Arabian Journal of Geosciences (2021) 14: 284
https://doi.org/10.1007/s12517-021-06661-y

ORIGINAL PAPER

Accurate artificial intelligence-based methods in predicting bottom-hole pressure in multiphase flow wells, a comparison approach

Mohammad Zolfagharroshan[1] · Ehsan Khamehchi[1]

Received: 21 November 2020 / Accepted: 29 January 2021 / Published online: 10 February 2021
© Saudi Society for Geosciences 2021

Abstract

Determining bottom-hole pressure in producing multiphase flow petroleum wells is a crucial issue that directly affects plans, infrastructures, and all the equipment required to develop an oil field. This study attempted to introduce the best method regarding accuracy and codes applicable in vertical oil and gas wells. Therefore, with the aid of genetic programming (GP), least-square support vector machines (LSSVM), and radial basis function (RBF) neural networks and using 450 producing wells data, three models are trained. Besides, five experimental and mechanistic correlations were used for taking more approaches into account to evaluate the accuracy of simulations. LSSVM and GP had the highest precision models among all the methods, proved by several analyses and calculation of the coefficient of determination (R^2), average relative error (ARE), average absolute relative error (AARE), and root mean square error (RMSE). The models mentioned above performed better than previously developed correlations, where Ansari et al.'s (1994) model had the nearest outcomes to GP results. In addition to the generated models, the analyses' results introduced a new correlation for predicting wellbore pressure using available measurements of the wellhead data.

Keywords Pressure gradient · Two-phase flow · Artificial intelligence

Introduction

Estimating the pressure gradient in multiphase flow oil and gas wells is more sophisticated than single-phase flow wells, avoiding direct measurements, e.g., using downhole gauges. The pipe's flowing fluid pressure gradient is a critical affecting parameter on completion and designing other infrastructures. Pressure drop through the pipelines depends on several parameters, including total mass flow rate, well geometry, fluid properties, thermodynamic parameters consisting of the temperature gradient, and bottom-hole pressure (Shoham 2006).

There are numerous models presented to model multiphase flow in well/pipeline like the studies of Beggs and Brill (1973), Chierici et al. (1974), Asheim (1986), Hasan and Kabir (1992), Petalas and Aziz (2000). These works developed gradually and are based on experimental and analytical solutions and are still used extensively in simulating applications. Mechanistic models are usually more complicated in describing physical processes while offering more advantages than previous experimental models, such as high accuracy predictions and more realistic hypotheses.

Over the past few years, applying artificial intelligence (AI) methods on big data sets helped scientists find meaningful correlations justifying different physical phenomena. Using computer tools like artificial neural networks (ANN) and genetic programming (GP) has led to developing models and recognizing patterns explaining various physical problems.

AI is used broadly in engineering, and science is. In petroleum engineering, we can address some of the previous works in various sections such as exploration and reservoir characteristics determination (Gharbi and Mansoori 2005; Ahmadi et al. 2014), drilling rate estimation (Naderi and Khamehchi 2018), and multiphase flow rate prediction in oil and gas wells (Khamehchi et al. 2020). Those are various examples of applying AI methods in the oil and gas industry, proposing better physical justification approaches and the least error compared with previous studies. For instance, in the Naderi and Khamehchi (2018) article, the authors proposed a robust

Responsible editor: Santanu Banerjee

✉ Ehsan Khamehchi
Khamehchi@aut.ac.ir

[1] Department of Petroleum Engineering, Amirkabir University of Technology, Hafez Ave, Tehran, Iran

method for minimizing drilling costs as a function of operational parameters such as depth, weight on bit, rotation speed, and standpipe pressure flow rate, mud weight, and bit rotational hours. The primary plus side of this study was the improvement of the accuracy with regard to error. This study has a similar theme to the study mentioned above.

Considering previous studies, still finding novel methods that have merits like calculating more precisely and shorter run-times, as the proposed genetic programming method in this study is desirable.

The main aim of this study is the estimation of bottom-hole pressure in two-phase flow oil- and gas-producing wells by applying three AI tools, namely, radial basis function (RBF) neural networks, genetic programming (GP), and least square support vector machines (LSSVM). These names are general terms, and each model has to be tuned for a specific problem to be developed. After optimizing the parameters and features of these tools for determining BHP, three models are constructed.

As extra work, some formerly presented correlations are also used alongside real and simulated data. One of the main features of this study is using field data sets recorded in Iran's production sites, in which all the data are recorded by downhole pressure gauges. Several statistical parameters will be calculated, consisting of the coefficient of determination (R^2), average relative error (ARE), average absolute relative error (AARE), and root mean square error (RMSE) to evaluate the results.

Pressure gradient calculation

Differences between density and shear stress of flowing fluid in pipelines/wells form a two-phase flow. Therefore, in thermal systems, oil and gas or water and vapor tend to move separately. In the producing wells without applying artificial lift, the fluid's pressure declines toward the surface, and the fluid velocity increases due to the expansion of fluid volume. Therefore, because of the extraction of light components from the liquid phase, depending on several factors such as buoyancy force, inertia, and interfacial tension, there might form diverse flow patterns in a two-phase flow pipeline (Zolfagharroshan and Khamehchi 2020).

Considering mass and momentum conservation equations for a control volume and solving it for a single-phase fluid, obtain the pressure gradient's general equation. The equation for two-phase flow is derived similarly, and the total pressure gradient of two-phase streams would be the sum of frictional, hydrostatic, and kinetic terms as the right-hand side of Eq. (1).

$$\left(\frac{\partial \mathrm{P}}{\partial \mathrm{L}}\right)_t = \left(\frac{\partial \mathrm{P}}{\partial \mathrm{L}}\right)_f + \left(\frac{\partial \mathrm{P}}{\partial \mathrm{L}}\right)_{el} + \left(\frac{\partial \mathrm{P}}{\partial \mathrm{L}}\right)_{acc} \quad (1)$$

where the kinetic term is generally insignificant, so it is neglected in this study (Watson 2016). As the hydrostatic component is proportional to the mixture's average density, it is the vertical flow's dominant term.

