Prediction performance advantages of deep machine learning algorithms for two-phase flow rates through wellhead chokes

Hossein Shojaei Barjouei¹ · Hamzeh Ghorbani² · Nima Mohamadian³ · David A. Wood⁴ · Shadfar Davoodi⁵ · Jamshid Moghadasi⁶ · Hossein Saberi⁷

Received: 17 September 2020 / Accepted: 4 January 2021 / Published online: 23 February 2021
© The Author(s) 2021

Abstract

Two-phase flow rate estimation of liquid and gas flow through wellhead chokes is essential for determining and monitoring production performance from oil and gas reservoirs at specific well locations. Liquid flow rate (Q_L) tends to be nonlinearly related to these influencing variables, making empirical correlations unreliable for predictions applied to different reservoir conditions and favoring machine learning (ML) algorithms for that purpose. Recent advances in deep learning (DL) algorithms make them useful for predicting wellhead choke flow rates for large field datasets and suitable for wider application once trained. DL has not previously been applied to predict Q_L from a large oil field. In this study, 7245 multi-well data records from Soroush oil field are used to compare the Q_L prediction performance of traditional empirical, ML and DL algorithms based on four influencing variables: choke size (D_{64}), wellhead pressure (P_{wh}), oil specific gravity (γ_o) and gas–liquid ratio (GLR). The prevailing flow regime for the wells evaluated is critical flow. The DL algorithm substantially outperforms the other algorithms considered in terms of Q_L prediction accuracy. The DL algorithm predicts Q_L for the testing subset with a root-mean-squared error (RMSE) of 196 STB/day and coefficient of determination (R^2) of 0.9969 for Soroush dataset. The Q_L prediction accuracy of the models evaluated for this dataset can be arranged in the descending order: DL > DT > RF > ANN > SVR > Pilehvari > Baxendell > Ros > Gilbert > Achong. Analysis reveals that input variable GLR has the greatest, whereas input variable D_{64} has the least relative influence on dependent variable Q_L .

Keywords Two-phase oil/gas flow rate · Wellhead choke-influencing variables · Deep versus machine learning · Large field datasets · Overfitting considerations

✉ David A. Wood
dw@dwasolutions.com

Hossein Shojaei Barjouei
hossein.shojaee@modares.ac.ir

Hamzeh Ghorbani
hamzehghorbani68@yahoo.com

Nima Mohamadian
nima.0691@gmail.com

Shadfar Davoodi
davoodis@hw.tpu.ru

Jamshid Moghadasi
j.moghadasi@put.ac.ir

Hossein Saberi
hsaberi97@yahoo.com

¹ Mechanical Engineering Department, Tarbiat Modares University, Tehran, Iran

² Young Researchers and Elite Club, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran

³ Young Researchers and Elite Club, Omidyeh Branch, Islamic Azad University, Omidyeh, Iran

⁴ DWA Energy Limited, Lincoln, UK

⁵ School of Earth Sciences and Engineering, Tomsk Polytechnic University, Lenin Avenue, Tomsk, Russia

⁶ Petroleum Engineering Department, Petroleum Industry University, Ahvaz, Iran

⁷ Department of Electrical Engineering, University of Guilan, Rasht, Iran

Abbreviations

a, b, c, d, e, f, g	Experimental coefficients calculated where sufficient data are available for specific reservoir systems	ω	Weight vector
AAPD	Absolute average percent deviation	$\phi(x)$	High-dimensional feature space
AI	Artificial intelligence	$\hat{y}_i$	Predicted output
ANN	Artificial neural network	$\bar{\epsilon}^{-\Pi}$	Average predicted oil flow rate prediction through orifice plate for data point
APD	Average percent deviation	$\epsilon_i^{-\Pi}$	Predicted oil flow rate prediction through orifice plate for data point i
B	Bias (threshold)	h^0	Network input
BS&W	Base sediment and water production	E_{MSE}	Mean square error
C	Regularization factor	L_K	Output layer neuron
D ₆₄	The size of choke (1/64 inch), GLR is the gas to liquid ratio	W_{ij}	Wright for the connection between ith input and jth neuron
DL	Deep learning	W^k	Weights' matrix
DNN	Deep neural network	b^k	Bias values' array
DT	Decision tree	g_t	Objective function gradient
FFANN	Feedforward artificial neural network	x_i	Input parameters
GLR	Gas to liquid ratio	x_i^l	Represents the value of attribute l of data record i
InB	In bag	$xmax^l$	The maximum values of the attribute l
k	Referred to as the ratio of specific heat	$xmin^l$	The minimum values of the attribute l
ML	Machine learning	$\frac{y_i}{\Phi}$	Output parameters
MSE	Mean square error	Φ_i	Average Value for Input Variable Φ
N	Number of samples in data set		Input Value of Data Point i for Input Variable Φ
OOB	Out-of-bag	α_i	Lagrangian multiplier
PDi	Percent deviation for i th data record	α_i^*	Lagrangian multiplier
P _{downstream}	Choke outlet pressure downstream of the choke	θ_t	Model's parameter at time step t
P _{upstream}	Pressure upstream of the choke	ξ^*	Slack variable
P _{wh}	The wellhead pressure	σ^k	Activation function
R	The dataset of data records	$\in$	Smoothing term
R	The correlation coefficient		
RBF	Radial basis function		
RF	Random forest		
RMSE	Root-mean-square error		
SD	Standard deviation		
SNN	Simple neural network		
SVM	Support vector machine		
SVR	Support vector regression		
T	Temperature		
T _{sc}	Standard temperature		
X	The value of variable x in a specific data record		
γ_g	Gas-specific gravity		
γ_o	Oil-specific gravity		
θ	Bias parameter		
σ	RBF kernel variance		
d	Polynomial degree		
f	Activation function		
k	Scale parameter		
k	Represent the number of layers		
t	Intercept		
η	Learning rate		
ξ	Slack variable		

Introduction

One of the key factors in determining the production performance of oil/gas reservoirs is to establish value of the variables influencing two-phase flow rate in oil and gas wells (Lak et al. 2014; Choubineh et al. 2017; Ghorbani et al. 2017b, c, 2019; Mirzaei-Paiaman and Salavati 2013). To determine the flow rate of each well (gas with condensate/oil), to control and stabilize the flow, and to prevent reservoir damage by creating a back pressure on the reservoir, a well reducer is typically installed (Ghorbani et al. 2019). Well reducers help to protect well equipment as well as the reservoir. The back pressure created by them helps to control pressure and flow rate (Mirzaei-Paiaman and Salavati 2013; Guo 2007; Al-Attar 2009; Chong et al. 2009; Nasriani et al. 2016; Mirzaei-Paiaman and Salavati 2012; Omana et al. 1969; Poettmann and Beck 1963). The increase in well-bore pressure associated with well reducers is due to the presence of oil-soluble gas or gas condensate, which forms in the wellbore due to pressure drop in the production tubing.

Wellhead reducers belong to two categories (Bairamzadeh and Ghanaatpisheh 2015; Gorjaei et al. 2015): (A) fixed chokes or positive chokes (Fig. 1) and (B) variable chokes (Fig. 2). In fixed chokes the bore or aperture diameter cannot be changed. In variable chokes the bore or aperture can be increased or decreased to adjust fluid flow rate through it (Gorjaei et al. 2015; Elhaj et al. 2015).

In fields with many wells, it is often not practical to measure flow rates directly at each well head choke. Rather, it is necessary to estimate flow at individual wells from available measurements of influencing variables. In this study, we evaluate and compare the liquid flow rate (Q_L) prediction performance of four machine learning (ML) algorithms: support vector regression (SVR), decision tree (DT), random forest (RF) and artificial neural network (ANN) with a deep learning neural network (DL) based on

four key influencing variables. The four influencing variables considered are choke size (D_{64}), wellhead pressure (P_{wh}), oil-specific gravity (γ_o) and gas–liquid ratio (GLR) are evaluated. The superior performance of the DL demonstrates that it is better suited than the ML algorithms to provide reliable and accurate Q_L predictions with large datasets. A dataset of 7245 data records (made available to readers) from ten wells in the offshore Soroush oil field (Iran) forms the basis of the analysis presented.

Theory of two-phase fluid flow through wellhead chokes

The principles of two-phase fluid flow are as follows:

Fig. 1 Image of fixed or positive wellhead chokes with 32/64 in or 1/2 in apertures

Fig. 2 Schematic diagram showing internal components of a variable wellhead choke (Ghorbani et al. 2017c; Latif et al. 2012)

(A) When a fluid flows through a flow pipeline, the fluid pressure is initially above the bubble point, i.e., the gas remains dissolved in the liquid.

(B) When the fluid is at the bubble point, gas bubbles begin to emerge from the liquid and it becomes a two-phase fluid.

(C) When the fluid pressure drops below the bubble point pressure, the fluid moves in a pipeline as two-phase flow.

Key reasons for involving wellhead chokes in a production flow stream are: (1) to create a pressure drop in the flow stream to prevent damage to well equipment, (2) to facilitate separation of gas from oil in separators; and (3) to create a back pressure on the reservoir to assist in maintaining reservoir pressure.

Although biphasic flowmeters were developed decades ago, their high-cost and ongoing calibration requirements to accurately calculate of two-phase fluid flow have limited their uptake. Calculating the exact GLR value inside the well reducer itself, while a well is online, is costly and requires the installation of accurate measurement sensors. In practice, estimated GLR data are typically used to calculate fluid flow rates through wellhead chokes.

Flow regimes in two-phase flow through wellhead chokes

For two-phase fluids (liquid and gas) passing through a choke, critical and subcritical flow regimes may occur (AlAjmi et al. 2015; Zarenezhad and Aminian 2011). *Critical flow* (sonic) occurs when the fluid flow reaches the speed of sound. In this type of flow, the fluid's flow rate has no effect on the downstream pressure and occurs in a state of downstream-pressure independence (Perkins 1993; Al-Attar 2008a, b). It is more difficult to estimate the fluid flow rates under *subcritical flow* regimes, where the ratio of outlet pressure to pre-reducing pressure (inlet pressure) is greater than 0.588 (Bairamzadeh and Ghanaatpisheh. 2015; Gould 1976; Safar Beiranvand et al. 2012; Nasriani and Kalantari ASL 2011). The ratio of the output pressure ($P_{\text{downstream}}$), inlet pressure (P_{upstream}), determined by Eq. (3) can distinguish between critical and sub-critical flow regimes, as displayed in Fig. 3 (Guo 2007; Al-Attar 2008a, b; Gould 1976; Safar Beiranvand et al. 2012; Nasriani and Kalantari ASL 2011; Guo and Ghalambor 2012; Ling 2012):

$$\frac{P_{\text{downstream}}}{P_{\text{upstream}}} = \left(\frac{2}{k+1} \right)^{\frac{k}{k-1}} \quad (1)$$

where $P_{\text{downstream}}$ = choke outlet pressure downstream of the choke; P_{upstream} = pressure upstream of the choke; and, $k = C_p/C_v$ and is referred to as the ratio of specific heat.

Fig. 3 Choke performance in terms of flow rate versus downstream to upstream pressure ratio (Kaydani et al. 2014)

One of the tasks that has been done in recent years is to use field data to calculate and predict, as well as to determine the parameters used in the oil and gas industry, for example in the following areas have been addressed: reservoirs (Ghorbani et al. 2017a); formation damage (Mohamadian and Ghorbani 2015), wellbore stability (Darvishpour et al. 2019), rheology and filtration (Mohamadian et al. 2018), production (Ghorbani and Moghadasi 2014a; Ghorbani et al. 2014b; 2017b); drilling fluid (Mohamadian et al. 2019).

