0.862	0.967	0.757	1.011	0.429	0.527
1.15	2	0.465	0.739	0.72	0.757	1.124	0.473	0.459
1.2	3	0.535	0.671	0.598	0.745	0.442	0.544	0.542
1.25	4	0.63	0.674	0.597	0.751	0.559	0.643	0.632
1.3	5	0.718	0.738	0.666	0.809	0.671	0.737	0.733
1.35	6	0.815	0.816	0.757	0.875	0.775	0.825	0.819
1.4	7	0.9	0.894	0.852	0.937	0.869	0.907	0.894
1.45	8	1	0.972	0.95	0.995	0.954	0.984	0.99
1.5	9	1.08	1.05	1.05	1.0496	1.031	1.057	1.065
1.6	10	1.135	1.127	1.151	1.102	1.103	1.127	1.137
1.7	11	1.2	1.203	1.254	1.153	1.164	1.191	1.203
1.8	12	1.25	1.28	1.358	1.201	1.218	1.251	1.244
1.9	13	1.3	1.356	1.464	1.248	1.268	1.307	1.299
2	14	1.34	1.432	1.571	1.293	1.315	1.358	1.344
2.2	15	1.36	1.508	1.679	1.336	1.358	1.402	1.356

factor for low and wide range of pseudo reduced temperature using machine learning approaches. Three models were developed using SVR, RBFNN and GBDT. The models were tested and the GBDT was found to have the highest R^2 score, lowest MSE and MAE of 0.9996, 0.00561 and 0.00010 respectively. To verify these claims, the models were compared to the actual Z-factor values, and the GBDT had the best match. In establishing the credibility of the proposed GBDT model, it was compared to four empirical correlations from literature. The mean absolute percentage errors (MAPEs) of the predictions made by each correlation were computed and plotted. The plot revealed that the GBDT model performed extremely well in predicting compressibility factor with an MAPE of about 1%. It can be concluded from this study that:

i. The proposed GBDT model, has been built to predict gas compressibility factor accurately within an extended range of pseudo-reduced temperature ($0.92 < T_{pr} < 3.0$).
ii. For the regression plot, the GBDT model performed the best given that almost all predicted Z-factor data points are within the line of best fit.
ii. A general trend was observed for the isotherm plots, for 0.92, 0.95.1.0, and 1.5 pseudo reduced temperatures, the isotherms on the lower portion of the plot significantly deviate from the ideal concave up, increasing relationship.
iv. At most of the selected pair of T_{pr} and P_{pr}, the GBDT model in general was seen to have predicted better than the other correlations.

For further studies, application of MLP type of ANN and advanced version of gradient boosting models such as extremely gradient boosting (XGboost), AdaBoost (Adaptive Boosting), CATBoost and Light GBM can be used in estimating Z-factor for natural gas.

Credit author statement

Emmanuel E. Okoro, Ekene Ikeora: Conceptualization, Methodology, Formal analysis, Writing – original draft, Supervision Emmanuel E. Okoro, Samuel E. Sanni, and Victor J. Aimihke: Data curation, Investigation, Project administration Samuel E. Sanni, Victor J. Aimihke Methodology, Writing – review & editing, Formal Oscar I. Ogali: Software, Methodology, Validation, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] I.I. Azubuike, S.S. Ikiensikimama, D.O. Orodu, Forecasting gas compressibility factor using artificial neural network tool for Niger-delta gas reservoir, in: Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition, NAIC, 2016, https://doi.org/10.2118/184382-ms.
[2] E. Heidaryan, J. Moghadasi, M. Rahimi, New correlations to predict natural gas viscosity and compressibility factor, J. Petrol. Sci. Eng. 73 (1–2) (2010) 67–72, https://doi.org/10.1016/j.petrol.2010.05.008.
[3] O. Al-Fatlawi, M.M. Hossain, J. Osborne, Determination of best possible correlation for gas compressibility factor to accurately predict the initial gas reserves in gas-hydrocarbon reservoirs, Int. J. Hydrogen Energy 42 (2017) 25492–25508, https://doi.org/10.1016/j.ijhydene.2017.08.030.
[4] V. Gaganis, D. Homouz, M. Maalouf, N. Khoury, K. Polychronopoulou, An efficient method to predict compressibility factor of natural gas streams, Energies 12 (2019) 1–20, https://doi.org/10.3390/en12132577.