There are many models to describe the pressure gradient of the two-phase flow. Earlier methods like Poettman and Carpenter (1952), Baxendell and Thomas (1961), and Fancher Jr and Brown (1962) did not consider two critical factors in the particular liquid holdup and flow patterns. Later on, other models have taken slippage between phases into account, like Hagedorn and Brown (1965). Lastly, the approaches developed in which the mentioned factors were considered in calculations were more robust than previous models. In this way, Ansari et al. (1994); Hasan and Kabir; and Petalas and Aziz models are well known and used in practical use.

Various pressure gradient methods were selected to be used in this study, and regarding assumptions and applications, five methods are chosen to be used in this simulation. The models are Hagedorn and Brown, Duns Jr and Ros (1963), modified Payne et al. (1979), Mukherjee and Brill (1985), and Ansari et al. (1994); except for Hagedorn and Brown, other methods are mainly different in liquid holdup computation and flow pattern recognition.

Artificial intelligence methods

Radial basis function networks

The artificial neural network (ANN) developed regarding biological neurons is one of the most powerful machine learning techniques. It is used extensively in different science and engineering areas, developed and known as one of the fastest and more accurate methods.

In ANN, in general, the weights are multiplied to inputs and summed with a bias term. The output is fed to a transfer (activation) function, which the type of function is selected by the user. The weights and biases are determined at last and reported.

RBFs are a type of neural network developed based on interpolation theories, which use a radial basis function as their middle layer's activation function. This network could approximate a more vast range of tasks regarding a better interpretation of their parameters than MLP networks (Jahanandish et al. 2011).

The output network is a linear combination of the neurons parameters and the RBF inputs value. Using RBF, prediction of time series, function approximation, classification, clustering, and system controls would be possible. The Schematic presented in Fig. 1 shows a general structure of this type of neural network, and like other types, it consists of an input layer, a hidden layer, and the output layer. The hidden layer has a nonlinear activation function that maps n-dimensional inputs data into m-dimensional data (which m is usually

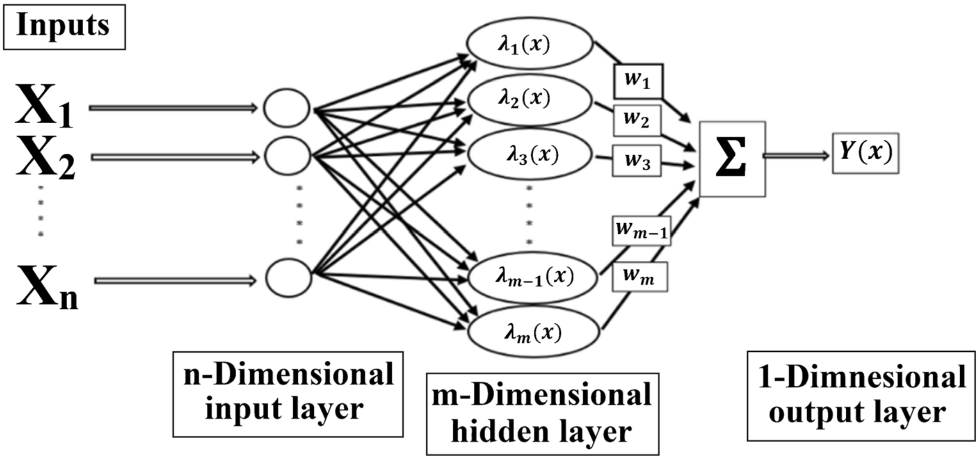

Fig. 1 Schematics of a simple RBF neural network (Malallah and Nashawi 2005)

higher than n) for better classification purposes (Malallah and Nashawi 2005).

The network's output is a scalar function of the input vector given by Eq. (2).

$$Y(x) = \sum_{i=1}^{m} \omega_i \lambda(||x - x_c||) \quad (2)$$

where m is the number of neurons in the hidden layer, x_c is the center vector of the neuron c, ω_i is the ith neuron's output weight, which can be positive or negative.

Functions that depend only on the distance from a center point are radially symmetric about this point. λ is the radial basis function and is a function of the norm value (Euclidean distance), which is denoted by || ||. In this study, a Gaussian type of RBF is used, which is the most common Kernel function in use and shown in Eq. (3). In this equation, σ is the spread parameter and has to be selected in an optimum manner.

$$\lambda(r) = \exp\left(\frac{-r^2}{2\sigma^2}\right), \quad r = ||x - x_c|| \quad (3)$$

In summary, each set of inputs ($[X_1, X_2, \ldots, X_n]$) entered into the hidden layer, and outputs from this layer (λ_i) multiplied by the specific weight (W_i). The result ($[Y_1, Y_2, \ldots, Y_m]$) is the weighted sum of the values calculated by the RBF networks, as the following equation.

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} & \ldots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} & \ldots & \lambda_{2m} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \lambda_{m1} & \lambda_{m2} & \lambda_{m3} & \ldots & \lambda_{mm} \end{bmatrix} \times \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ . \\ . \\ W_m \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ Y_N \end{bmatrix} \quad (4)$$

The mean square error (MSE) was used to evaluate the RBF performance. MSE is a risk function to measure the average squared differences between the estimated and the measured values. MSE defines as Eq. (5).

$$MSE = \sum_{i=1}^{N} \frac{\left(Y_i^{Measured} - Y_i^{Estimated}\right)^2}{N} \quad (5)$$

where N is the number of the data points, $Y_i^{Measured}$ and $Y_i^{Estimated}$ are measured, and estimated target values, respectively.

Before building an RBF model, all tuning parameters have to be optimized, which takes some time to reach these values by running the program several times (Jahanandish et al. 2011).

In this study, we applied the gradient descent algorithm, which adapts weights, centers, and deviations—the testing set of data used to assess the model's accuracy to predict unseen data. For the current study, 70 and 20% of the production data set were assigned for training and testing data, respectively, and a 10% for validation.

Genetic programming

Genetic programming (GP) is one of the latest evolutionary algorithms, presented in 1990 by Koza, one of the most robust machine learning tools. GP is as powerful as it can be compared with ANN, and one of the most critical features of this tool is generating models instead of producing answers.

In this technique, at first, the populations of chromosomes were produced randomly. They represent the functions which were applied to the independent data. It has to be noted that GP usually needs a larger population size compared with typical algorithms; for instance, if other algorithms require 10 to 100 population size, in GP, this quantity would be in the order of thousands to millions.

