Energy Reports 8 (2022) 2233–2247

Contents lists available at ScienceDirect

Energy Reports

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

Research paper

A robust approach to pore pressure prediction applying petrophysical log data aided by machine learning techniques

Guodao Zhang [a], Shadfar Davoodi [b], Shahab S. Band [c,*], Hamzeh Ghorbani [d,e,**], Amir Mosavi [f,g,h,***], Massoud Moslehpour [i,j,****]

[a] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[b] School of Earth Sciences & Engineering, Tomsk Polytechnic University, Lenin Avenue, Tomsk, Russia
[c] Future Technology Research Center, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan
[d] Young Researchers and Elite Club, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran
[e] Faculty of General Medicine, University of Traditional Medicine of Armenia (UTMA), Yerevan, Armenia
[f] Institute of Information Society, University of Public Service, 1083 Budapest, Hungary
[g] John von Neumann Faculty of Informatics, Obuda University, Budapest, Hungary
[h] Institute of Information Engineering, Automation and Mathematics, Slovak University of Technology in Bratislava, Slovakia
[i] Department of Business Administration, Asia University, 500, Lioufeng Rd., Wufeng, Taichung 41354, Taiwan
[j] Department of Management, California State University, San Bernardino, 5500 University Parkway, San Bernardino, CA 92407, USA

ARTICLE INFO

Article history:
Received 27 October 2021
Received in revised form 28 December 2021
Accepted 4 January 2022
Available online 31 January 2022

Keywords:
Pore pressure
Machine learning algorithms
Petrophysical data
Decision tree algorithm

ABSTRACT

Determination of pore pressure (PP), a key reservoir parameter that is beneficial for evaluating geomechanical parameters of the reservoir, is so important in oil and gas fields development. Accurate estimation of PP is also essential for safe drilling of oil and gas wells since PP data are used as the input for safe mud window determination. In the present study, empirical equations along with machine learning methods, namely random forest algorithm, support vector regression (SVR) algorithm, artificial neural network (ANN) algorithm, and decision tree (DT) algorithm, are employed for PP prediction applying well log data. To this end, 2827 data records collected from three wells (Well A, Well B, and Well C) drilled in one of the Middle East oil fields are used. The dataset of Wells A and B is used for models' training, validating, and testing, while Well C dataset is applied for evaluating the models' generalizability in PP prediction in the field under study. To construct the predictive algorithms, 12 input variables are initially considered in the study. A feature selection analysis is conducted to find the most influential input variables set for developing PP predictive models. The results obtained suggest that the 9-input-variable set is the most efficient combination of inputs used in the ML models construction. Among all the four ML algorithms proposed, the DT algorithm presents the most accurate predictions for PP, delivering R^2 and RMSE values of 0.9985 and 14.460 psi, respectively. Furthermore, the model generalization analysis results reveal that the 9-input-variable DT model developed can be used for PP prediction throughout the field of study since it presented an excellent accuracy performance in predicting PP when applied to Well C dataset.

1. Introduction

Pore pressure (PP) is one of the essential parameters in different processes of drilling and exploration, including well design, well stability analysis, and mud program design (Hu et al., 2013; Yu et al., 2020; Bahmaei and Hosseini, 2020; Zhang et al., 2020). Accurate determination of PP helps in selective production and injection, hydrocarbon migration path mapping, and prevention of drilling mud loss during drilling (Mousavipour et al., 2020; Ahedor et al., 2020). Inaccurate estimation of PP can cause severe problems during drilling operation such as loss of drilling mud into the formation being drilling, which may result in the kick of

* Corresponding author at: Future Technology Research Center, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan.
** Corresponding author at: Young Researchers and Elite Club, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran.
*** Corresponding author at: Institute of Information Society, University of Public Service, 1083 Budapest, Hungary.
**** Corresponding author at: Department of Business Administration, Asia University, 500, Lioufeng Rd., Wufeng, Taichung 41354, Taiwan.
E-mail addresses: guodaozhang@zjut.edu.cn (G. Zhang), davoodis@hw.tpu.ru (S. Davoodi), shamshirbands@yuntech.edu.tw (S.S. Band), hamzehghorbani68@yahoo.com (H. Ghorbani), amir.mosavi@kvk.uni-obuda.hu (A. Mosavi), writetodrm@gmail.com (M. Moslehpour).

https://doi.org/10.1016/j.egyr.2022.01.012

the formation fluid in the well and finally well blowout that will, in turn, cause Irreparable damages to the drilling rig and well crew (Zhang et al., 2020; Mahetaji et al., 2020; Maddahi et al., 2020).