Tangeren et al. (1949) conducted fundamental studies on fluid flow through constraints, focusing on bubble point pressure and critical flow conditions. They achieved these conditions by adding gas bubbles to incompressible fluids to prevent upstream and downward pressure transfer associated with a restriction. Gilbert (1954), based on 260 data records for chokes ranging in size from 6/64 inches to 18/64 inches, developed empirical Eq. (2) for critical flow (Table 1). Various modifications to the Gilbert equation have been proposed, including those of (Baxendell 1958), Ros (1960), Achong (1961), and Pilehvari (1981) (Table 1) Eq. (3–6). These modifications are based on observations of data from different regions and a wide range of chokes size. They achieve variable prediction accuracy for two-phase flow rates through the wellhead chokes.

Safar Beiranvand et al. (2012) evaluated 748 production flow data records from 10 oil wells in Iran, with choke sizes ranging from 16/64 to 40/64, based on which they proposed Eq. (7) for flow rate prediction (Table 1). In that relationship, the percentage of water production, measured as base solids and water (BS&W), is added to the Gilbert Equation (1954). Ghorbani et al. (2019) based on 82 wellhead choke data measurements spanning liquid flow rates from < 100 to 30,000 stock tank barrels/day (STB/D) proposed flow rate prediction Eq. (8) (Table 1), which also included BS&W in its formulation. Mirzaei-Paيمان and Salavati (2013), based on 102 production flow data records, proposed Eq. (9) (Table 1), in which oil density and gas-specific gravity are added to the Gilbert equation.

Table 1 Empirical equations proposed for fluid flow across oil field chokes

Authors/year	Equation	Formula	Coefficient
Gilbert (1949)	(2)	$Q_L = a \frac{P_{wh}^b D_{64}^c}{GLR^d}$	a = 0.1, b = 1, c = 1.89, d = 0.546
Baxendell	(3)	$Q_L = a \frac{P_{wh}^b D_{64}^c}{GLR^d}$	a = 0.1046, b = 1, c = 1.93, d = 0.546
Ros	(4)	$Q_L = a \frac{P_{wh}^b D_{64}^c}{GLR^d}$	a = 0.05747, b = 1, c = 2.00, d = 0.500
Achong	(5)	$Q_L = a \frac{P_{wh}^b D_{64}^c}{GLR^d}$	a = 0.26178, b = 1, c = 1.88, d = 0.650
Pilehvari	(6)	$Q_L = a \frac{P_{wh}^b D_{64}^c}{GLR^d}$	a = 0.021427, b = 1, c = 2.11, d = 0.313
Safar Beiranvand et al. (2012)	(7)	$Q_L = a \frac{P_{wh}^b D_{64}^c (1-BS\&W\%)^d}{GLR^e}$	a = 0.0382, b = 1, c = 2.151, d = 0.52965, e = 0.5154
Ghorbani et al. (2019)	(8)	$Q_L = \frac{P_{wh}^b D_{64}^c (1-BS\&W\%)^d}{a GLR^e}$	a = 1.3522, b = 1, c = 1.7056, d = -0.164, e = 0.74042
Mirzaei-Paيمان et al. (case-4) (2013)	(9)	$Q_L = a \frac{P_{wh}^b D_{64}^c T_{go}^d}{GLR^f}$	a = 0.052439, b = 1, c = 1.9108, d = 0.3988, e = 0.1711, f = 0.5220
Choubineh et al. (2017)	(10)	$Q_L = a \frac{P_{wh}^b D_{64}^c T_{go}^d \left(\frac{T}{T_{sc}}\right)^f}{GLR^g}$	a = 0.067,662, b = 1, c = 2.08,918, d = 0.625,862, e = 1.583,074, f = 0.000,453, g = 0.508,714

Choubineh et al. (2017), considering information from 235 production flow-test data records, proposed Eq. (10) (Table 1). The advantage of Eq. (10), compared to the other formulations described, is that it takes fluid temperature into account.

Machine learning (ML) algorithms

Developments over recent years of industrial automation and intelligent machine monitoring and recording technologies, mean that large digital datasets are now routinely available for system behavior pattern analysis. Machine learning (ML) systems are flexible intelligent computer algorithms that provide data-driven tools for improving automated learning and prediction capabilities for many systems ML algorithms achieve cognition and learning through data-record variable relationships and pattern recognition to build an intelligent model for making predictions and decisions (Bonaccorso 2017).

In recent years, several researchers have applied ML methods to flow measurement through wellhead or production-unit choke in oil and gas fields. Table 2 compiles some such recently published ML studies and the flow rate prediction accuracies they have achieved.

Development from shallow to deep learning models

The structure of simple neural networks (SN), including artificial neural networks (ANN), is inspired by the neural communications and cognitive structures observed in animal and human brains (Barrow 1996). Such networks consist of a large number of interconnected neural processing units (neurons or nodes), each of which, when configured in a sequential processing system, creates a specific communication pattern in response to a specific input or stimulus (Pouyanfar

et al. 2018). Environmental perception sensors derived from raw input information activate neurons in the input layer of SN network. Subsequently, other neurons located in the next layer, hidden inside the network, are activated based on the weighted communications they receive from the input-layer neurons (Schmidhuber 2015). Improvements in computer capacities and speeds now make it possible to rapidly process information passed through neural networks involving more neurons and/or more hidden (Pouyanfar et al. 2018). Deep ANN or DNN architectures represent an improved version of traditional ANN (Rolnick et al. 2017) which mainly consist of an input layer, one hidden layer, and an output layer. On the other hand, DNN consist of a feed-forward neural network with a large number of neurons, most of which are distributed across multiple hidden layers (Saikia et al. 2020). Deep learning methodologies exploit new training methods and communication algorithms among the neurons, helping them to develop high prediction performance and reliability from multi-layered learning models.

Machine learning (ML) versus deep learning (DL)

Big data analysis

In technical terms, DL is a powerful subset of ML methods that has been developed in a similar way to meet challenges and limitation of many classical machine-learning techniques when confronted with processing large data sets (“big data”) characterized by having many data records and many input variables, and/or unstructured data fields such as involved in image and speech recognition algorithms (Nguyen et al. 2019). DL techniques tend to offer improved efficiency and accuracy as the size of the data set increases (Zhang et al. 2018). For smaller datasets, classical ML algorithms often outperform DL networks, certainly in terms of execution

Table 2 Some applications of machine learning methods to predict two-phase flow through wellhead choke

Machine learning application to predict multiphase flow through wellhead choke							
Author and year	Field	Machine learning method	Input variables	Total dataset	R ²	Standard division (SD)	Reported error
Rashid et al. (2019)	Oil field	Radial based function (RBF) neural network- Genetic algorithm (GA) optimization	Upstream press. (Pup) Gas liquid ratio (GLR) Choke size (D)	276	0.986	0.554	RMSE=292.378 & AAPD=2.6103
Khamis et al. (2020)	Oil field	Artificial neural network (ANN) Fuzzy logic (FL) Support vector machine (SVM) Functional network (FN)	Upstream press. (Pup) Gas liquid ratio (GLR) Choke size (D) Temperature (T)	347	0.84 0.81 0.997 0.81		
Ghorbani et al. (2018)	Oil field	Genetic Algorithm (GA) & Excel solver optimizer	Liquid flow rate (QL) Wellhead press. (Pwh) Gas liquid ratio (GLR) Choke size (D) Water cut	182	0.997	563.85	RMSE=562.52 & AAPD=7.33
Ghorbani et al. (2017c)	Gas condensate field	Fire fly optimization algorithm	Qg—GOR- Pup- Gas specific gravity- choke size	92	0.9677	0.0031	RMSE=0.0071 & AAPD=20.21
Choubineh et al. (2017)	Oil field	Artificial neural network (ANN) ANN- teaching—learning-based optimization (ANN-TLBO) algorithm Nonlinear regression	Choke size (D), Wellhead pressure (Pwh), Oil specific gravity, Gas specific gravity, Gas liquid ratio (GLR), Temperature (T)	113	0.947 0.981		RMSE=1227 & AARE=14.85 RMSE=714 & AARE=6.5
Gorjaei et al. (2015)	Oil field	Least squares support vector machine- Particle swarm optimization (PSO-LSSVM)	Choke size (D), Wellhead pressure (Pwh), Gas liquid ratio (GLR)	276	0.963 0.976		RMSE=1077 & AARE=9.4 ARE=0.80 & AARE=7.99

speed. On the other hand, DL networks often require little or no supervision (additional training) once they are initially trained/calibrated, whereas ML models typically require more supervision and re-training as part of their development and application.

Feature extraction

Classical ML algorithms usually cannot be implemented directly on a raw data records incorporating many variables (high dimensionality). Such algorithms require a raw-data, pre-processing step, typically incorporating feature extraction/selection (Zheng and Casari 2018). The performance of classical ML techniques is highly dependent on the user-based presentation quality of the input data. Inconsistent quality of the data presentation and/or the allocation of inappropriate parametric weights often leads to poor results and generally can compromise ML performance (Pouyanfar et al. 2018). Effective ML feature selection depends upon a human user possessing sufficient knowledge and understanding of a system and the potential data inconsistency problems it incorporates. In contrast, DL algorithms automatically perform feature selection without the need for human user intervention. Hence, DL models with higher processing capacity and enhanced feature recognition capabilities are often able to provide accurate prediction models either without or with much less human assistance/input (Goodfellow et al. 2016).

Economic and interpretation perspectives

Implementation of DL algorithms on multi-dimensional and complex “big data” systems requires the use of powerful computational processors. These can be costly as they need

to be capable of conducting and storing the large numbers of calculations and/or intermediate data manipulations required in a reasonable timeframe. On the other hand, ML algorithms can typically be executed using standard, low-cost PC-grade computer systems. Since the feature selection process in most classical ML models is not performed automatically, users typically gain a better understanding of the various factors associated with the input variables that influence the ML algorithms prediction performance, which can be helpful when tuning the control parameters associated with an ML model's configuration. On the other hand, users tend to be blind to the internal functions used by DL models in the automated feature selection, making them behave, in many cases, like “black boxes”.

Figure 4 shows the generic differences in the network architecture and implementation usually associated with ML and DL algorithms.

There are only a limited number of studies available that have focused the unique capabilities of DL algorithms on topics related to fluid flow measurement. Loh et al. (Loh et al. 2018) trained and then tested the long short-term memory (LSTM) deep learning model to predict real-time production rate using two gas-well datasets that were experiencing salt-deposition issues. Ezzatabadipour et al. 2017 implemented a multilayer perceptron (MLP) algorithm with three hidden layers, configured as a deep learning model, to predict patterns of multiphase flow, based on 5676 data records of gas and liquid two-phase flow in inclined pipes. Wen et al. (Wen et al. 2019) introduced a modified recurrent (RU-Net) algorithm which is a deep convolutional neural network (CNN) extending the U-Net architecture, originally developed for biomedical image segmentation, to successfully predict multiphase flow in porous media involving CO₂ injection.

Fig. 4 Generic differences in the network architecture and execution of ML and DL algorithms

They used 128 feature maps composed of 2-D plans extracted from radially plotted coordinates of a wellbore subjected to CO₂-injection the inputs for their model included series of permeability, injection duration, injection rate, and saturation fields planes.