GP considers every individual population as a tree; the tree structure is known as a gene. Simply, this tree can be a mathematical or logical formula, or even a computer program, which produces an output, which the eligibility can be analyzed.

Trees can be quickly evaluated recursively. Figure 2 shows a simple schematic of the tree (gene) structure (Koza et al. 2006).

If we assume that the two independent parameters are Z and M and a hypothetical mathematical expression is $(5 \times M) + (Z - 8)$, this expression is called a gene. It is a simple tree. When the algorithm progresses, more genes are added, and the model's capability to estimate the model improves.

There are two types of unintentional changes in genetics that might happen to a newborn child in nature: crossover and mutation. These mechanisms that nature tries to prevent them from happening as a creature's features are accidental and not from the parents. Still, it can accelerate the mechanism of evolution. So these genetic operators are also used in some evolutionary algorithms, including GP.

As the other evolutionary methods, reproduction exists in GP, so individuals are chosen as parents from the current generation to form the next generation. Tournament selection is the most commonly used selection method, which involves running several “tournaments” among a few individuals.

Then, each tournament winner is selected for crossover. The crossover allows new individuals to be created, and the mutation is an asexual operation that operates on only one individual. Note that the mutation operation affects increasing the population's genetic diversity by creating new individuals (Coello et al. 2007). The algorithm will be stopped if the last population of genes can predict the response, with minimum error. Then, the response model will be the weighted summation of all genes with a bias term.

In this study, a free open MATLAB-based software platform called GPTIPS (the Genetic Programming Toolbox for Identification of Physical Systems) was used to establish a symbolic nonlinear correlation for bottom-hole pressure, developed by Searson et al. (2010) for multigene regression application. The general form of the symbolic regression model is as the following equation.

$$\widehat{Y} = f(x_1, x_2, x_3 \ldots x_n) \tag{6}$$

where $\widehat{Y}$ is the model predictor of Y, which is desirable to predict, and X_1, $X_2 \ldots$, X_n are independent variables, and f is the sum of nonlinear regression functions.

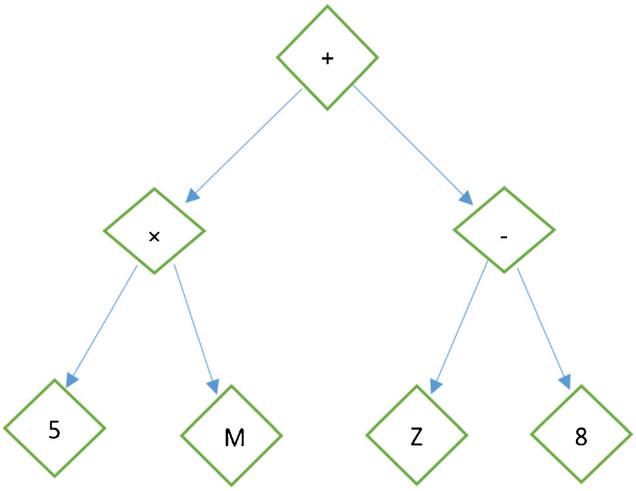

Fig. 2 Schematic of a tree (gene) structure representing the model term (5 × M) + (Z − 8) (Koza et al. 2006)

In GPTIPS, some adjustable parameters must be specified by the users. These parameters are population number, the maximum number of genes, the depth of genes, the maximum number of iterations, and essential mathematical operators and functions. These parameters should be adjusted, so there is a need to carry out several runs to attain the appropriate values.

The algorithm continues the trial and error until the last gene population predicts that the response meets the least error value.

Least square support vector machines

Support vector machines (SVM) are a type of statistical machine learning method applicable in clustering and regression problems. This tool is extensively used in classification and pattern recognition.

Input parameters are transferred into a high-dimensional space using kernel functions; then, the SVM algorithm makes a hyperplane in the new space. Finally, the data is separated from the possible highest distance, which is called an insensitive zone.

In other words, by using nonlinear transfer functions, SVM maps m-dimensional input data into the n-dimensional feature space. To categorize data with the maximum possible gap, SVM constructs a hyperplane in the feature space and finds the solution by solving a quadratic problem.

This study used the least square support vector machine (LSSVM), which finds a desirable solution by solving a linear system of equations. These machines are a modified version of SVM in a more straightforward form to create a model.

To classify a set of binary sample data as [(x_1, y_1), (x_2, y_2), (x_3, y_3)… (x_n, y_n)], same as Vapnik's formula, if the following condition in Eq. (7) is satisfied, this set of data is linearly separable in the feature space.

$$y_i\left(w^T\phi(x_i) + b\right) \geq 1 \quad , \quad i = 1, 2, 3\ldots, N \tag{7}$$

where y_i is 1 or −1, w^T is the weight vector transposed, $\phi(x_i)$ is the transfer function with a nonlinear form to maps input data into a high-dimensional feature space, and b is a bias term.

The LSSVM is formulated as an optimization problem, which is defined as the following equation.

$$Min\ F = \frac{1}{2}\, w^T w + \frac{C}{2} \sum_{i=1}^{N} \xi_i^2 \tag{8}$$

Subjected to

$$y_i\left(w^T\phi(x_i) + b\right) = 1 - \xi_i \quad , \quad i = 1, 2, 3\ldots, N \tag{9}$$

where C and ξ are penalty factors and the estimation error, respectively. This optimization problem can be solved considering the Lagrange function and Karush–Kuhn–Tucker

Table 1 Range of production data

Variable	Symbol	Minimum	Maximum	Standard deviation
Wellhead pressure (*psi*)	P_{wh}	30	3343	949.34
Oil flow rate ($\frac{bbl}{day}$)	Q_o	0	13,006.56	2760.75
Gas flow rate ($\frac{scf}{day}$)	Q_g	0	12,876.49	2181.82
Water flow rate ($\frac{bbl}{day}$)	Q_w	0	13,741	2503.11
Tubing inside diameter (inch)	D_i	1.49	6.375	1.802
Oil API gravity	API	11.2	54	10.21
Gas specific gravity	γ_g	0.6	1.4	0.16
Wellbore depth (ft)	D	918	13,125	3871.45
Surface temperature (°F)	T_s	70	286	56.12
Bottom-hole temperature (°F)	T_d	80	292	67.12