The pore pressure, also called the formation pressure, is the pressure of the fluids inside the formation pore, resulting from the hydraulic potential (Oloruntobi et al., 2020). In a drilling operation, pore pressure is regarded as a safe pressure only if the hydrostatic pressure of the drilling fluid in the wellbore falls between the formation pressure and formation fracture pressure (Darvishpour et al., 2019; Richards et al., 2020). Formation pressure can be classified in two forms: (i) natural pressure (i.e., drilling mud hydrostatic pressure exceeds the formation pressure and less than the fracture pressure of the formation) (ii) abnormal pressure (i.e., drilling mud pressure is less than the formation pressure) (Bahmaei and Hosseini, 2020; Li et al., 2012).

1.1. Literature review

In recent years, various methods have been employed by many researchers to predict and estimate the PP accurately, discussed as follows:

The first study made on the prediction of PP was presented by Terzaghi in 1943, in which an empirical equation was developed to estimate the PP. He designed an experiment to examine the effect of rock compaction on overburden pressure. According to the theory proposed, the overburden pressure was neutralized by the PP exerted by the vertical stresses, and all the effects of stress, including changes in the elastic wave velocity, were considered as effective stresses.

Hottmann and Johnson, in 1965, conducted a study to predict PP, considering the properties of shales and the deviation in sound velocity recorded by sonic logs (Hottmann and Johnson, 1965).

In 1943, Terzaghi et al. developed a relationship displaying that overburden pressure is a function of the PP and effective stress in the rock matrix (see Eq. (1)) (Terzaghi, 1943). Having effective stress available and estimating overburden pressure through Eq. (2), PP can be estimated using Eq. (1).

$$S_{over} = S_{eff} + \mathrm{PP} \tag{1}$$

$$S_{over} = \int_0^h \rho g h \, dh \tag{2}$$

Where S_{over} stands for overburden stress, S_{eff} represents effective stress, PP denotes PP and ρ is bulk density, g denotes gravitational acceleration, and h signifies depth.

Biot Willis also proposed an empirical relationship (see Eq. (3)) between overburden pressure, effective stress, and PP, considering a coefficient called Biot, which represents changes in the volume of the pore fluid to those of the total rock volume. Biot coefficient equals 1, where the fluid can readily flow through the pores in the rock. It should be noted that Boit's equation is valid only for homogeneous rocks, and applying the Biot coefficient to heterogeneous rocks leads to a great deal of error (Biot and Willis, 1957).

$$\mathrm{PP} = \frac{S_{over} - S_{eff}}{\beta} \tag{3}$$

Where PP represents the PP, β signifies the Biot coefficient, and S_{over} and S_{eff} represent the overburden and effective stresses, respectively.

In addition to these studies, other empirical models to predict PP have been developed using shear wave velocity, resistance, and compression log data (Eaton, 1975; Bowers, 1995; Yoshida et al., 1996). In 1975, Eaton proposed two empirical models for predicting PP using compressional pressure wave logs, shear wave, and resistivity logs (Eqs. (4) and (5)) (Yoshida et al., 1996; Shen et al., 2017; Farsi et al., 2021a).

$$\mathrm{PP} = S_{over} - (S_{over} - S_{hyd}) \left(\frac{\varDelta_n}{\varDelta_t}\right)^q \tag{4}$$

$$\mathrm{PP} = S_{over} - (S_{over} - S_{hyd}) \left(\frac{R_t}{R_n}\right)^q \tag{5}$$

Where S_{over} represents overburden stress, S_{hyd} denotes hydrostatic pressure gradient stress, $\varDelta_n$ signifies sonic log measured in shale based on compressional log, $\varDelta_t$ stands for sonic log measured in shale based on compressional log, and R_n and R_t denote resistivity log in normal pore pressure profile and resistivity log, respectively.