Many research articles have applied ML methods to two-phase flow rate prediction through wellhead chokes. However, to the authors' knowledge no results have so far been published applying DL models to make such predictions. In this study, a comparison of the flow rate prediction performance of DL and ML models is provided using a substantial dataset (7245 data records from three wells). It reveals that DL models can outperform ML models for the dataset evaluated.

Methodology

Workflow diagram

A workflow diagram (Fig. 5) summarizes the steps involved in constructing, evaluating and comparing prediction performance of DL and ML algorithms in two-phase flow rate prediction through wellhead chokes. The process sequence

begins with data collecting followed by data variable characterization, including the determination of the maximum and minimum values for each data variable involved. This information is used to normalize all data variables to range between +1 and -1 by applying Eq. (11).

$$x_i^l = \left(\frac{x_i^l - x_{min}^l}{x_{max}^l - x_{min}^l} \right) * 2 - 1 \quad (11)$$

where x_i^l = the value of attribute for data record l ; x_{min}^l = the minimum value of the attribute among all the data records in the dataset; and x_{max}^l = the maximum value of the attribute among all the data records in the dataset.

The normalized data records are then allocated to training and testing subsets. In this study, 80% of the data records comprising a dataset are allocated to the training subset and the remaining 20% of the data records are allocated to the testing subset. The testing subset is held independently and in no way involved in the algorithm's training procedure; it is used solely to test the flow-rate prediction performances of the trained algorithms. The relative prediction performances of the trained algorithms applied to the testing subsets are then used to identify the algorithm that provides the best prediction performance for the dataset.

Fig. 5 Schematic diagram of the workflow sequence applied for comparing the prediction performance of ML and DL algorithms

Learning network algorithms evaluated

Learning networks combining various algorithms are now routinely applied to solve a wide range of petroleum engineering challenges, including: reservoirs (Ghorbani et al. 2020a), production (Choubineh et al. 2017; Ghorbani et al. 2017c; 2018; 2019; 2020a; 2020b), process (Ghorbani et al. 2018) and drilling (Mohamadian et al. 2021; Rashidi et al. 2020) drilling trajectories (Atashnezhad et al. 2014); and fluid processing (Wood 2018). In this study, we evaluate and compare the following ML and DL algorithms.

Support vector regression (SVR)—ML model

Cortes and Vapnik (1995) developed the support vector machine (SVM) algorithm in 1991 on the basis of statistical learning theory. SVM has been used to solve a wide range of classification, regression and time series prediction problems (Cortez and Vapnik 1995; Cao and Tay 2003; Drucker et al. 1997; Kuo et al. 2013; Shao et al. 2020; Smola and Schölkopf 2004; Rui et al. 2019; Vapnik 2013; Ahmad et al. 2020). For classification SVM applies kernel functions to map input nonlinear vectors into higher dimensions and a hyperplane is constructed within the feature space separating data into two classes. SVM strives to maximize boundary separation between the classes and the hyperplane for the training subset by establishing the support vectors. A regression-based SVR model (Brereton et al. 2010; Pan et al. 2009) is constructed here to predict two-phase flow rate through wellhead chokes.

SVR models require dataset to define their input variables $[x_i] \in X = R^n$ and their corresponding output (or dependent) variable $y_i \in Y = R$ where $i = 1, 2, 3, \dots, N$. N is the total number of data records present in the dataset. The dependent variable of each data record, y_i , is the prediction target. That prediction is achieved by accurately fitting a regression function $y = f(x)$. The target values are approximated by SVR using a learning function expressed as Eq. (12).

$$f(x, \omega) = \omega\phi(x) + b \quad (12)$$

where $f(x)$ = SVR target prediction; $\omega \in R^n$ = the weight vector; $b \in R$ = the threshold (bias); and, $\phi(x)$ = the high-dimensional feature space mapping from low-dimensional space x . Coefficients ω and b are derived by minimizing of the regularized risk function (Eq. 13).

$$\frac{1}{2}\omega^2 + C\frac{1}{l}\sum_{i=1}^n L_\epsilon(y_i, f(x_i, \omega)) \quad (13)$$

where $\|\omega\|^2$ = smoothness or flatness of the function; C = a regularization factor which is the measure of the trade-off extent between the model flatness and empirical error; and $\frac{1}{l}\sum_{i=1}^n L_\epsilon(y_i, f(x_i))$ = an empirical error that is measured using an ϵ -insensitive loss function expressed by Eq. (14).

$$L_\epsilon(Y_i, f(x_i, \omega)) = \begin{cases} 0, & y_i - f(x_i, \omega) \leq \epsilon \\ y_i - f(x_i, \omega) - \epsilon, & \text{otherwise} \end{cases} \quad (14)$$

where the range of ϵ values is defined (Eq. 10) such that the loss is equal to zero when the value predicted falls within the range, while the loss is equal to the difference between ϵ of the range and predicted value when that predicted value does not fall within the range.

The ω and b coefficients can be estimated using transformation of the regularized risk function (Eq. 9) into the original objective function (Eq. 15) by introducing two positive constants, ξ and ξ^* , also known as slack variables.

$$\begin{aligned} & \text{Min} \frac{1}{2}\omega^2 + C\frac{1}{l}\sum_{i=1}^l (\xi_i - \xi_i^*) \\ & \text{Subjected to} \begin{cases} y_i - f(x_i, \omega) - b \leq \epsilon + \xi_i \\ f(x_i, \omega) + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* > 0 \end{cases} \end{aligned} \quad (15)$$

This convex optimization expression (Eq. 11) can be solved by applying a Lagrangian multiplier strategy (Shao et al. 2020; Smola and Schölkopf 2004; Rui et al. 2019; Vapnik 2013; Ahmad et al. 2020). Equation (16) defines the transformed equation with these multipliers.

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) (\phi(x_i) \cdot \phi(x)) + b \quad (16)$$

in which the α_i and α_i^* represent the Lagrangian multipliers. The input vectors x_i that correspond to $\alpha_i - \alpha_i^* = 0$ are called support vectors established using the variable information associated with the training subset's data records. An appropriate kernel function also needs to be defined in order to overcome the computation complexity involving high-dimensional space (Rui et al. 2019). Any function that satisfies Mercer's condition might be employed as a kernel function (Vapnik 2013). The four kernel functions commonly used with SVR (Smola and Schölkopf 2004; Vapnik 2013) are listed in Table 3.

Among all the kernel functions applied, the radial basis function (RBF) kernel, known also as the Gaussian kernel, is the most common kernel applied with SVM and possesses an ability of anti-interference to noise in data (Kuo et al. 2013; Liu and Xu 2014; Wu et al. 2018; Vapnik et al. 1996; Hashemitaheri et al. 2020). In this study, the RBF kernel is used with the SVR algorithm to predict two-phase flow rate (Q_l) through a wellhead choke. The control parameter values applied to that SVR model are listed in Table 4.

Table 3 Kernel functions commonly used with SVM and SVR algorithms

Kernel function	Mathematical expression	Definition of parameters
Polynomial	$K(x, x_i) = \left(t + \frac{x_i^T x}{c}\right)^d$	d = degree of polynomial t = intercept
Sigmoid	$K(x, x_i) = \tanh(kx_i^T x + \theta)$	k = scale parameter θ = bias parameter
Radial basis function (RBF)	$K(x, x_i) = \exp\left(-\frac{\ x_i - x_j\ ^2}{2\sigma^2}\right)$	σ^2 = variance of RBF (Gaussian) kernel
Linear	$K(x, x_i) = x_i^T x$	

Table 4 SVR wellhead choke flow model control parameter values

Parameters	Status
Kernel function	RBF
ϵ range	0.1
C range	100,000
Cross-validation	Not applied
γ range (RBF)	0.05

- Distinguishing variables between input (or attribute) variables and dependent (or target) variables.
- Dividing the data records at “child” nodes, based on defined rules, using a splitting algorithm that assesses the attribute variables.
- Conducting further splitting, whereby each child node acts as parent node from which further nodes and decision-tree layers are generated.

Decision tree

Decision trees are a popular machine-learning method widely applied to evaluate diverse datasets, both classification and regression (Nie et al. 2011; Tsai and Chiou. 2009; Lorena et al. 2007). Groups of data records are organized into a hierarchical structure consisting of nodes and ramifications governed by a set of rules. This makes them applicable to classification and numerical (regression) datasets, although they are more widely used for classification problems (Lorena et al. 2007; Osei-Bryson 2004; Ortuño et al. 2015). For solving classification problems, a class label is given to each leaf of the tree with the data classes discriminated by rules and allocated to specific leaves. Three steps are involved in constructing decision trees for machine learning applications:

A simple decision tree is illustrated in Fig. 6 distinguishing the different types of nodes involved. From the root node (top layer) the entire dataset of data records is initially split into two subsets, which become the child (or decision) nodes forming the second layer of the tree. The child nodes are then divided further into sub-nodes, forming additional layers of child nodes, and ultimately reaching a final layer with terminal nodes (or leaves). A splitting algorithm is applied to ensure homogeneity (consistency) of the data records allocated to each child node. The data splits applied at each node are established by the algorithm to achieve the highest degree of homogeneity at each sub-node. Clear distinguished (homogeneous) groups of data records are established in the layer with the terminal nodes.

The tree is developed (layers added) until the training subset data records are perfectly classified by the tree so that they are allocated to the correct leaf with 100% accuracy. However, when the trained decision tree is applied to another

Fig. 6 Typical structure of a decision-tree machine-learning model

dataset (e.g., an independent testing subset) it is typically unable to classify those data records perfectly, indicating overfitting by the initial training process. The more layers and nodes involved in a decision tree the greater the risk of overfitting (Czajkowski and Kretowski 2016; Liu et al. 2016; Fakhari and Moghadam 2013).

In this study, a the scikit learn (sklearn) decision tree module (Scikit-Learn 2020a, b) is coded in Python applying the “gini” criterion to establish feature importance and the “best” splitter is applied to decide which feature and the value of the threshold to apply in making each split (Table 5). This is used to predict the two-phase flow rate (Q_l) through a wellhead choke.

Random forest

The random forest algorithm represents an extension to the decision tree algorithm, as it constructs and develops multiple decision trees to evaluate. It is a supervised, predictive algorithm suitable for the classification and regression applications based on training and testing subsets of data records for which the input and dependent variables distinguished (Zhou et al. 2020; Grape et al. 2020). The multiple decision trees are constructed in parallel, each using relatively few layers/nodes. This approach reduces the likelihood of overfitting, compared to algorithms based on individual decision trees. It also decreases the variance and bias of the prediction results, without compromising decision accuracy, by assessing the predictions of all decision trees collectively (Ahmad 2018; Breiman 2001).

To train random forest ML models, subsets of data records are bootstrapped (i.e., randomly samples with replacement) from the full dataset. Each subset of bootstrapped data records can then be used to develop an unpruned classification or regression decision tree. Not all input variables (M) are used in the development of each tree as candidates for splitting; rather, a small, randomly selected number of the available input variables (G) are selected randomly for each tree and then used for the splitting operation. Multiple decision trees are built iteratively until a specified number of

trees (K) are available. In regression (numerical) problems, predictions of the dependent variable are made through aggregation of predictions (bagging) from all the individual regression trees developed. The bagging processing reduces complexity of the individual trees and tends to diminish the chances overall of the algorithm overfitting the training subset. The prediction function for the random forest algorithm is expressed as Eq. (17).

$$\hat{f}_{RF}^K(x) = \frac{1}{K} \sum_{k=1}^K T_i(x) \quad (17)$$

where K = number of individual regression trees; X = input variable vector; and, $T_i(x)$ = prediction from a single regression tree for the i^{th} data record.