Table 2 A limited part of used production data presented

Wellhead pressure, P_{wh} (psi)	Oil flow rate, Q_o $\left(\frac{\text{bbl}}{\text{day}}\right)$	Gas flow rate, Q_g $\left(\frac{\text{scf}}{\text{day}}\right)$	Water flow rate, Q_w $\left(\frac{\text{bbl}}{\text{day}}\right)$	Tubing inside diameter, D_i (inch)	Oil gravity (API)	Gas specific gravity, γ_g	Wellbore depth, D (ft)	Surface temperature, T_s (°F)	Bottom-hole temperature, T_d (°F)	Measured bottom-hole pressure, P_{bh} (psi)
1840	95.41	50.66	107.59	6.375	51	0.65	13,125	142	292	6543.3
2745	4877.6	4019.1	1219.4	2.994	32.1	1.02	5524	167	202	4734.3
2546	693.36	583.12	7010.6	3.5	28	0.81	2767	228	236	3702.8
250	124	23.2	0	1.995	23	0.6	2900	80	121	1153
3343	441.1	211.73	3969.9	2.992	32.1	1.02	5524	151	202	5707
627	602.4	450.42	150.6	1.995	54	0.752	7450	101	155	2700.6
2423	0	3615.2	47	1.995	30	0.635	8758	112	218	3089
676	0	0	369	1.994	34	0.6	1685	150	154	1454.6
584	113.6	71.909	170.4	1.995	34.8	0.662	3646	103	116	1766.8
1108	224.4	136.21	149.6	1.995	34.8	0.662	3646	107	116	2338.6
310	25.2	957.6	478.8	1.994	34	0.6	1685	113	143	495
36	0	45.9	35	1.38	30	1	1414	80	80	155
1377	103.2	76.265	154.8	1.994	34	0.6	1685	140	154	1950.4
652	37.2	15.661	148.8	1.994	34	0.6	1685	146	154	1358.1
2340	3393.4	2229.5	6889.6	6.375	51	0.65	13,125	273	292	7058.8
1514	5142	3728	0	2.992	32.1	1.02	5524	192	202	3330
1122	310.2	196.67	206.8	1.994	34	0.6	1685	139	154	1607.8
1887	2228.8	1885.5	10,153	6.366	24.6	0.81	6198	216	230	4271.1
1048	422	294.56	0	1.994	34	0.6	1685	141	154	1376.5
2883	621.2	534.23	7143.8	2.992	32.1	1.02	5524	147	202	5403.6
3050	4427.4	2417.3	2830.6	6.366	24.6	0.81	6198	212	230	5299.7
1635	125	424.1	0	1.995	54	0.752	7350	80	154	3290
1990	2402	1801.5	3317	2.992	32	1.03	5933	179	202	4435.5
1799	5273.6	2689.6	4492.4	3.958	29.5	0.81	7958	20	239	4817
35	0	18.7	97.5	1.049	30	1	1428	80	80	256
304	0	0	1600	3.375	51	0.65	13,125	267	292	6063
167	2265.6	824.68	1510.4	5.82	11.2	0.65	7100	168	182	2253.1
645	141	141	0	2.376	44	78	12,441	100	192	4508
2933	3569.4	2056	3040.6	3.375	48.6	0.67	12,795	220	290	7292.6
240	414	339.5	0	1.995	51	1	12,721	80	200	3132

Table 3 RBF neural network parameters used in this study

Parameter	Type/value
Activation function	Gaussian
Spread value	0.5
Number of training data	325
Number of testing data	90

(KKT) conditions. The Lagrange function is constructed as below:

$$L_{svm} = \frac{1}{2}\, w^T w + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\delta_i\left\{\left[w^T\phi(x_i)+b\right]+\xi_i - y_i\right\} \tag{10}$$

where the Lagrange multiplier is the δ_i; to find optimal conditions of the problem, implicit derivation should be applied for w, b, ξ, and δ_i as follows:

$$\begin{cases} \dfrac{\partial L_{svm}}{\partial \mathrm{w}} = 0 \rightarrow w = \sum \delta_i \phi(x_i) \\ \dfrac{\partial L_{svm}}{\partial \xi_i} = 0 \rightarrow \delta_i = C\xi_i \;\; , i = 1,2,3\ldots,N \\ \dfrac{\partial L_{svm}}{\partial \mathrm{b}} = 0 \rightarrow \sum_{i=1}^{N} \delta_i = 0 \\ \dfrac{\partial L_{svm}}{\partial \delta_i} = 0 \rightarrow \; y_i = w\phi(x_i) + b + \xi_i \;\; , i = 1,2,\ldots,N \end{cases} \tag{11}$$

The following SVM equation is obtained by solving the above system of equations.

$$Y(x) = \sum \delta_i \phi_{i,j} + b \tag{12}$$

where $\phi_{i,\,j}$ is the kernel function, defined as Eq. (13) (Yuan et al. 2015).

$$\phi_{i,j} = K\left(x_i, x_j\right) = \phi(x_i)^T \phi\left(x_j\right) \tag{13}$$

In summary, for the present study, the same as the RBF neural network, we used the RBF kernel function due to its regression reduction ability, as introduced in Eq. (3). To determine the LSSVM tuning parameters (C and σ), MSE was applied as an objective function. The training, testing, and validation set of data were selected as 70, 20, and 10% of the total data set, respectively.

Table 4 GP optimum characteristics used in this study

Parameter	Type/value
Population number	1000
Maximum number of iterations	200
Maximum number of genes	8
Maximum gene depth	6
Used functions	Plus (+)—minus (−)—multiplication (×)—division (/)—square root (√)—natural logarithm (log)—square power (a^2)—exponential power (e^x)

Table 5 The optimal value for the LSSVM parameters

Parameter	Value
Penalty factor (C)	791.5198
RBF kernel tuning parameter (σ)	3.4392

The following steps of building an LSSVM model for this study are carried out:

1. The RBF kernel function is selected;
2. Tuning parameters are optimized;
3. The model are trained and tested;
4. The capability of the model is evaluated;
5. The algorithm is repeated to reach the best LSSVM model;
6. Stopping the process and saving the best model.

After reaching optimal parameters, namely, penalty factor (C) and RBF kernel tuning parameter (σ), all collected data were randomly used in the program.