Reviewing the results presented in most of the articles in which empirical model proposed for prediction of the parameters involved in the oil and gas industry, it was found that these equations provide accurate predictions for the parameter of interest only for the field, the data of which are used in the development of the empirical equations (i.e., empirical models are a field-specific) (Abad et al., 2021a; Naveshki et al., 2021). To overcome this issue, recently, many studies have been performed to develop predictive models for forecasting diverse parameters within the oil and gas industry applying artificial intelligence techniques (Hazbeh et al., 2021a). In the following, some of the intelligent predictive models having been developed in the recent decade are reviewed.

In 2010, Wang et al. conducted a study on the prediction of the PP, where they used three methods: trend line method (TLM), the original Fillippone formula method (OFFM), and hybrid genetic algorithm without a mutation rate (HGANM) (Wang et al., 2010). Analyzing the prediction results for the three methods used, it was found that the proposed HGANM method provided the best prediction performance accuracy.

Later in 2013, the Propagation Artificial Neural Network (BPANN) method was used by Hu et al. to predict the PP based on the data gathered from 5 wells within two different fields. The study's outcome showed a considerable degree of error involved in the predictions made by the model, where the average error reported for the model was 7.15% (Hu et al., 2013).

Abidin, in 2014, used ANN algorithm technique for the prediction of PP. The prediction performance accuracy of their intelligent model was adequately high, where the reported error for PP predictions was equal to 5.0048% (Abidin, 2014).

In 2017, Haris et al. employed the probabilistic neural network (PNN) method to predict PP applying bulk density, Vp/ Vs ratio, P-impedance (Zp), and S-impedance (Zs) as input parameters to the predictive model. Their results displayed the PNN model proposed delivered a high degree of prediction performance accuracy, where the PNN model's precision was 98% higher than relationships developed based on seismic data (Haris et al., 2017).

In 2018, Rashidi and Asadi utilized a set of drilling data, mechanical specific energy (MSE), and drilling efficiency (DE), collected from three wells drilled a sandstone reservoir in Iran to predict the PP of the formation applying an ANN. Analyzing the results of the proposed ANN proved that the model proposed can make accurate predictions. They also stated that the predictive ANN might be applied for real-time PP prediction while a well is being drilled (Rashidi and Asadi, 2018).

Karmakar and Maiti, in 2019, based on 357 well log data records, developed predictive models applying Bayesian neural network (BNN) optimized by the Scaled Conjugate Gradient (SCG) and Hybrid Monte Carlo (HMC) to predict the PP in well U1343E located at Bering Sea slope region of the IODP. The outcome of their study showed that the BNN model presents a high degree of accuracy in PP prediction by delivering reduction error (RE) around 0.98 (Karmakar and Maiti, 2019).

In 2020, four inelegant predictive models, namely gradient boosting machines (GBM), support vector machine (SVM), multilayer perceptron (MLP), and random forest algorithm, were employed by Yu et al. to predict the PP. Comparison of the proposed models' prediction accuracy results displayed that the RF algorithm outperformed the other three models in terms of prediction performance accuracy (Yu et al., 2020).

Recently, in 2021, Abdelaal et al. based on 3100 drilling data recodes, developed three models such as support vector machines, functional networks, and random forest to predict PP during the drilling operation. The proposed model used four input variables, including the rate of penetration (ROP), mud flow rate (Q), standpipe pressure (SPP), and rotary speed (RS). Comparing the prediction results of the models proposed showed that the RF algorithm was the best model among all four models in terms of prediction accuracy (R = 0.98 and AAPE = 2%) (Abdelaal et al., 2021).

The present paper aims to develop four intelligent predictive models for predicting PP by petrophysical data. The four newly configured algorithms developed include Random Forest (RF) algorithm, support vector regression (SVR) algorithm, artificial neural network (ANN) algorithm, and decision tree (DT) algorithm. The models are developed based on a combination of 9 input variables with the highest degree of influence on the PP, conducting a sensitivity analysis. The input variables involved in the development of the predictive models include (LLS), computed gamma ray (CGR), shear-wave velocity (Vs), neutron porosity (NPHI), sonic compression transit time (DT), spectral gamma-ray (SGR), photoelectric absorption factor (PEF), deep resistivity (ILD), and bulk density (RHOB). To the best knowledge of the authors, the DT algorithm has never been applied for PP prediction so far. So, this is the first-ever made DT model employed for predicting PP. After training and testing, a comprehensive comparison is performed on the statistical indicators used for the models' prediction accuracy performance to find the best predictive model in terms of precision of predictions.