An out-of-bag error estimation (OOB) is progressively performed as the forest of individual regression trees is constructed. OOB is derived using the unselected data records (the OOB subset) as a test for the k th tree once it is trained in the bagging process. The OOB subset provides a progressive unbiased estimation of generic prediction error prior verifying the prediction accuracy of the aggregated results using the independent testing subset. The relative importance of each input variable, in relation to the dependent variable predictions, can be assessed from the aggregated results. This is helpful in reducing dimensionality of the model by eliminating low-contribution input variables to enhance the model's efficiency on datasets with high dimensionality (Ahmad et al. 2017). There are various ways to achieve this. One is to switch two input variables in the tree solutions, while keeping the remaining variables constant, and recording the mean reduction in the prediction accuracy. This makes it possible to rank the relative importance of each input variable to prediction accuracy achieved for the dependent variable (Ahmad 2018). The random forest algorithm is illustrated in Fig. 7.

In this study, a regression random forest model, as described, was established using the Scikit Learn Random Forest Regressor (Scikit-Learn 2020a, b) to predict two-phase flow rate (Q_l) through a wellhead choke. The control parameters for the random forest model are listed in Table 6.

Table 5 Decision tree wellhead choke flow regression model control parameter values

Parameters	Value
Maximum depth	100
Criterion	gini
Splitter	best
Objective function	Mean squared error (MSE)
Example prediction time	0.012925 (s)

Artificial neural network (ANN)

ANN is one of the most widely used ML algorithms for solving nonlinear problems across many fields of engineering (Shahbaz et al. 2019) despite their reputation of acting black boxes due to their hidden layers of regression-like calculations (Wood 2018). There are several types of ANN algorithms applied with feed-forward neural network (FFNN) composed of single-hidden-layer and multi-hidden-layer perceptron's (MLP) both being popular. In this

Fig. 7 Schematic diagram of generic configuration of random forest algorithm

Table 6 Random forest wellhead choke flow regression model control parameter values

Parameters	Value
Maximum depth	1000
Random state	0
Number of decision trees	1000
Objective function	Mean squared error (MSE)
Example prediction time	7.352307 (s)

study, a single-hidden-layer ANN (Fig. 8) is constructed for predicting two-phase flow rate (Q_i) through a wellhead choke.

The information passed from the neurons of one layer to the neurons of the next layer in the ANN is adjusted by weight and bias vectors. The neurons in the hidden layer perform information processing and send processed signals forward to the output layer adjusted by an activation function. Equation (18) expresses the signal adjustments made as information flows through the ANN

$$y_i = f\left(\sum_{j=1}^n W_{ij}x_j + b_i\right) \quad (18)$$

where f = the activation function; b_j = the hidden layer bias; x_i = input for the i^{th} variable i ; and, W_{ij} = connection weight between the i^{th} input and j^{th} neuron.

The neural network is trained, typically with a backpropagation algorithm, to improve its prediction performance, by adjusting the weights and bias values applied to the hidden layer. It does this by minimizing the mean squared error between actual and predicted values collectively for all data records the training subset, as defined by Eq. (19).

$$E_{MSE} = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2 \quad (19)$$

where $\hat{y}_i$ = actual value of dependent variable for the i^{th} data record; y_i = predicted value of dependent variable for the i^{th} data record; and m = the number of data records in the training subset.

The prediction performance and convergence speed/efficiency of ANN can often be improved by applying alternative optimization algorithms to backpropagation. Many

Fig. 8 Indicative structure of a feed-forward artificial neural network (ANN) with a single hidden layer with multiple neurons

optimizers have been applied for this purpose including Adam, RMSprop, Adagrad, Adadelta, Momentum, and Nesterov Accelerated Gradient. In this study, the RMSprop optimizer is applied, because it is an enhanced gradient-based algorithm that progressively divides the calculated gradient by a rolling average of recently attained gradient values (Selvam 2018). The initial learning rate of RMSprop is updated for different weights according to Eq. (20) and (21).

$$E[g^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2 \quad (20)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + f}} g_t \quad (21)$$

where $E[g^2]_t$ = rolling average gradient at iteration t ; g_t = gradient of objective function at iteration t ; θ_t = objective function value at iteration t ; f = smoothing term (avoiding division by zero); and η = learning rate, for which a default value of 0.001 is typically applied.

Equation (21) shows that the learning rate η is divided by exponentially decaying rolling average of squared gradients (Kartal and Özveren 2020).

The network structure of the single-hidden-layer ANN model has on input layer that includes the same number of neurons as input variables, one output layer with 1 neuron for the dependent variable prediction, and one hidden layer

with 500 neurons. The model is configured and executed in Keras (Keras 2020), a deep learning package coded in Python, running on top of the machine learning platform TensorFlow (TensorFlow 2020). The control parameter values of the constructed ANN model are listed in Table 7.

Certain control parameters involved require some clarification:

Table 7 Control parameters for the one-hidden layer ANN model constructed to predict two-phase flow rate (Q_t) through wellhead choke

Control Parameters	Status
Number of hidden layers	1
Number of neurons in the hidden layer	500
Activation function used input to hidden layer	SELU (Scaled Exponential Linear Unit)
Activation function used hidden to output layer	SELU
Objective function minimized for training subset	MSE
Optimization algorithm	RMSprop
Minimum delta	0.01% of Q_L mean value
Patience (number of iterations)	25
Number of iterations	237
Learning rate	0.01

min_delta This is used to define the algorithm's stopping criteria. The minimum change in the objective function qualifies as an improvement. This depends on the scale of the dependent variable and will be dataset specific.

Patience: This is also used in relation to the algorithm's stopping criteria number of iterations with no improvement after which training of the algorithm is terminated.

Activation function: The Scaled Exponential Linear Unit (SELU) function is applied as it facilitates self-normalization. By internalizing normalization, it tends to speed up convergence.

Learning rate: It controls the magnitude of change to apply to the model in each iteration based on magnitude of the prediction error. This value is used to adjust the values of the weights and biases.

Selection of suitable values for “min-delta” and “patience”, and identifying a high-performing activation function, requires some trial-and-error analysis with each dataset as the optimum selection depends on the character of the data records.

A flow diagram summarizing the sequence of steps through which the ANN progresses to make its dependent-variable predictions is displayed in Fig. 9.

Deep neural network (DNN)

Deep neural network (DNN) or deep learning (DL) algorithms represent a further development of ANN, containing multiple hidden layers each typically configured with a large number of neurons. A typical structure of DL algorithm containing three hidden layers is illustrated in Fig. 10. As with ANN, the hidden layers contain a specified number of neurons, with weights, biases, and activation functions applied. In the DL algorithm, each layer's output is communicated as input for the next layer by applying Eq. (22) beginning with the input layer of the network where network is $h^0 = x$ (Lee et al. 2018).

$$h^k = (W^k h^{k-1} + b^k) \sigma^k \quad (22)$$

where k = number of DNN network layers; h^k = the array of output for the k layer; W^k = neuron weights for the k layer; b^k = bias for the k layer; and, σ^k = activation function for the k layer.

The final layer's output, $\hat{y}$, is the dependent variable's prediction. The purpose of using the activation functions is to emphasize the nonlinearity of relationships between the input and output variables (Taqi et al. 2018). Key characteristics of activation factors are that they should be differentiable making them easier to apply particularly with gradient-based optimizers (Bengio 2009; Asghari et al. 2020; Nwankpa et al. 2018). The activation functions that are widely used with DNN are listed in Table 8.

Fig. 9 Flow diagram of single layer ANN built for prediction of QL through wellhead chock

In DL algorithms, the dissimilarity (error) between the predictions and the actual values of dependent variables is minimized by conducting training with a training dataset of substantial size. An objective function that represents this dissimilarity is defined to train DL algorithms. The objective functions commonly used to train DL algorithms on a dataset with n samples are listed in Table 9.

From the outset of the training process, the values of the weights applied to neurons and biases applied to each layer of the DL network are initialized to each have values of 1. The DL model begins by making predictions based on random changes to the initial weights. Then, using an optimization algorithm, the weight and bias values are iteratively updated in order to minimize the objective function.

The typical control parameters applied to configure DL models include the number of layers, number of neurons within each layer, objective function used for training, activation function applied to each layer, optimization algorithm, minimum delta, patience (i.e., early stopping criteria), learning rate, and split percentages between training and testing subsets. Most of these control parameters need to be established by trial and error to suit specific data sets (Ng 2016). The control parameters selected for the developed DL model are listed in Table 10.

The sequence of steps through which the DL progresses to make its dependent-variable predictions is the same as that described for the ANN model (Fig. 9).

Fig. 10 Indicative structure of a deep learning neural network (DNN) with three hidden layers with multiple neurons

Table 8 The most common activation functions

Activation function	Formula
Sigmoid	$s(x) = \frac{1}{1+e^{-x}}$
tanh	$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
Relu (Rectified linear unit)	$y(x) = \max(0, x)$
Selu (Scaled exponential linear unit)	$y(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ ae^x - \alpha & \text{if } x \leq 0 \end{cases}$
Leaky-relu	$y(x) = \max(ax, x) : 0 < a \ll 1$
Linear	$y(x) = x$

Sorush oil field 10-well dataset evaluated

Data collection

The prediction of two-phase flow rate through wellhead chokes is conducted using a dataset of 7245 data records from 10 production wells drilled in the Sorush oil field (well numbers SR#17 to SR#26) and valorization by

113 datapoint of 12 well from South of Iran collected in Choubineh et al. 2017 article. This oil field is located in the Persian Gulf offshore southwest Iran in Bushehr province, 83 km southwest of Kharg Island (Fig. 11). Sorush oil field was discovered in 1962 with its discovery well achieving maximum daily production of 14,000 barrels. The Sorush oil field extends over an area of about 260 square kilometers and with more than 15 billion barrels of oil in place.

The wellhead data measurements from the ten Sorush oil field wells and other data provided by Choubineh et al. 2017 include: choke size (D_{64}) is measured in 1/64 in, wellhead pressure (P_{wh}) measured in pounds per square inch gauge (psig), oil specific gravity (γ_o) relative to water, gas to liquid ratio (GLR) measured in Scf/STB and two-phase flow rate (Q_L) measured in stock-tank barrels per day. The first four represent input variables, and the last represent the dependent variable for the ML and DL models (Table 11). Values for the input and dependent variables for all data records are made available in an Excel supplementary file (Appendix 1).

Table 9 Common objective functions minimized during the training process of deep learning (DL) models

Objective function	Abbreviation	Formula
Mean of squared error	MSE	$J(y, \hat{y}) = \frac{1}{n} \sum_i (\hat{y}_i - y_i)^2$
Mean of absolute error	MAE	$J(y, \hat{y}) = \frac{1}{n} \sum_i (\hat{y}_i - y_i)$
Mean of absolute error%	MAPE	$J(y, \hat{y}) = \frac{1}{n} \sum_i \frac{ \hat{y}_i - y_i }{y_i} * 100$
Mean of squared logarithmic error	MSLF	$J(y, \hat{y}) = \frac{1}{n} \sum_i (\log(\hat{y}_i + 1) - \log(y_i + 1))^2$

Table 10 Control parameters for the multi-layer DL model constructed to predict two-phase flow rate (Q_L) through wellhead choke

Control Parameters	Status
Number of hidden layers	3
Number of neurons in hidden layers, 1, 2, and 3	500, 100, 50
Activation function used in input / hidden layers	SELU
Activation function applied to output layer	SELU
Objective function used training	MSE
Optimization algorithm	RMSprop
Minimum delta	0.01% of Q_L mean value
Patience (number of iterations)	25
Number of iterations	418
Learning rate	0.1

Variable data analysis

Cumulative distribution functions (CDF) for the five dataset variables are displayed in Fig. 12 providing insight to the form of distributions displayed by these variables across all 7245 data records in Soroush Oil field. The percentage of the CFD values less than a specific variable value is calculated according to Eq. (23).

$$F_X(x) = P(X \leq x), \text{ for all } x \in R \quad (23)$$

where P = percentage of data records with values in a distribution less than a specific data record; x = data variable value range; X = the value of variable x in a specific data record; and R = the dataset of data records.