Field data

Production data gathered from Iranian oil fields, with various fluid properties and well geometries, are used to create computer-based models. There are 450 data points from the bottom-hole pressure of producing wells and fluid and pipe parameters used as input randomly in training models. The real-time production data includes wellhead pressure (P_{wh}), oil flow rate (Q_o), gas flow rate (Q_g), water flow rate (Q_w), tubing inside diameter (D_i), oil API gravity (API), gas specific gravity (γ_g), wellbore depth (D), surface temperature (T_s), bottom-hole temperature (T_d), and bottom-hole pressure (P_{bh}). Table 1 summarizes the minimum, maximum, average, and standard deviation of all data points for developing models in the RBF neural network, GP, and LSSVM.

In Table 2, a small part of recorded production data is shown, which downhole gauges measure the bottom-hole pressure data.

Results and discussion

RBF neural network, GP, and LSSVM optimum parameters

Parameters have to be selected to construct a model so that prediction error reaches its minimum value, so there is often a

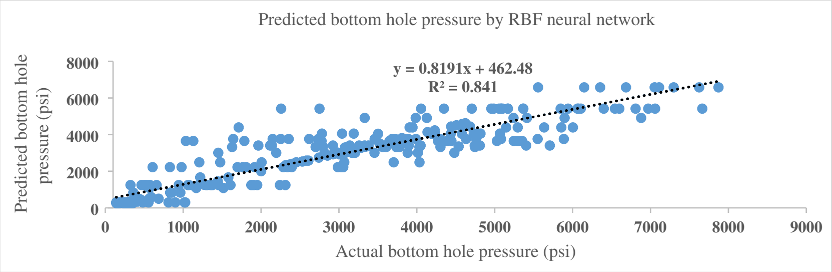

Fig 3 Predicted bottom-hole pressure versus actual bottom-hole pressure by trained RBF neural network

need to train the models several times. The RBF neural network's parameters are shown in Table 3. The spread value, which controls the smoothness of the interpolating function, is the main tuning parameter.

GP's main parameters are mentioned previously, and after carrying out the first analysis, we set the population number and the maximum number of iterations to 1000 and 200, respectively. Then we tuned the maximum number of genes from 3 to 10 and the maximum depth of the genes from 2 to 10 in a total of 72 run-step to obtain the minimum prediction error. The smallest error was received in a maximum of eight genes and a maximum of six gene depth. Table 4 summarizes the optimal GP parameters used in this study.

LSSVM model parameters are the penalty factor (C) and RBF kernel tuning parameter (σ). In this study, the optimum value for these parameters was obtained using coupled simulated annealing (CSA) optimization algorithm, shown in Table 5.

Accuracy of the RBF neural network, LSVM, and GP models

Statistic parameters were computed to investigate the prediction ability of the developed models. In this regard, to assess the model performance, determination coefficient (R^2), average relative error (ARE), average absolute relative error (AARE), and root mean square error (RMSE) used. Regression goodness of fit is determined by R^2, which varies between 0 and 1, which the $R^2 = 1$ is the perfect situation. The following equations are the definition of the statistical parameters used in this study.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(P_i^{actual} - P_i^{predicted}\right)^2}{\sum_{i=1}^{N}\left(P_i^{actual} - P_{actual}^{average}\right)^2} \tag{14}$$

$$ARE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(P_i^{actual} - P_i^{predicted}\right)}{P_i^{actual}} \tag{15}$$

$$AARE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\left(P_i^{actual} - P_i^{predicted}\right)}{P_i^{actual}}\right| \tag{16}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(P_i^{actual} - P_i^{predicted}\right)^2}{N}} \tag{17}$$

where P_i^{actual}, $P_i^{predicted}$, and $P_{actual}^{average}$ are actual, predicted, and the actual average of bottom-hole pressure, respectively, and N is the total number of data points.

Computer-based models results and comparison

Figures 3, 4, and 5 show the predicted bottom-hole pressure versus actual values by the three used computer models in this study. LSSVM and GP models outperform the RBF neural network to estimate bottom-hole pressure, based on R^2 value. In the same way, the LSSVM model predicts bottom-hole pressure more accurately than the GP model. The latter is not unnatural, as it is said that SVM is more suitable for small data sets, but sometimes they act more accurately than other algorithms in practical uses (Fayazi et al. 2014).

As it is mentioned, GP can produce models instead of answers, so the model presented by GP based on the inputs

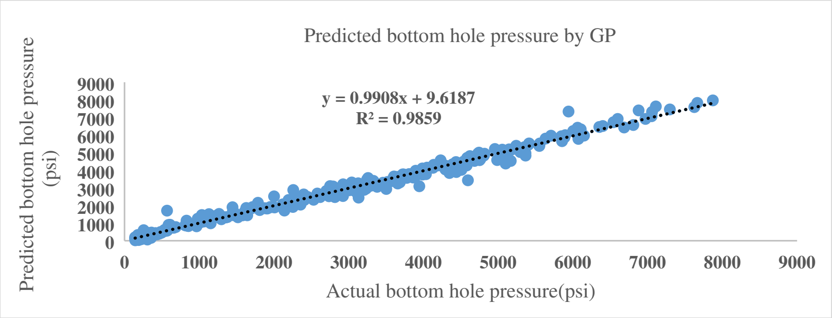

Fig 4 Predicted bottom-hole pressure versus actual bottom-hole pressure by genetic programming (GP)

Fig 5 Predicted bottom-hole pressure versus actual bottom-hole pressure by least square support vector machine (LSSVM)

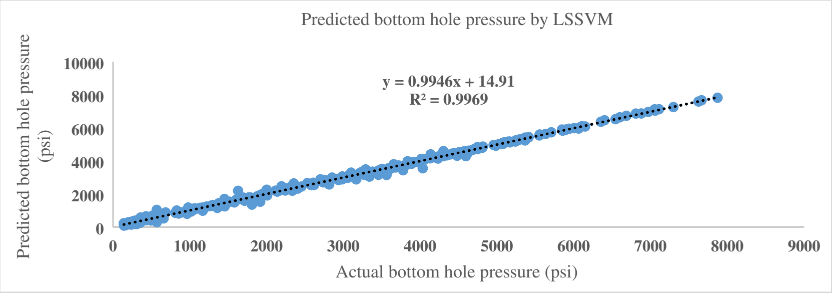

Fig 6 Actual values of bottom-hole pressure of the wells and estimated values by correlations namely: (**A**) Ansari et al., (**B**) Mukherjee and Brill, (**C**) Duns and Ros, (**D**) Beggs and Brill, and (**E**) Hagedorn and Brown methods