[5] N. Chukuigwe, S. Ikiensikimama, I. Okafor, Digital transformation of the standing and Katz compressibility factor chart for natural gases, in: Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition, NAIC, 2020, https://doi.org/10.2118/203755-ms.
[6] E. Sanjari, E.N. Lay, An accurate empirical correlation for predicting natural gas compressibility factors, J. Nat. Gas Chem. 21 (2012) 184–188, https://doi.org/10.1016/S1003-9953(11)60352-6.
[7] M. Mohamadi-Baghmolaei, R. Azin, S. Osfouri, R. Mohamadi-Baghmolaei, Z. Zarei, Prediction of gas compressibility factor using intelligent models, Nat. Gas. Ind. B 2 (4) (2015) 283–294, https://doi.org/10.1016/j.ngib.2015.09.001.
[8] A. Khosravi, L. Machado, R.O. Nunes, Estimation of density and compressibility factor of natural gas using artificial intelligence approach, J. Petrol. Sci. Eng. 168 (2018) 201–216, https://doi.org/10.1016/j.petrol.2018.05.023.
[9] M. Farzaneh-Gord, H. Reza, B. Mohseni-Gharesafa, A. Toikka, I. Zvereva, Accurate determination of natural gas compressibility factor by measuring temperature, pressure and joule-thomson coefficient: artificial neural network approach, J. Petrol. Sci. Eng. (2021) 108427, https://doi.org/10.1016/j.petrol.2021.108427.
[10] S.J. Evans, How digital engineering and cross-industry knowledge transfer is reducing project execution risks in oil and gas, in: Society of Petroleum Engineers – Offshore Technology Conference, 2019, https://doi.org/10.4043/29458-ms.
[11] D. Koroteev, Z. Tekic, Artificial intelligence in oil and gas upstream: trends, challenges, and scenarios for the future, Energy and AI 3 (2021), 100041, https://doi.org/10.1016/j.egyai.2020.100041.
[12] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, H. Oza, Application of machine learning and artificial intelligence in oil and gas industry, Petroleum Research 6 (4) (2021) 379–391.
[13] N. Azizi, R. Behbahani, M.A. Isazadeh, An efficient correlation for calculating compressibility factor of natural gases, J. Nat. Gas Chem. 19 (6) (2010) 642–645, https://doi.org/10.1016/S1003-9953(09)60081-5.
[14] O.O. Festus, S.S. Ikiensikimama, Evaluation of compressibility factor correlations for Niger delta gas reservoirs, in: Society of Petroleum Engineers - Nigeria Annual International Conference and Exhibition, NAIC, 2010, pp. 161–169, https://doi.org/10.2118/136967-ms.
[15] M. Kamyab, J.H.B. Sampaio, F. Qanbari, A.W. Eustes, Using artificial neural networks to estimate the Z–factor for natural hydrocarbon gases, J. Petrol. Sci. Eng. 73 (3–4) (2010) 248–257, https://doi.org/10.1016/j.petrol.2010.07.006.
[16] B.D. Al-Anazi, G.R. Pazuki, M. Nikookar, A.F. Al-Anazi, The prediction of the compressibility factor of sour and natural gas by an artificial neural network system, Petrol. Sci. Technol. 29 (4) (2011) 37–41, https://doi.org/10.1080/10916460903330080.
[17] M. Baniasadi, A. Mohebbi, M. Baniasadi, A new correlation based on artificial neural networks for predicting the natural gas compressibility factor, J. Eng. Thermophys. 21 (4) (2012) 248–258, https://doi.org/10.1134/S1810232812040030.
[18] E. Sanjari, E.N. Lay, Estimation of natural gas compressibility factors using artificial neural network approach, J. Nat. Gas Sci. Eng. 9 (2012) 220–226, https://doi.org/10.1016/j.jngse.2012.07.002.
[19] E.M. Shokir, M.N. El-Awad, A.A. Al-Quraishi, O.A. Al-Mahdy, Compressibility factor model of sweet, sour, and condensate gases using genetic programming, Chem. Eng. Res. Des. 90 (6) (2012) 785–792, https://doi.org/10.1016/j.cherd.2011.10.006.