Fig. 11 Map showing the location of the Soroush oil field, offshore southwest Iran**Table 11** Data record statistical characterization of the variables in this study

Dataset Variables for Prediction Two-phase Flow Rate from Wellheads in the oil Fields						
Field	Variables	Wellhead Choke Diameter	Wellhead Pressure	Oil Specific Gravity	Gas to Liquid Ratio	Two-Phase Flow Rate
	Symbol	D64	Pwh	Yo	GLR	QL
	Units	inch/64	psig	%	Scf/STB	STB/day
7245 dataset records from Soroush Oil Field	Mean	65.7	319	1.00	135.3	11,667
	Std. Deviation	15.2	169	0.06	42.9	3625
	Variance	231.5	28,718	0.00	1840.1	13,137,230
	Minimum	33.8	131	0.93	3.0	660
	Maximum	111.9	1024	1.07	217.0	23,700
113 dataset records from Oil Field of South Iran	Mean	54.5	1280	0.86	858.6	8549
	Std. Deviation	16.9	349	0.02	440.3	5376
	Variance	282.4	120,507	0.00	192,106.3	28,641,000
	Minimum	24.0	50	0.81	107.0	1324
	Maximum	80.0	2940	0.92	3660.0	22,150

Fig. 12 The cumulative distribution function (CDF) of the Soroush field dataset variables for the prediction of two-phase flow rate through wellhead choke. The variables are choke size (D_{64}), wellhead pressure (P_{wh}), oil-specific gravity (γ_o), gas–liquid ratio (GLR), and two-phase flow rate through wellhead choke (Q_L ; dependent variable).

The thinner blue line represents the cumulative distribution functions for the data variables. The thicker red line represents a normal distribution with the same mean and standard deviation across all 7245 data records in Soroush Oil field

Inspection of the CFDs (Fig. 12) reveals the following: for

- D_{64} For ~13% of the data records $D(1/64in)$ are <48.64 in. For ~76% of the data records $48.64 < D_{64} < 84.22$ in. For ~11% of the data records $D(1/64in) > 84.22$ in. The D_{64} data approximately follows a normal distribution with a slightly negative skew.
- P_{wh} For ~29% of the data records $P_{wh} < 218.59$ Psig. For ~62% of the data records $218.59 < P_{wh} < 553.03$ Psig. For ~11% of the data records $P_{wh} > 553.03$ Psig. The P_{wh} data deviate somewhat from a normal distribution and are more negatively skewed than the choke diameter distribution and display a proportionally higher standard deviation.
- γ_o For ~45% of the data records $\gamma_o < 0.9293$. For ~38% of the data records $0.9293 < \gamma_o < 1.05795$. For ~17% of the data records $\gamma_o > 1.05795$. The γ_o data do not follow

a normal distribution but are approximately symmetrical and display a relatively low standard deviation in relation to its mean value.

- GLR For ~39% of the data records $GLR < 118$ Scf/STB. For ~38% of the data records $118 < GLR < 150$ Scf/STB. For ~24% of the data records $GLR > 150$ Scf/STB. The GLR data follow a normal distribution quite closely displaying a moderate standard deviation.
- Q_L For ~49% of the data records $Q_L < 11,623$ STB/D. For ~42% of the data records $11,623 < Q_L < 16,710$ STB/D. For ~9% of the data records $Q_L > 16,710$ STB/D. The Q_L data approximately follow a normal distribution and are essentially symmetrical with a moderate standard deviation.

Of the input variables, γ_o and P_{wh} are the most asymmetrical and distributed least like normal distributions. The characteristics of the variable distributions are shown in

more detail in Fig. 13, with the relative shapes of the distributions highlighting that these variables are nonlinearly related.

Measurements to determine prediction errors

The entire dataset of 7245 data records in Soroush oil field is divided for each case run into a training subset (80%) and a training subset (20%). Statistical analysis of the prediction errors for the two-phase flow rate through choke (Q_L) associated with each of the ML and DL algorithms is conducted using the following widely used metrics: percentage deviation (PD), average percentage deviation (APD), average absolute percentage deviation (AAPD), standard deviation (STD), mean squared error (MSE), root-mean-squared error (RMSE), and coefficient of determination (R^2). The computation formulas for these statistical measures are expressed in Eq. (24–31).

Percentage difference (PD):

$$PD_i = \frac{\xi_{(\text{Measured})} - \xi_{(\text{Predicted})}}{\xi_{(\text{Measured})}} \times 100 \quad (24)$$

Average percent deviation (APD):

$$APD = \frac{\sum_{i=1}^n PD_i}{n} \quad (25)$$

Absolute average percent deviation (AAPD):

$$AAPD = \frac{\sum_{i=1}^n |PD_i|}{n} \quad (26)$$

Standard deviation (SD):

$$SD = \sqrt{\frac{\sum_{i=1}^n (PD_i - Dimean)^2}{n - 1}} \quad (27)$$

$$Dimean = \frac{1}{n} \sum_{i=1}^n (\xi_{\text{Measured } i} - \xi_{\text{Predicted } i}) \quad (28)$$

Fig. 13 Variable value versus dataset index number highlighting the ranges and extreme values associated with each variable recorded for the Soroush oil field dataset. The variables displayed are: choke size

(D_{64}), wellhead pressure (P_{wh}), oil-specific gravity (γ_o), gas to liquid ratio (GLR), and two-phase flow rate through wellhead choke (Q_L) across all 7245 data records in Soroush Oil field

Mean square error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (\xi_{Measured\ i} - \xi_{Predicted\ i})^2 \quad (29)$$

Root-mean-square error (RMSE):

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n}} \quad (30)$$

where n = number of data records; x_i = measured dependent variable value for the i^{th} data record; and y_i = predicted dependent variable value for the i^{th} data record.

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^N (\xi_{Predicted\ i} - \xi_{Measured\ i})^2}{\sum_{i=1}^N \left(\varphi_{Predicted\ i} - \frac{\sum_{i=1}^n \xi_{Measured\ i}}{n} \right)^2} \quad (31)$$

Results

Two-phase flow rate prediction accuracies achieved by ML, DL, and traditional mathematical models

In order to make comparisons between ML, DL, and traditional empirical methods for two-phase flow rate (Q_L) prediction, and to determine the best performance accuracy achieved for all 7245 data records in Soroush oil field, the following method is applied:

Two-phase flow rate (Q_L) prediction accuracies achieved by the training subset (~80%), the testing subset (~20%), and the complete dataset (5796 data records) are presented in Tables 12, 14, respectively. Analysis of these results reveals that the relative performance accuracy of ML, DL, and

traditional empirical methods can be ranked in the following descending order: DL > ML > traditional empirical methods, respectively. The deep learning model achieves the highest Q_L prediction accuracy for the three sample sets evaluated: RMSE < 195.90 STB/D and R^2 = 0.9969 for the testing subset; RMSE < 161.23 STB/D and R^2 = 0.9981 for the training subset; and, RMSE < 144.21 STB/D and R^2 = 0.9978 for the total dataset.

Among the empirical models considered, the Pilehvari model achieved the best relative Q_L performance accuracy: RMSE < 1522.6 STB/D and R^2 = 0.7012 for the testing subset; RMSE < 1507.1 STB/D and R^2 = 0.7331 for the training subset; and, RMSE < 1510.1 STB/D and R^2 = 0.7178 for the total dataset.

Analysis of results

Figure 14 and Tables 12, 13, 14 display RMSE and R^2 comparisons of performance accuracy of ML, DL, and traditional mathematical models. Analysis of these results confirms that the Q_L prediction accuracy of the DL model substantially outperforms the other models considered. The poorer Q_L prediction accuracy of the empirical models identifies that they can be ranked in descending order of accuracy, as follows: Pilehvari > Baxendell > Ros > Gilbert > Achong.

Close inspection of the model's prediction results (Tables 12, 13, 14) reveals that the deep learning model achieves exceptionally high Q_L prediction accuracy (RMSE = 195.9 STB/D; AAPD = 1.025% for the testing subset) and is substantially more accurate than the four ML models applied to the Soroush field dataset. The near-perfect R^2 value is achieved due to the large number of closely spaced samples in the dataset. Figure 15 displays the predicted versus measured Q_L values for each data record in each subset evaluated by the five models. The Q_L prediction performance

Table 12 Q_L Prediction performance compared for ML, DL, and traditional mathematical (empirical) models applied to the training subset (5796 data records from Soroush oil field) for the Soroush field dataset of wellhead choke recordings

Two-phase Liquid Flow Rate Prediction Accuracy for the Training Subset (5796 Data Records; ~80%)							
Models		APD	AAPD	SD	MSE	RMSE	R^2
Units		(%)	(%)	(STBD)	(STBD)	(STBD)	
SVR	Support Vector Regression	-0.613	1.748	671.1	450,304	671.0	0.9663
DT	Decision Tree	-0.307	1.383	560.7	314,388	560.7	0.9765
RF	Random Forest	-0.375	1.630	590.7	359,597	599.7	0.9731
ANN	Artificial Neural Network	0.028	5.388	701.2	498,963	706.4	0.9641
DL	Deep Learning	-0.013	0.690	161.3	25,995	161.2	0.9981
Mathematical Models	Gilbert	3.649	12.763	1562.4	4,865,742	2205.8	0.6742
	Baxendell	2.159	11.620	1004.3	3,992,151	1998.0	0.7193
	Ros	2.882	11.870	1620.7	4,061,329	2015.3	0.6915
	Achong	0.619	11.746	1705.1	5,614,888	2369.6	0.6430
	Pilehvari	-0.169	9.845	1218.0	2,271,312	1507.1	0.7331

Fig. 14 Two-phase flow rate (Q_L) prediction accuracy expressed in terms of R^2 and RMSE and compared for the full dataset (7245 data records) of Soroush oil field data records for all the prediction methods evaluated

Table 13 Q_L Prediction performance compared for ML, DL, and traditional mathematical models applied to the independent testing subset (1449 data records from Soroush oil field) for the Soroush field dataset of wellhead choke recordings

Models units		APD (%)	AAPD (%)	SD (STBD)	MSE (STBD)	RMSE (STBD)	R^2
SVR	Support Vector Regression	−0.396	1.907	599.8	372,144	610.0	0.9698
DT	Decision Tree	0.008	1.795	579.8	347,930	589.9	0.9720
RF	Random Forest	−0.027	1.804	591.6	362,341	601.9	0.9707
ANN	Artificial Neural Network	−0.287	5.416	644.4	434,714	659.3	0.9658
DL	Deep Learning	−0.043	1.025	192.6	38,378	195.9	0.9969
Mathematical Models	Gilbert	3.480	12.813	2248.8	4,802,960	2191.6	0.6620
	Baxendell	1.941	11.770	1494.7	4,021,592	2005.4	0.6953
	Ros	2.678	11.973	1736.4	4,005,564	2001.4	0.6827
	Achong	−0.224	12.143	1480.3	6,242,742	2498.5	0.6126
	Pilehvari	−0.304	10.272	1239.6	2,318,273	1522.6	0.7012