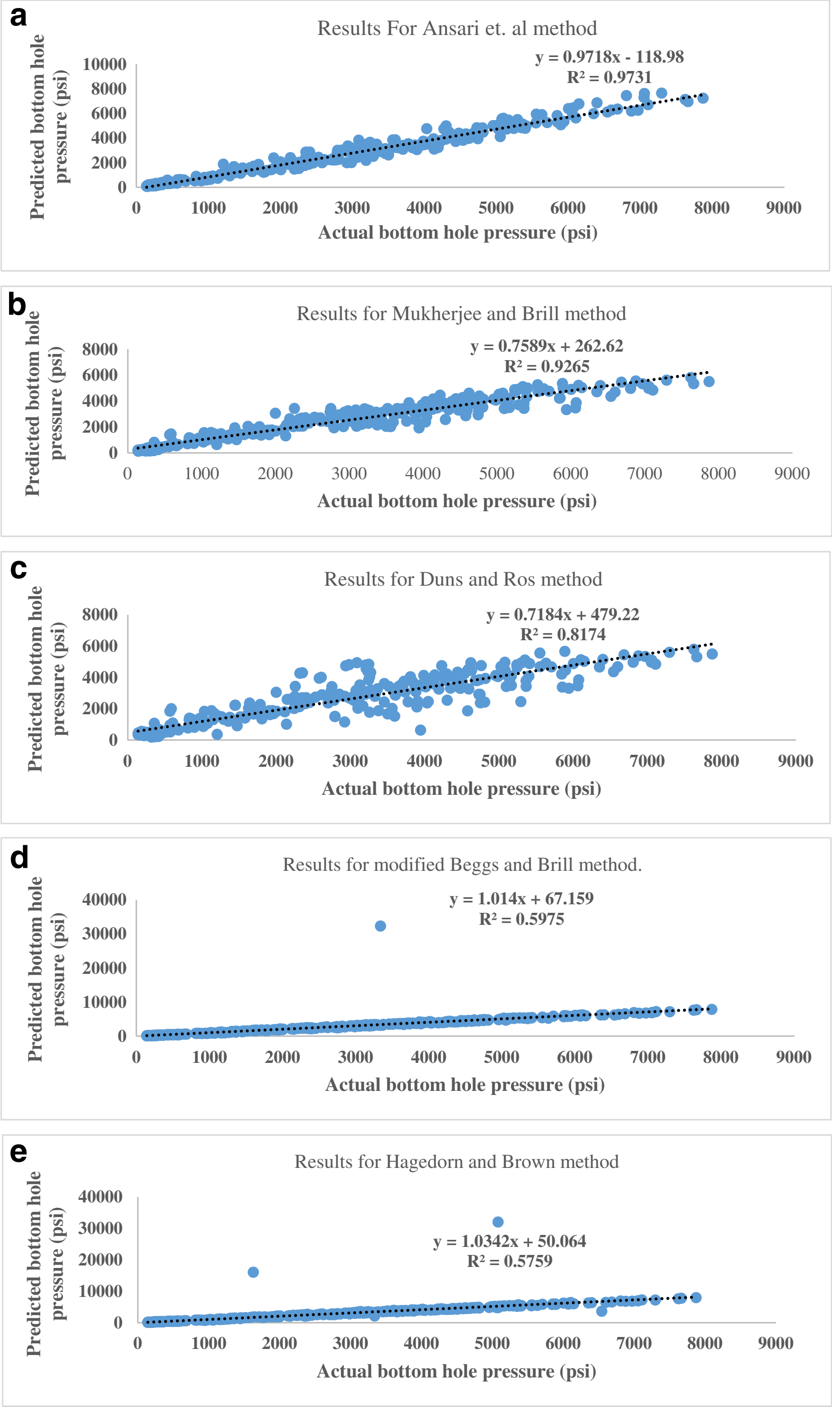

Table 6 Summary of calculated statistical parameters of all used models in this study

Model	R^2	ARE (%)	AARE (%)	RMSE
LSSVM	0.9968	−1.2331	7.7541	112.5652
GP	0.9858	0.0099	12.4586	239.3826
Ansari et. al	0.9731	12.0050	16.4373	383.9738
Mukherjee and Brill	0.9265	7.6632	18.4531	763.5344
RBF neural network	0.8417	−13.22	28.98	800.5431
Duns and Ros	0.8174	−10.4097	33.8313	935.6895
Modified Beggs and Brill	0.5975	−3.4899	8.4114	1676.0500
Hagedorn and Brown	0.5759	−3.4965	12.1311	1855.3117

(production data) is as Eq. (18), with regard to the R^2 value, as shown in Fig. 4. The Eq. (18) units are the same as in Table 2 and are not in the SI unit systems.

$$\begin{aligned} P_{bh} = {} & 0.8744\left(P_{wh} + \log\left(\gamma_g^{Q_o}\right) - \mathrm{T_s}\right) + 33.32\log\left(\frac{\sqrt{P_{wh} + Q_w}}{Q_g T_f \log\left(Q_g\right)}\right) \\ & -809.4\log(D_i) + 45.17\frac{\log\left(\frac{1}{P_{wh}} - P_{wh}T_f\right)}{P_{wh}^2 Q_o Q_g} \\ & + 5.925\log(\log(\log(P_{wh}).Q_o D_i)) - \frac{330.7\left(T_f - D_i^3\right)}{Q_w + T_s} \\ & + \frac{809.4\log\left(Q_o Q_g (P_{wh} + Q_w)\right)}{P_{wh}} + 0.03102D\log\left(\sqrt{Q_w + T_f} + T_s^3\right) \\ & + \frac{0.08744Q_g}{D_i^2} + 0.009942D\log\left(\frac{Q_g^3}{P_{wh}}\right)^2 + 248.1 \end{aligned} \tag{18}$$