[20] A. Chamkalani, A. Mae'soumi, A. Sameni, An intelligent approach for optimal prediction of gas deviation factor using particle swarm optimization and genetic algorithm, J. Nat. Gas Sci. Eng. 14 (2013) 132–143, https://doi.org/10.1016/j.jngse.2013.06.002.
[21] A. Kamari, A. Hemmati-Sarapardeh, S. Mirabbasi, Prediction of sour gas compressibility factor using an intelligent approach, Fuel Process. Technol. 116 (2013) 209–216, https://doi.org/10.1016/j.fuproc.2013.06.004.
[22] H. Fatoorehchi, H. Abolghasemi, R. Rach, An accurate explicit form of the hankinson-thomas-phillips correlation for prediction of the natural gas compressibility factor, J. Petrol. Sci. Eng. 117 (2014) 46–53, https://doi.org/10.1016/j.petrol.2014.03.004.
[23] A. Fayazi, M. Arabloo, A.H. Mohammadi, Efficient estimation of natural gas compressibility factor using a rigorous method, J. Nat. Gas Sci. Eng. 16 (2014) 8–17, https://doi.org/10.1016/j.jngse.2013.10.004.
[24] M.M. Ghiasi, A. Shahdi, P. Barati, M. Arabloo, Robust modeling for efficient estimation of compressibility factor in retrograde gas condensate systems, Ind. Eng. Chem. Res. 53 (2014) 12872–12887, https://doi.org/10.1021/ie404269b.
[25] C. Li, Y. Peng, J. Dong, Prediction of compressibility factor for gas condensate under a wide range of pressure conditions based on a three-parameter cubic equation of state, J. Nat. Gas Sci. Eng. 20 (2014) 380–395, https://doi.org/10.1016/j.jngse.2014.07.021.
[26] M.A. Mahmoud, Development of a new correlation of gas compressibility factor (Z-factor) for high pressure gas reservoirs, J. Energy Resour. Technol. 136 (2014), 012903, https://doi.org/10.1115/1.4025019.
[27] A. Sarrafi, A.E.F. Monfared, E.G. Ravandi, R. Pouramiri, Using fuzzy logic for the accurate determination of the compressibility factor of hydrocarbon gases, Energy Sources, Part A Recovery, Util. Environ. Eff. 37 (20) (2015) 2231–2239, https://doi.org/10.1080/15567036.2012.676706.
[28] M. Shateri, S. Ghorbani, A. Hemmati-Sarapardeh, Application of wilcoxon generalized radial basis function network for prediction of natural gas compressibility factor, J. Taiwan Inst. Chem. Eng. 50 (2015) 131–141, https://doi.org/10.1016/j.jtice.2014.12.011.
[29] I.I. Azubuike, S.S. Ikiensikimama, O.D. Orodu, Natural gas compressibility factor measurement and evaluation for high pressure high temperature gas reservoirs, Int. J. Sci. Eng. Res. 7 (7) (2016) 1173–1181.

[30] A. Kamari, F. Gharagheizi, A.H. Mohammadi, D. Ramjugernath, A corresponding states-based method for the estimation of natural gas compressibility factors, J. Mol. Liq. 216 (2016) 25–34, https://doi.org/10.1016/j.molliq.2015.12.103.
[31] N. Azizi, M. Rezakazemi, M.M. Zarei, An intelligent approach to predict gas compressibility factor using neural network model, Neural Comput. Appl. 31 (2017) 55–64, https://doi.org/10.1007/s00521-017-2979-7.
[32] E.E. Okoro, D. Honfre, K.C. Igwilo, A. Mamudu, Measurement of the best Z-factor correlation using gas well inflow performance data in Niger-delta, Int. J. Appl. Eng. Res. 12 (12) (2017) 3507–3522.
[33] A. Salem, A. Elgibaly, M. Attia, A. Abdulraheem, Comparing 5-different artificial intelligence techniques to predict Z-factor, in: Society of Petroleum Engineers - SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition 2018, SATS, 2018, pp. 23–26, https://doi.org/10.2118/192354-ms.
[34] G.K. Ekechukwu, O.D. Orodu, Novel mathematical correlation for accurate avoids using hybrid models of gas compressibility factor, Nat. Gas. Ind. B 6 (6) (2019) 629–638, https://doi.org/10.1016/j.ngib.2019.09.001.