2. Methodology

In computer language, the term Artificial Intelligence (AI) is referred to as the intelligence of machine; where in most research work, this tool is called as knowledge and design of intelligent factors that resemble the natural intelligence of human minds (Legg and Hutter, 2007; Russell and Norvig, 2002). In other words, AI is a tool that can present similar behaviors as those of intelligent human behaviors, including understanding complex situations, simulating human thought processes and reasoning methods, and acquiring knowledge and reasoning to solve problems (Poole et al., 1998; Shamshirb et al., 2020; Hassanpouryouzb et al., 2021).

Nowadays, AI algorithms are widely utilized for solving various challenges of engineers in different sectors of science and technology (Choubineh et al., 2017; Ghorbani et al., 2017, 2019, 2020; Ranaee et al., 2021; Farsi et al., 2021b; Hazbeh et al., 2021b; Shamshirb et al., 2019). The origin and main ideas of AI algorithms could be sought in philosophy, linguistics, mathematics, psychology, neuroscience, physiology, control theory, probability, and optimization (Lieder and Griffiths, 2020). They have found numerous applications in computer science, engineering, biology, medicine, social sciences, etc. (Abad et al., 2021b; Rajabi et al., 2021; Shamshirb et al., 2021).

In the present study, four intelligent algorithms are employed to predict the pore pressure, which are RF, SVR, ANN, and DT. The theoretical descriptions of these algorithms are provided in the following.

Fig. 1 displays the workflow schematic used in this study, in which the steps taken for construction, evaluation, and comparison of the prediction accuracy achieved by the intelligent algorithms used to predict pore pressure are presented. As shown in Fig. 1, the first five steps are allocated to dataset preparation, where the collected data are first sorted and filtered to remove the outlier data. Next, the minimum and maximum values for each variable are specified, and then the data are normalized within a numerical range between +1 and −1, applying Eq. (6).

$$d_i^l = \left(\frac{d_i^l - dmin^l}{dmax^l - dmin^l}\right) * 2 - 1 \tag{6}$$

Where d_i^l stand for the value of the attribute for *I*th data, $dmin^l$ and T$dmax^l$ represent the minimum and maximum values of attribute *l* among the entire data points within the dataset, respectively.

Subsequently, the dataset with normalized data is divided into three smaller sets of data, including training (70% of the entire data points in the dataset), testing (15% of the entire data points in the dataset), and validation (15% of the entire data points in the dataset).

After that, the statistical accuracy parameters (APR, AAPR, STD, RMSE & R^2) are calculated for measured values of pore pressure (calculated by Eaton's method, which was appeared to be realistic based on verification made based on the limited RFT PP data) and those predicted by the four artificial intelligence algorithms and empirical models considered in this study, and the best model in terms of prediction accuracy is found. Finally, the best algorithm discovered is then employed for the prediction of pore pressure using a new dataset gathered to form a different well to test its generalizability.

2.1. Decision tree algorithm

One method of machine learning (ML) widely used to evaluate datasets is the DT (Larestani et al., 2022; Lorena and de Carvalho, 2007). In this ML method, a set of data is organized in a hierarchical structure consisting of nodes and strings, the data of which are then classified and prepared by a set of rules for a numerical process (i.e., regression) (Lorena and de Carvalho, 2007). One of the issues with the DT algorithm is that it sometimes takes problem when classifying data. To solve a class-type problem, a class label is given to each tree's leaf, which is assigned to specific leaves according to data classification rules (Larestani et al., 2022). To construct a DT learning algorithm, the input variables (or attributes) and target variables are first distinguished. Next, the data are assigned to "child" nodes based on the defined rules. By splitting further, each child node will act like a parent node from which more layers and nodes will be developed (Osei-Bryson, 2004). In the DT method, first, the entire set of data is split into two subsets, which are decision (child) nodes building the decision tree's second layer. Then, the decision nodes are further split into sub-nodes, which build extra layers of decision nodes. This splitting process continues till reaching a final layer containing leaf (terminal) nodes. The data are classified by the decision tree, and the tree continues to develop until all the data recodes are assigned to the correct leaf.