Table 14 Q_L Prediction performance compared for ML, DL, and traditional mathematical models applied to the entire dataset (7245 data records from Soroush oil field) for the Soroush field dataset of wellhead choke recordings

Models Units		APD (%)	AAPD (%)	SD (STBD)	MSE (STBD)	RMSE (STBD)	R^2
SVR	Support vector regression	−0.570	1.780	659.3	434,679	659.3	0.9670
DT	Decision tree	−0.245	1.780	566.6	321,106	566.7	0.9756
RF	Random forest	−0.306	1.665	600.1	360,155	600.1	0.9726
ANN	Artificial neural network	−0.038	5.396	697.9	487,019	697.9	0.9644
DL	Deep learning	−0.019	0.757	168.8	28,481	168.8	0.9978
Mathematical Models	Gilbert	3.620	11.482	2196.0	4,370,300	2090.5	0.6631
	Baxendell	2.120	11.024	1259.3	3,959,439	1989.8	0.7244
	Ros	2.561	11.781	1681.5	4,052,142	2013.0	0.6894
	Achong	0.449	12.419	1579.0	5,740,491	2395.9	0.6467
	Pilehvari	−0.211	10.392	1230.6	2,280,496	1510.1	0.7178

Fig. 15 Two-phase flow rate (Q_L) prediction versus measured value for each data record in the training and testing subsets and the full dataset evaluated for the Soroush oil field wellhead measurements related to all 7245 data records from Soroush oil field

of the five proposed models enables them to be ranked in the following order: DL > DT > RF > ANN > SVR.

The Q_L prediction results for each model applied to the full dataset (7245 data records) can be compared more closely by superimposing their results on a single predicted versus measured Q_L graph (Fig. 16). This reveals that the predictions from the ML models demonstrate significant scatter in comparison with the ML model that closely follows the $X = 1$ line with much less dispersion in Fig. 15. In detail, the ML models show a tendency to overestimate at the lower end of the Q_L range ($< \sim 7500$ STB/D) and underestimate at the upper end of the Q_L scale ($> 20,000$ STB/D). That tendency is particularly apparent for the RF and ANN models and partly explains their lower accuracy performance overall.

Figure 17 compares the prediction performance of the models on a sample-by-sample basis for all data records in terms of relative prediction error (PD%; Eq. 20) for both the training and testing subsets for the Soroush oil field. Whereas the DL model shows consistently low relative percentage errors across the sample index range, the ML models show some high PD values at various points throughout the sample index range. The higher prediction errors are on the negative side (i.e., underestimation of Q_L). The ANN model shows more PD errors more negative than -2% than the other models (Fig. 16). On the other hand, the SVR, DT and RF models show a PD% for one data record more negative than -5%. In terms of their Q_L values both subsets of data records are spread across the entire Q_L value range. They are displayed sequentially in Fig. 17 for illustrative purposes only. The range ($-0.14762\% \geq < 0.142185\%$) of relative percentage errors achieved by the DL model is almost an order of magnitude less than the other four models, emphasizing the

superiority of its Q_L predictions for the Soroush oil field dataset evaluated.

The histograms included in Figs. 18 display the Q_L prediction errors for the ML and DL algorithms. The superiority of the DL model is clear in terms of the lower prediction errors it generates. The best performing ML model is DT which consistently outperforms the RF model. On the other hand, the SVR and ANN ML models consistently show poorer Q_L prediction performance than the other ML models when applied to this dataset.

Figure 19 distinguishes critical flow from subcritical flow regimes for the flow measured through the wellhead chokes of the wells in the Soroush oil field dataset. If the downstream pressure to upstream pressure ratio is ≤ 0.588 , the flow regime is identified as critical. On the other hand, if this ratio > 0.588 , the flow regime is identified as subcritical. Analysis of Fig. 19 clearly reveals that this ratio is ≤ 0.588 for almost all the data records evaluated for the Soroush oil field. Therefore, the flow through the reducers related to these 10 wells can be confirmed as conforming to a critical flow regime.

Development and generalization of the deep learning model

The DL model developed in previous section has addressed solely the Soroush oil field (10 production wells) dataset. In order to evaluate the accuracy of the DL algorithms for general application to other oil fields, a published dataset of 113 data records from other fields (Choubineh et al. 2017) has been evaluated. This independent dataset includes data records from twelve oil wells located in South Iran. Statistical measures of Q_L prediction accuracy achieved for these data records are displayed in Table 15. A Comparison of the results of Table 15 with those displayed in Tables 12, 13, 14 confirms the high two-phase flow rate-prediction accuracy achieved by the developed DL model, trained with Soroush oil field data, when applied to other wells from other oil field.

Figure 20 displays the actual versus the Q_L values predicted by the DL model trained with Soroush oil field training subset (5796 data records), applied to the complete datasets available for 12 oil wells from in South Iran. The performance accuracy achieved by this model confirms its reliability for application across the for 12 oil wells from South Iran. The method used is clearly suitable for application to other fields. However, it would be prudent to initially test the Soroush-calibrated DL model with some direct Q_L measured data for the specific fields to which it is applied in order to establish whether some recalibrations are required.

Figure 21 reveals the prediction performance of the Soroush-trained DL model applied to the 113 independent data records available for 12 oil wells from South Iran.

Fig. 16 Two-phase flow rate (Q_L) prediction versus measured value for each data record in the full dataset evaluated for the Soroush oil field wellhead measurements related to 7245 data records from Soroush oil field

Fig. 17 Two-phase flow rate (Q_L) prediction relative percentage errors compared for the training and testing subsets of the five prediction models evaluated for all 7245 data records from Soroush oil field

Those data records are displayed sequentially in Fig. 21 for illustrative purposes only. The range of relative percentage errors achieved for this independent dataset is $(-0.08611\% > = < 0.159105\%)$.

Discussion

It is useful to determine the relative influence of the input variables on the dependent variable (Q_L) values. Calculating the nonparametric Spearman's correlation coefficient (ρ) is a useful measure to establish this. ρ is expressed over the range -1 (perfect negative correlation) to 1 (perfect positive correlation), with a zero value indicating no correlation and

implying relatively low or no impact (Gauthier 2001). ρ is calculated for ranked data using Eq. (32).

$$\rho = \frac{\sum_{i=1}^n (O_i - \bar{O})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2 \sum_{i=1}^n (M_i - \bar{M})^2}} \quad (32)$$

where O_i = the value of data record i for input variable O ; $\bar{O}$ = the average value of the input variable O ; M_i = the value of data record i for input variable M ; $\bar{M}$ = the average of the input variable M ; and n = the number of data points in the population.

Figure 22 displays the p values for the relationships between Q_L and the four input variables considered in this

Fig. 18 Error (STBD) histograms with fitted normal distributions (red line) for DL and ML models used to predict Q_L for all 7245 data records from Soroush oil field

study. It is apparent that GLR has the greatest relative impact on Q_L , whereas D (1/64) has the least influence. On the other hand, falling between the two input variables mentioned, Pwh and γ_o have almost equal influence on Q_L , the former directly and the latter inversely.

Many ML and DL algorithms display a tendency to overfit their training subsets, particularly with datasets consisting of relatively small numbers of data records. The substantial number of data records (7245) involved in the Soroush oil field dataset and the allocation of 20% of them (1449) to the independent testing subset have helped to avoid overfitting in this study. A comparison of the Q_L prediction accuracy achieved by the training (Table 12) and testing (Table 13) subsets

indicates that overfitting has not impacted the prediction performance of any of the algorithms and traditional mathematical (empirical) models evaluated. Indeed, for the SVR and ANN a slightly higher prediction accuracy is achieved for the testing subset than for the training subset in terms of RMSE and R^2 . For the other algorithms the prediction accuracy achieved by the testing subset is only slightly worse than the training subset, and the dispersion and magnitude of errors recorded across both subsets for each algorithm are similar (Fig. 17). The results of Fig. 19 identify that a critical flow regime prevails through the Soroush field wellheads. Figures 20, 21, and Table 15 demonstrate that the DL mode

Fig. 19 Determination of the flow regimes in two-phase flow through wellhead chokes for the Soroush oil field dataset

developed for the Soroush oil field can be successfully applied to wells from other fields.

Although overfitting has not been a problem for the large Soroush oil field dataset, that is unlikely to be the case for smaller datasets for these algorithms, particularly those characterized by a large number of neurons, layers or nodes. All of the algorithms evaluated lack transparency in readily revealing the details of how each individual data record prediction is generated. Future progress with DL algorithms needs to focus on improving their transparency and developing more clarity in their trade-off between prediction accuracy, complexity and overfitting risks. Despite this lack of calculation transparency, this study highlights the high-prediction performance of the DL algorithm with respect to flow rates through wellhead chokes across a wide range of magnitudes and multiple wells in the Soroush oil field. These results suggest sufficient reliability associated with the DL predictions to justify its incorporation with automated flow measurement systems for individual wells in that field. With careful calibration and training and sufficient data records to avoid overfitting this method offers the potential to improve flow rate prediction through wellheads in other oil fields.

Fig. 20 Cross-plot of predicted versus measured Q_L values compared for the DL model trained with Soroush oil field training subset (5796 data records) and applied to the 113 data records available for 12 oil wells from South Iran

Fig. 21 Relative error (%) Q_L prediction by Soroush-trained DL model applied to the 113 data records available for 12 oil wells from South Iran

Conclusions

In total, 7245 data records compiled for ten production wells from the Soroush oil field wells (SR#17 to SR#26) offshore Iran demonstrate the comparative performance of four traditional machine learning (ML), traditional mathematical (empirical) models, and one deep learning (DL) algorithm in predicting flow rate through wellhead chokes. The DL technique is a new technique that has not previously been applied for two-phase flow rate (Q_L) prediction.

Table 15 Q_L prediction accuracy of the DL model, trained with Soroush oil field subset (5796 data records), applied to the complete datasets available for 12 oil wells from South Iran (113 data records) treating them an independent testing subset

QL prediction accuracy measures for DL model trained with Soroush oil field subset (5796 data records) for 12 oil wells from in South Iran						
Models		APD	AAD	SD	MSE	RMSE
Units		(%)	(%)	(STBD)	(STBD)	(STBD)
DL	Deep learning	0.018	0.572	178.5	30,909	175.8
						0.9983

Fig. 22 Relationships of the two-phase flow rate prediction through wellhead choke individually with its input calculation variables: upstream choke size (D_{64}), wellhead pressure (P_{wh}), oil-specific gravity (γ_o), and gas–liquid ratio (GLR)

The input variables choke size (D_{64}), wellhead pressure (P_{wh}), oil-specific gravity (γ_o), and gas–liquid ratio (GLR) are assessed by the algorithms to derive predictions of two-phase flow rate through the wellhead chokes (Q_L). The Q_L value ranges from 660 to 23,700 stock tank barrels/day (STB/day) for this dataset and shows nonlinear relationships with its influencing variables. Spearman's correlation coefficients reveal that the input variables GLR and γ_o are inversely related to Q_L , whereas input variables P_{wh} and D_{64} display positive correlations with Q_L . D_{64} shows the lowest correlation coefficient with Q_L of the four input variables evaluated. Prevailing flow through the wellhead chokes of the Soroush oil field conforms to a critical flow regime.