[35] M. Maalouf, N. Khoury, D. Homouz, K. Polychronopoulou, Accurate Prediction of Gas Compressibility Factor using Kernel Ridge Regression. 2019 Fourth International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), 1–4. https://doi.org/10.1109/actea.2019.8851106.
[36] H.M. Sidrouhou, M. Korichi, S. Dada, Evaluation of correlations of compressibility factor (z) of natural gas for Algerian gas reservoirs, Energy Proc. 157 (2019) 655–669, https://doi.org/10.1016/j.egypro.2018.11.231.
[37] Z. Tariq, M. Mahmoud, New correlation for the gas deviation factor for high-temperature and high-pressure gas reservoirs using neural networks, Energy Fuel. 33 (3) (2019) 2426–2436, https://doi.org/10.1021/acs.energyfuels.9b00171.
[38] L. Lin, S. Li, S. Sun, Y. Yuan, M. Yang, A novel efficient model for gas compressibility factor based on GMDH network, Flow Meas. Instrum. 71 (2020) 101677, https://doi.org/10.1016/j.flowmeasinst.2019.101677.
[39] J. Ogbunike, T. Adeyemi, Development of a novel compressibility factor correlation for high pressure - high temperature HPHT reservoirs using stochastic and robust optimization approach, in: Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition 2020, NAIC, 2020, https://doi.org/10.2118/203619-ms.
[40] Y. Wang, J. Ye, S. Wu, An accurate correlation for calculating natural gas compressibility factors under a wide range of pressure conditions, Energy Rep. 8 (2) (2022) 130–137, https://doi.org/10.1016/j.egyr.2021.11.029.
[41] K.B. Orodu, E.E. Okoro, O.K. Ijalaye, O.D. Orodu, Gas compressibility factor explicit correlations for range of pseudo reduced temperature and pressure, Flow Meas. Instrum. 67 (2019) 176–185, https://doi.org/10.1016/j.flowmeasinst.2019.05.003.
[42] H. Liu, Y. Wu, P. Guo, Z. Liu, Z. Wang, S. Chen, B. Wang, Z. Huang, Compressibility Factor Measurement and Simulation of Five High-Temperature Ultra-high-pressure Dry and Wet Gases, vol. 500, Fluid Phase Equilibria, 2019, 112256, https://doi.org/10.1016/j.fluid.2019.112256.
[43] D.A. Otchere, T.O.A. Ganat, V. Nta, E.T. Brantson, T. Sharma, Data analytics and Bayesian Optimised Extreme Gradient Boosting approach to estimate cut-offs from wireline logs for net reservoir and pay classification, Appl. Soft Comput. 120 (2022), 108680, https://doi.org/10.1016/j.asoc.2022.108680.
[44] G.I. Diaz, A. Fokoue-Nkoutche, G. Nannicini, H. Samulowitz, An effective algorithm for hyperparameter optimization of neural networks, IBM J. Res. Dev. 61 (4) (2017) 1–11, https://doi.org/10.1147/JRD.2017.2709578.
[45] T.H. Ahmed, G.V. Cady, A.L. Story, A generalized correlation for characterizing the hydrocarbon heavy fractions, in: Society of Petroleum Engineers - SPE Annual Technical Conference and Exhibition, ATCE, 1985, https://doi.org/10.2118/14266-ms.
[46] J.K. Sing, D.K. Basu, M. Nasipuri, M., Kundu, Improved K–means Algorithm in the Design of RBF Neural Networks. IEEE TENCON 2003 Conference on Convergent Technologies for Asia-Pacific Region. Pp. 841-845. https://doi.org/10.1109/TENCON.2003.1273297.
[47] M. Kubat, Decision trees can initialize radial-basis function networks, IEEE Trans. Neural Network. 9 (5) (1998) 813–821, https://doi.org/10.1109/72.712154.
[48] A. Geron, Hands-on Machine Learning with Scikit-Learn, Keras and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2019, ISBN 978-1492032649.
[49] J. VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media, Inc., 2017, ISBN 978-1-491-91205-8.
[50] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, ISBN 9780262035613.
[51] M. Golan, C.H. Whitson, Concepts of Well Performance Engineering (Chp. 1). Well Performance, second ed., Prentice Hall Inc., Norway, 1991, pp. 1–108.