Empirical and mechanistic models comparison

This section includes five powerful pressure drop methods, developed as correlations and maps, and has been extensively used in the petroleum industry. Any modifications to the used equations have not been done. The plotted results in predicting bottom-hole pressure for the data set are presented in Fig. 6. The value of R^2 is calculated for each model; therefore, it can be seen that the Ansari et al. model could predict the pressure values more precisely than the other four pressure drop methods, which were applied to vertical two-phase flow wells. The mentioned model has an R^2 value of 0.9731, which is the nearest to 1 compared with other methods. The furthest estimation of pressures belonged to Hagedorn and Brown method by R^2 value of 0.5759. This error could have originated from this fact that this method considers neither liquid holdup nor flow pattern. However, this method is still used in oil and gas wells to compute the pressure drop, but this method is not suitable for this range of fluid parameters, presented in Table 2. Mukherjee and Brill and Duns and Ros methods were in second and third places in predicting pressure by R^2 value of 0.9265 and 0.8174, respectively. Even though Mukherjee and Brill's model is generally used in deviated wells, it could indicate high accuracy pressure values. Beggs and Brill's method did not have an excellent performance here, and the most probable reason for this is that this model is developed for horizontal flows. Also, the experimental works in this method are carried out on water and air.

Overall comparison of proposed models

The analysis based on Eq. (14), R^2 value, carried out, and LSSVM, GP, and Ansari et al.'s methods generated the most accurate results in previous parts. In this part, other statistical parameters, ARE (%) and AARE (%), of all models, computed and plotted in Fig. 7. Lastly, in Fig. 8, the calculated RMSE values are plotted, and all the calculated parameters are listed in Table 6.

Regarding Table 6, the results of Figs. 7 and 8 are the same as the results in the previous section analysis, R^2 value, as there were not many changes in the top

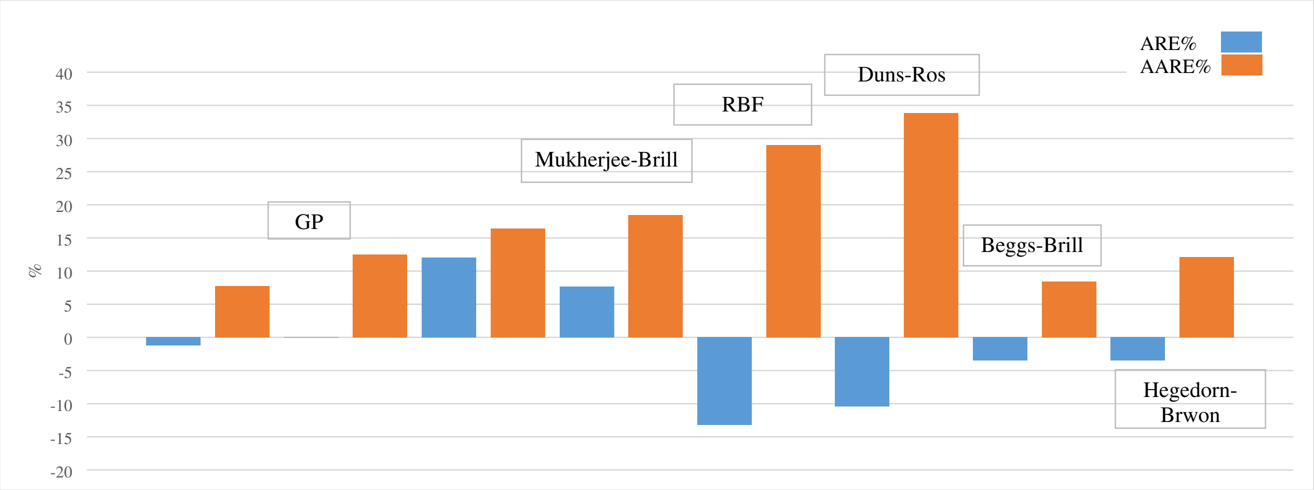

Fig 7 Comparison of ARE (%) and AARE (%) values of all applied models on wells data in this study

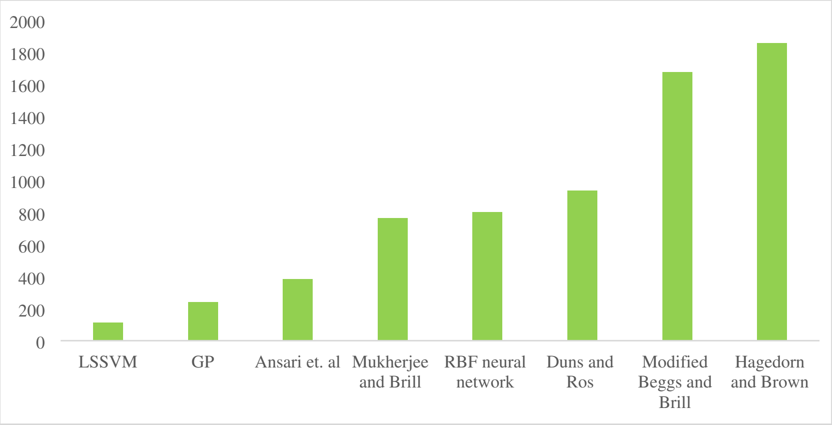

Fig 8 RMSE value for all developed models

three methods, and the origination of the differences is the same as explained before.

Conclusion

In summary, in this paper, we attempted to use two-phase flow and artificial intelligence concepts, and computer programming to develop several models predicting bottom-hole pressure in vertical wells producing oil and gas, based on available production data. The analysis was carried out on the whole data set of production wells and the best models that successfully predicted pressure introduced which were LSSVM and genetic programming that proposed a better precision than other previously developed experimental and mechanistic works. Besides the results and genetic programming, which was the second best tool in estimating pressure by the R^2 value of 0.98, let us present a new correlation for predicting bottom-hole pressure in two-phase flow producing wells, regarding the properties of fluids.

Data Availability Filed data of all production wells are available, but they are not allowed to be shared. The computer codes are also available. However, following the steps explained in the manuscript, a user can generate the codes. The referenced multiphase flow correlations are also available in all related books, e.g., “Shoham, Ovadia. Mechanistic Modeling Of Gas/Liquid Two-Phase Flow In Pipes. Spe, 2005.”