In addition to the deep learning neural network (DL), the ML algorithms evaluated are support vector regression (SVR), decision tree (DT), random forest (RF), artificial neural network (ANN). The traditional mathematical (empirical) models evaluated are those proposed by Gilbert (1954), Baxendell (1958), Ros (1960), Achong (1961), and Pilehvari (1981). The models are all used for two-phase flow rate prediction through wellhead choke for the Soroush oil field dataset. The large set of data records and the allocation of a substantial number of those records to an independent testing subset (1449 records or 20% of the entire dataset) enable the DL and ML algorithms to avoid overfitting with respect to this dataset. Based on six statistical measures of prediction accuracy, the DL algorithm substantially outperforms the four ML algorithms and five traditional mathematical models in predicting Q_L for this dataset.

The DL algorithm predicts Q_L for the testing subset with a root-mean-squared error (RMSE) of 196 STB/day and coefficient of determination (R^2) of 0.9969 (in Soroush oil field). On the other hand, the RMSE for the four ML algorithms is greater than 590 STB/day. In order to test the suitability of the DL model for general application to predict Q_L in other oil fields, 113 independent data records from 12 wells from fields from South Iran were evaluated with the Soroush-trained DL model. The RMSE and R^2 for DL applied to that independent well dataset are 175.

STB/day and 0.9983, respectively. These results suggest that the DL algorithm is a suitable choice for automating flow rate estimates through wellhead chokes, although for smaller datasets in other fields careful testing and calibration with measured Q_L data are required to eliminate the risk of overfitting.

Appendix 1

A supplementary Excel file containing the variable values for each of the 7245 data records from the ten wells from the Soroush oil field evaluated in this study is made available for readers to download.

Supplementary Information The online version contains supplementary material available at (<https://doi.org/10.1007/s13202-021-01087-4>)

Acknowledgments This research was supported by Tomsk Polytechnic University under Grant Number VIU-CPPSND- 214/2020.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest with respect to the content of this manuscript.

Ethical approval The authors have adhered to a strict code of ethics in preparing and submitting this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Achong, I. (1961). Revised bean performance formula for Lake Maracaibo wells. Internal Company Report, Shell Oil Co., Houston, TX, USA
- Ahmad MW, Mourshed M, Rezgui Y (2017) Trees vs Neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy Build* 147:77–89. <https://doi.org/10.1016/j.enbuild.2017.04.038>
- Ahmad MW, Reynolds J, Rezgui Y (2018) Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees. *J clean product* 203:810–821. <https://doi.org/10.1016/j.jclepro.2018.08.207>

- Ahmad MS, Adnan SM, Zaidi S, Bhargava P (2020) A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens. *Constr Build Mater* 248:118475. <https://doi.org/10.1016/j.conbuildmat.2020.118475>
- AlAjmi MD, Alarifi SA, Mahsoon A.H (2015) Improving multiphase choke performance prediction and well production test validation using artificial intelligence: a new milestone. In SPE digital energy conference and exhibition. 2015. Society of Petroleum Engineers. 8 pages. <https://doi.org/https://doi.org/10.2118/173394-MS>
- Al-Attar H (2008) Performance of wellhead chokes during sub-critical flow of gas condensates. *J Petrol Sci Eng* 60(3–4):205–212. <https://doi.org/10.1016/j.petrol.2007.08.001>
- Al-Attar H (2008) Performance of wellhead chokes during sub-critical. *J Petrol Sci Eng*. <https://doi.org/10.1016/j.petrol.2007.08.001>
- Asghari V, Leung YF, Hsu S-C (2020) Deep neural network-based framework for complex correlations in engineering metrics. *Adv Eng Inform* 44:101058. <https://doi.org/10.1016/j.aei.2020.101058>
- Atashnezhad A, Wood DA, Fereidounpour A, Khosravanian R (2014) Designing and optimizing deviated wellbore trajectories using novel particle swarm algorithms. *J Nat Gas Sci Eng* 21:1184–1204. <https://doi.org/10.1016/j.jngse.2014.05.029>
- Bairamzadeh S., Ghanaatpisheh EJSI (2015). A new choke correlation to predict liquid flow rate. *Sci. Intl (Lahore)* 27(1), pp. 271–274. www.academia.edu/download/40252273/2937730771_a_271-274--SINA-Chem_Engn--IRAN.pdf
- Barrow, H. 1996. Connectionism and neural networks. *Artificial Intelligence Handbook of Perception and Cognition* Elsevier. pp. 135–155. DOI: <https://doi.org/https://doi.org/10.1016/B978-012161964-0/50007-8>
- Baxendell P (1958) Producing Wells on Casing Flow-An Analysis of Flowing Pressure Gradients. *Trans AIME* 213(01):202–206. <https://doi.org/10.2118/983-G>
- Bengio Y (2009) Learning Deep Architectures for AI. *Found Trends Machine Learn* 2(1):1–127. <https://doi.org/10.1561/2200000006>
- Bonaccorso, G. 2017. *Machine learning algorithms*. Packt Publishing Ltd, Birmingham, U.K. pp 360. ISBN: 9781785889622
- Breiman L (2001) Random forests. *Machine Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- Brereton RG, Lloyd GR (2010) Support vector machines for classification and regression. *Analyst* 135(2):230–267. <https://doi.org/10.1039/B918972F>
- Cao L-J, Tay FEH (2003) Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans Neural Networks* 14(6):1506–1518. <https://doi.org/10.1109/TNN.2003.820556>
- Chong D et al (2009) Structural optimization and experimental investigation of supersonic ejectors for boosting low pressure natural gas. *Appl Therm Eng* 29(14–15):2799–2807. <https://doi.org/10.1016/j.applthermaleng.2009.01.014>
- Choubineh A et al (2017) Improved predictions of wellhead choke liquid critical-flow rates: modelling based on hybrid neural network training learning-based optimization. *Fuel* 207:547–560. <https://doi.org/10.1016/j.fuel.2017.06.131>
- Cortez C, Vapnik V (1995) Support vector networks. *Machine Learn* 20:273–297. <https://doi.org/10.1007/BF00994018>
- Czajkowski M, Kretowski M (2016) The role of decision tree representation in regression problems: an evolutionary perspective. *Appl Soft Comput* 48:458–475. <https://doi.org/10.1016/j.asoc.2016.07.007>
- Darvishpour A et al (2019) Wellbore stability analysis to determine the safe mud weight window for sandstone layers. *Petrol Explorat Develop* 46(5):1031–1038. [https://doi.org/10.1016/S1876-3804\(19\)60260-0](https://doi.org/10.1016/S1876-3804(19)60260-0)
- Drucker H et al (1997) Support vector regression machines. *Adv Neural Inf Process Syst* 28(7):779–784
- Elhaj MA, Anifowose F, Abdulraheem A (2015). Single gas flow prediction through chokes using artificial intelligence techniques. in SPE Saudi Arabia Section Annual Technical Symposium and Exhibition. Society of Petroleum Engineers. 14 pages. <https://doi.org/https://doi.org/10.2118/177991-MS>
- Ezzatabadipour M et al. (2017). Deep learning as a tool to predict flow patterns in two-phase flow. <https://arxiv.org/abs/1705.07117> [Accessed 27 Aug 2020]
- Fakhari A, Moghadam AME (2013) Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. *Appl Soft Comput* 13(2):1292–1302. <https://doi.org/10.1016/j.asoc.2012.10.019>
- Gauthier TD (2001) Detecting trends using Spearman's rank correlation coefficient. *Environ Forensics* 2(4):359–362. <https://doi.org/10.1006/enfo.2001.0061>
- Ghorbani H, Moghadasi J (2014a) Development of a new comprehensive model for choke performance correlation in Iranian oil wells. *Adv Environ Biol*, 8(17), 877–882. <http://www.aensiweb.net/AENSIWEB/aeb/aeb/September%202014/877.882.pdf>
- Ghorbani H et al. (2014b). Development of a New Comprehensive Model for Choke Performance Correlation in Iranian Gas Condensate Wells. 8(17), 308–313
- Ghorbani, H., et al. (2017a). The Exposure of New Estimating Models for Bubble Point Pressure in Crude Oil of One of The Oil fields in Iran. *Am J Oil Cheml Technol*, 178–193.
- Ghorbani H et al. (2017b). Developing a New Multiphase Model for Choke Function Relation for Iran's Gas Wells. *Am J Oil Chem Technol* 194–202.
- Ghorbani H et al (2017c) Prediction of gas flow rates from gas condensate reservoirs through wellhead chokes using a firefly optimization algorithm. *J Nat Gas Sci Eng* 45:256–271. <https://doi.org/10.1016/j.jngse.2017.04.034>
- Ghorbani H et al (2018) Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared. *Petroleum*. <https://doi.org/10.1016/j.petlm.2018.09.003>
- Ghorbani H et al (2019) Predicting liquid flow-rate performance through wellhead chokes with genetic and solver optimizers: an oil field case study. *J Petrol Explorat Product Technol* 9(2):1355–1373. <https://doi.org/10.1007/s13202-018-0532-6>
- Ghorbani H et al (2020) Performance comparison of bubble point pressure from oil PVT data: Several neurocomputing techniques compared. *Experiment Computl Multiphase Flow* 2(4):225–246. <https://doi.org/10.1007/s42757-019-0047-5>
- Gilbert, W. (1954) Flowing and gas-lift well performance. In *Drilling and production practice*. American Petroleum Institute
- Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. MIT Press. <https://www.deeplearningbook.org/> [Accessed 27 Aug 2020]
- Gorjaei RG et al (2015) A novel PSO-LSSVM model for predicting liquid rate of two-phase flow through wellhead chokes. *J Nat Gas Sci Eng* 24:228–237. <https://doi.org/10.1016/j.jngse.2015.03.013>
- Gould TL (1976) Discussion of paper: An evaluation of critical multiphase flow performance through wellhead chokes, by Ashford P.E. *J Petrol Technol* 26:843–850
- Grape, S., et al. (2020). Determination of spent nuclear fuel parameters using modelled signatures from non-destructive assay and Random Forest regression. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, pp. 163979. DOI: <https://doi.org/https://doi.org/10.1016/j.nima.2020.163979>

- Guo B. (2007). Petroleum production engineering, a computer-assisted approach Gulf Professional Publishing, 312 pages. eBook ISBN: 9780080479958
- Guo B, Ghalambor A (2012) Natural gas engineering handbook (second edition). Elsevier 472 pages. DOI: <https://doi.org/https://doi.org/10.1016/C2013-0-15534-1>
- Hashemitaheeri M, Mekarthy SMR, Cherukuri H (2020) Prediction of specific cutting forces and maximum tool temperatures in orthogonal machining by support vector and Gaussian process regression methods. *Procedia Manufacturing* 48:1000–1008. <https://doi.org/10.1016/j.promfg.2020.05.139>
- Kartal F, Özveren U (2020) A deep learning approach for prediction of syngas lower heating value from CFB gasifier in Aspen plus. *Energy*. <https://doi.org/10.1016/j.energy.2020.118457>
- Kaydani H et al (2014) Wellhead choke performance in oil well pipeline systems based on genetic programming. *J. Pipeline Syst. Eng. Pract.* 5(3):06014001. [https://doi.org/10.1061/\(ASCE\)PS.1949-1204.0000165](https://doi.org/10.1061/(ASCE)PS.1949-1204.0000165)
- Keras (2020). Deep learning package coded in Python and running on TensorFlow. <https://keras.io/about/> [Accessed 26 Aug 2020]
- Khamis M et al (2020) Optimization of choke size for two-phase flow using artificial intelligence. *J Petrol Explorat Product Technol* 10(2):487–500. <https://doi.org/10.1007/s13202-019-0734-6>
- Kuo B-C et al (2013) A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification. *IEEE J Select Topics Appl Earth Observat Remote Sens* 7(1):317–326. <https://doi.org/10.1109/JSTARS.2013.2262926>
- Lak A et al (2014) Choke modeling and flow splitting in a gas-condensate offshore platform 21:1163–1170. <https://doi.org/10.1016/j.jngse.2014.07.020>
- Latif F, Griston-Castrup S, Al Kalbani A (2012). Field evaluation of MOV adjustable steam chokes. In SPE Western Regional Meeting, 2012. Society of Petroleum Engineers, 10 pages. <https://doi.org/10.2118/153726-MS>
- Lee S et al (2018) Background information of deep learning for structural engineering. *Arch Comput Methods Eng* 25(1):121–129. <https://doi.org/10.1007/s11831-017-9237-0>
- Ling K (2012) Modifications to equations of Gas flow through choke. in SPE Latin America and Caribbean Petroleum Engineering Conference, 2012. SPE-151547-MS. Society of Petroleum Engineers. DOI: <https://doi.org/https://doi.org/10.2118/151547-MS>
- Liu J et al (2016) Representing conditional preference by boosted regression trees for recommendation. *Inf Sci* 327:1–20. <https://doi.org/10.1016/j.ins.2015.08.001>
- Loh K, Omrani PS, van der Linden RJ (2018). Deep learning and data assimilation for real-time production prediction in natural gas wells. <https://arxiv.org/abs/1802.05141> [Accessed 27 Aug 2020]
- Lorena AC, de Carvalho AC (2007) Protein cellular localization prediction with support vector machines and decision trees. *Comput Biol Med* 37(2):115–125. <https://doi.org/10.1016/j.combiomed.2006.01.003>
- Mirzaei-Paiaman A, Salavati S (2012) The application of artificial neural networks for the prediction of oil production flow rate. *Energy Sources Part A Recov Utilizat Environ Effects* 34(19):1834–1843. <https://doi.org/10.1080/15567036.2010.492386>
- Mirzaei-Paiaman A, Salavati SJES, Part A: Recovery, Utilization and E. Effects. (2013). A new empirical correlation for sonic simultaneous flow of oil and gas through wellhead chokes for Persian oil fields. 35(9), pp. 817–825. <https://doi.org/10.1080/15567031003773304>
- Mohamadian N et al (2018) Rheological and filtration characteristics of drilling fluids enhanced by nanoparticles with selected additives: an experimental study. *Adv Geo-Energy Res* 2(3):228–236
- Mohamadian N et al (2019) A hybrid nanocomposite of poly (styrene-methyl methacrylate-acrylic acid)/clay as a novel rheology-improvement additive for drilling fluids. *J Polym Res* 26(2):33. <https://doi.org/10.1007/s10965-019-1696-6>
- Mohamadian N et al (2021) A geomechanical approach to casing collapse prediction in oil and gas wells aided by machine learning. *J Petrol Sci Eng* 2021(196):107811. <https://doi.org/10.1016/j.petrol.2020.107811>
- Mohammadian N, Ghorbani H (2015) An investigation on chemical formation damage in Iranian reservoir by focus on mineralogy role in shale swelling potential in Pabdeh and Gurpi formations. *Adv Environ Biol* 9(4):161–166
- Nasriani HR, Kalantari ASL (2011). Two-phase flow choke performance in high rate gas condensate wells. in SPE Asia Pacific Oil and Gas Conference and Exhibition, 2011. Society of Petroleum Engineers, 9 pages. DOI: <https://doi.org/https://doi.org/10.2118/145576-MS>
- Nasriani HR, Moradi M, Abad D, Kalantariasl A (2016). A new correlation for prediction of critical two-phase flow through wellhead chokes. in 78th EAGE Conference and Exhibition 2016. European Association of Geoscientists & Engineers
- Ng A. (2016). Machine learning yearning: Technical strategy for AI engineers in the era of deep learning (draft version) 118 pages. <https://d2wvfoqc9gyqzf.cloudfront.net/content/uploads/2018/09/Ng-MLY01-13.pdf> [Accessed 27 Aug 2020]
- Nguyen G et al (2019) Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artif Intell Rev* 52(1):77–124. <https://doi.org/10.1007/s10462-018-09679-z>
- Nie G et al (2011) Credit card churn forecasting by logistic regression and decision tree. *Expert Syst Appl* 38(12):15273–15285. <https://doi.org/10.1016/j.eswa.2011.06.028>
- Nwankpa, C., et al. (2018). Activation functions: Comparison of trends in practice and research for deep learning. <http://arxiv.org/abs/1811.03378> [Accessed 27 Aug 2020]
- Omana, R., et al. (1969) Multiphase flow through chokes. In Fall Meeting of the Society of Petroleum Engineers of AIME, 1969. Soc Petrol Eng. 16 pages. DOI: <https://doi.org/https://doi.org/10.2118/2682-MS>
- Ortuño FM et al (2015) Comparing different machine learning and mathematical regression models to evaluate multiple sequence alignments. *Neurocomputing* 164:123–136. <https://doi.org/10.1016/j.neucom.2015.01.080>
- Osei-Bryson K-M (2004) Evaluation of decision trees: a multi-criteria approach. *Comput Oper Res* 31(11):1933–1945. [https://doi.org/10.1016/S0305-0548\(03\)00156-4](https://doi.org/10.1016/S0305-0548(03)00156-4)
- Pan Y et al (2009) A novel QSPR model for prediction of lower flammability limits of organic compounds based on support vector machine. *J Hazard Mater* 168(2–3):962–969. <https://doi.org/10.1016/j.jhazmat.2009.02.122>
- Perkins TK (1993). Critical and subcritical flow of multiphase mixtures through chokes. *SPE Drilling & Completion* 8(04), 6 pages. DOI: <https://doi.org/https://doi.org/10.2118/20633-PA>
- Pilehvari AA (1981). Experimental study of critical two-phase flow through wellhead chokes. University of Tulsa
- Poettmann F, Beck R (1963) New charts developed to predict gas-liquid flow through chokes. *World Oil* 184(3):95–100
- Pouyanfar S et al (2018) A survey on deep learning: algorithms, techniques, and applications. *ACM Comput Surv* 51(5):1–36. <https://doi.org/10.1145/3234150>
- Rashid S et al (2019) Prediction of critical multiphase flow through chokes by using a rigorous artificial neural network method. *Flow Measurment Instrum* 69:101579. <https://doi.org/10.1016/j.flowmeasinst.2019.101579>
- Rashidi S et al (2020) Shear modulus prediction of embedded pressurize salt layers and pinpointing zones at risk of casing collapse in oil and gas wells. *J Appl Geophys*. <https://doi.org/10.1016/j.jappgeo.2020.104205>

- Rolnick D, et al. (2017). Deep learning is robust to massive label noise (version 3). <https://arxiv.org/abs/1705.10694> [accessed 27 August 2020]
- Ros N (1960) An analysis of critical simultaneous gas/liquid flow through a restriction and its application to flow metering. *Appl Scientif Res* 9(1):374, 0003–6994. <https://doi.org/10.1007/BF00382215>
- Rui J, et al. (2019). Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. *J Petrol Sci Eng*, 180, pp. 699–706, 0920–4105. Doi: <https://doi.org/10.1016/j.petrol.2019.06.014>
- Safar Beiranvand M et al (2012) New multiphase choke correlations for a high flow rate Iranian oil field. *Mech Sci* 3(1):43–47. <https://doi.org/10.5194/ms-3-43-2012>
- Saikia P et al (2020) Artificial neural networks in the domain of reservoir characterization: a review from shallow to deep models. *Comput Geosci* 135:104357. <https://doi.org/10.1016/j.cageo.2019.104357>
- Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks* 61, pp. 85–117, 0893–6080. DOI: <https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003>
- Scikit-Learn (2020). Decision tree module for Python. <https://scikit-learn.org/stable/modules/tree.html> [Accessed 25 Aug 2020]
- Scikit-Learn (2020). Random forest regressor module for Python. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html> [accessed 25 August 2020]
- Selvam RK (2018). Adam and RMSProp Optimizer - Implementation and Testing. <https://www.sravikiran.com/GSOC18/2018/08/02/adamandrmsprop/> [Accessed 26 August 2020]
- Shahbaz M et al (2019) Artificial neural network approach for the steam gasification of palm oil waste using bottom ash and CaO. *Renewable Energy* 132:243–254. <https://doi.org/10.1016/j.renene.2018.07.142>
- Shao M, et al. (2020). Prediction of energy consumption in hotel buildings via support vector machines. *Sustainable Cities and Society*, pp. 102128. DOI: <https://doi.org/https://doi.org/10.1016/j.scs.2020.102128>
- Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. *Statist comput* 14(3):199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Tangren R, Dodge C, Seifert H (1949) Compressibility effects in two-phase flow. *J Appl Phy* 20(7):637–645 0021–8979. <https://doi.org/10.1063/1.1698449>
- Taqi AM et al. (2018) The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance. in 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). 2018.
- TensorFlow (2020). An open-source machine learning platform. <https://github.com/tensorflow/tensorflow> [Accessed 26 Aug 2020]
- Tsai C-F, Chiou Y-J (2009) Earnings management prediction: a pilot study of combining neural networks and decision trees. *Expert Syst Appl* 36(3):7183–7191. <https://doi.org/10.1016/j.eswa.2008.09.025>
- Vapnik V (2013) The nature of statistical learning theory. Springer, New York, p 314
- Vapnik V, Golowich SE, Smola AJ (1996) Support vector method for function approximation, regression estimation and signal processing. In: *Adv Neur Inf Process Syst* 9:281–288
- Wen G, Tang M, Benson SM (2019). Multiphase flow prediction with deep neural networks. <https://arxiv.org/abs/1910.09657> [Accessed 27 Aug 2020]
- Wood DA (2018) Transparent OpenBox learning network provides insight to complex systems and a performance benchmark for more opaque machine learning algorithms. *Adv Geo-Energy Res* 2(2): 148–162. <https://doi.org/10.26804/ager.2018.02.04>.
- Wu T et al (2018) Automatic lateralization of temporal lobe epilepsy based on MEG network features using support vector machines. *Complexity*. <https://doi.org/10.1155/2018/4325096>
- Zarenezhad B, Aminian AJPS, Technology (2011). An artificial neural network model for design of wellhead chokes in gas condensate production fields. 29(6), pp. 579–587, 1091–6466. DOI: <https://doi.org/https://doi.org/10.1080/10916460903551065>
- Zhang Q et al (2018) A survey on deep learning for big data. *Inform Fus* 42:146–157. <https://doi.org/10.1016/j.inffus.2017.10.006>
- Zheng A, Casari A (2018) Feature engineering for machine learning: principles and techniques for data scientists O'Reilly Media, Inc. 217 pages. https://www.repath.in/gallery/feature_engineering_for_machine_learning.pdf [accessed 27 Aug 2020]
- Zhou X et al (2020) Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. *Reliabil Eng Syst Safet*. <https://doi.org/10.1016/j.res.2020.106931>

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