Code availability All codes are available, and it should be mentioned that there had been considerable efforts and time spent on developing the codes. They cannot be shared completely.

Declarations This paper follows the ethical standards of the Arabian Journal of Geosciences. This work is original and has not been published or submitted anywhere else (completely or partially). All the figures and explanations by other researchers cited properly. The paper focused on the scientific aspect of a problem in petroleum wells and contains no sensitive data or misinformation for the readers.

Nomenclature *API*, oil gravity; b, bias term; C, penalty factor; D, wellbore depth; D_i, tubing inside diameter; L, length; P, p, pressure; P_{bh}, bottom-hole pressure; P_{wh}, wellhead pressure; $\frac{\partial P}{\partial L}$, pressure gradient; Q_o, oil flow rate; Q_g, gas flow rate; Q_w, water flow rate; T_d, bottom-hole temperature; T_s, surface temperature; W, weight vector, weight values; x, the symbol for input data; Y, the symbol for a function, the outputs values of simulations; ξ, error estimation term; $\phi(x)$, transfer function, the Kernel function; δ, Lagrange multiplier; σ, spread parameter; ω, weight values; λ, radial basis function, outputs of hidden layer**Unit conversion** Bbl, = 0.158987 m^3; °F, (0 °C × 9/5) + 32 = 32 °F; ft, = 30.48 cm; Inch, = 2.54 cm; Psi , = 6894.76 Pa

References

Ahmadi MA, Ahmadi MR, Hosseini SM, Ebadi M (2014) Connectionist model predicts the porosity and permeability of petroleum reservoirs by means of petro-physical logs: Application of artificial intelligence. J Pet Sci Eng 123:183–200

Ansari AM, Sylvester ND, Sarica C, Shoham O, Brill JP (1994) A comprehensive mechanistic model for upward two-phase flow in wellbores. SPE Prod Facil 9(02):143–151

Asheim H (1986) MONA, an accurate two-phase well flow model, based on phase slippage. SPE Prod Eng 1(03):221–230

Baxendell PB, Thomas R (1961) The calculation of pressure gradients in high-rate flowing wells. J Pet Technol 13(10):1–023

Beggs DH, Brill JP (1973) A study of two-phase flow in inclined pipes. J Pet Technol 25(05):607–617

Chierici GL, Ciucci GM, Sclocchi G (1974) Two-phase vertical flow in oil wells-prediction of pressure drop. J Pet Technol 26(08):927–938

Coello CAC, Lamont GB, Van Veldhuizen DA (2007) Evolutionary algorithms for solving multi-objective problems, vol 5. Springer, New York, pp 79–104

Duns Jr H, Ros NCJ (1963) Vertical flow of gas and liquid mixtures in wells. In *6th world petroleum congress*. World Petroleum Congress

Fancher Jr GH, Brown KE (1962) Prediction of pressure gradients for multiphase flow in tubing. In *Fall Meeting of the Society of Petroleum Engineers of AIME*. Society of Petroleum Engineers

Fayazi A, Arabloo M, Shokrollahi A, Zargari MH, Ghazanfari MH (2014) State-of-the-art least square support vector machine application for accurate determination of natural gas viscosity. Ind Eng Chem Res 53(2):945–958

Gharbi RB, Mansoori GA (2005) An introduction to artificial intelligence applications in petroleum exploration and production. J Pet Sci Eng 49(3-4):93–96

Hagedorn AR, Brown KE (1965) Experimental study of pressure gradients occurring during continuous two-phase flow in small-diameter vertical conduits. J Pet Technol 17(04):475–484

Hasan AR, Kabir CS (1992) Two-phase flow in vertical and inclined annuli. Int J Multiphase Flow 18(2):279–293

Jahanandish I, Salimifard B, Jalalifar H (2011) Predicting bottom-hole pressure in vertical multiphase flowing wells using artificial neural networks. J Pet Sci Eng 75(3-4):336–342

Khamehchi E, Zolfagharroshan M, Mahdiani MR (2020) A robust method for estimating the two-phase flow rate of oil and gas using wellhead data, Springer

Koza JR (1990) Genetic programming: a paradigm for genetically breeding populations of computer programs to solve problems, vol 34. Stanford University, Department of Computer Science, Stanford

Koza JR, Keane MA, Streeter MJ, Mydlowec W, Yu J, Lanza G (2006) Genetic programming IV: Routine human-competitive machine intelligence (vol. 5). Springer Science & Business Media, Berlin

Malallah A, Nashawi IS (2005) Estimating the fracture gradient coefficient using neural networks for a field in the Middle East. J Pet Sci Eng 49(3-4):193–211

Mukherjee H, Brill JP (1985) Empirical equations to predict flow patterns in two-phase inclined flow. Int J Multiphase Flow 11(3):299–315

Naderi M, Khamehchi E (2018) Application of optimized least square support vector machine and genetic programming for accurate estimation of drilling rate of penetration. Int J Energy Optim Eng (IJEOE) 7(4):92–108

Payne GA, Palmer CM, Brill JP, Beggs HD (1979) Evaluation of inclined-pipe, two-phase liquid holdup and pressure-loss correlation using experimental data (includes associated paper 8782). J Pet Technol 31(09):1–198

Petalas N, Aziz K (2000) A mechanistic model for multiphase flow in pipes. J Can Pet Technol 39. https://doi.org/10.2118/00-06-04

Poettman FH, Carpenter PG (1952) The multiphase flow of gas, oil, and water through vertical flow strings with application to the design of gas-lift installations. In Drilling and production practice. American Petroleum Institute

Searson DP, Leahy DE, Willis MJ (2010) GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. In *Proceedings of the International multiconference of engineers and computer scientists* (Vol. 1, pp. 77-80)

Shoham O (2006) Mechanistic modeling of gas-liquid two-phase flow in pipes. Richardson, TX: Society of Petroleum Engineers

Watson A (2016) Geothermal engineering. Springer-Verlag New York

Yuan X, Chen C, Yuan Y, Huang Y, Tan Q (2015) Short-term wind power prediction based on LSSVM–GSA model. Energy Convers Manag 101:393–401

Zolfagharroshan M, Khamehchi E (2020) A rigorous approach to scale formation and deposition modeling in geothermal wellbores. Geothermics 87:101841