Ease of preparation and interpretation is one of the essential advantages of the DT algorithm. In addition, there is less need for data preparation and processing in the construction of the algorithm since the missing or outlying data are filtered by the algorithm. However, the DT model highly suffers from a lack of generalizability (i.e., when the model trained is utilized to another independent dataset, it is usually not able to classify these data records within the dataset fully). Another disadvantage of this algorithm is that the increased number of layers and nodes may

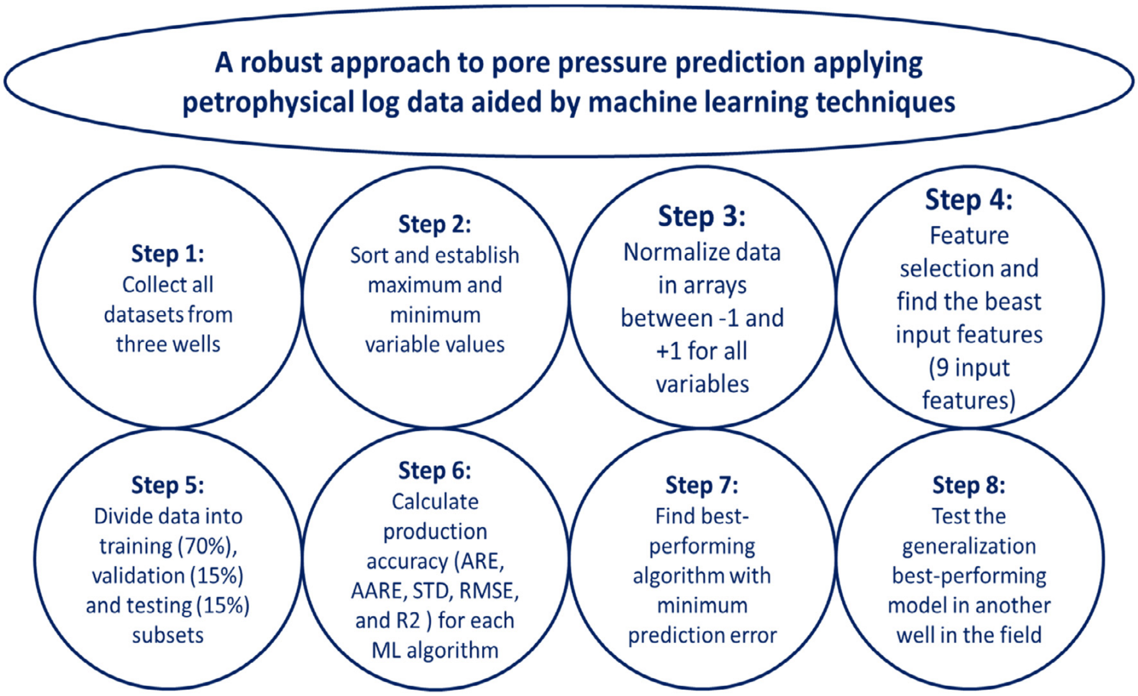
A robust approach to pore pressure prediction applying petrophysical log data aided by machine learning techniques
Step 1: Collect all datasets from three wells
Step 2: Sort and establish maximum and minimum variable values
Step 3: Normalize data in arrays between -1 and +1 for all variables
Step 4: Feature selection and find the beast input features (9 input features)
Step 5: Divide data into training (70%), validation (15%) and testing (15%) subsets
Step 6: Calculate production accuracy (ARE, AARE, STD, RMSE, and R2) for each ML algorithm
Step 7: Find best-performing algorithm with minimum prediction error
Step 8: Test the generalization best-performing model in another well in the field

Fig. 1. Schematic of workflow chart used to predict pore pressure.

Table 1
Control parameters of the DT regression model developed to predict PP.

Control parameters	Value
Maximum depth	105
Criterion	gini
Splitter	best
Objective function	Mean squared error
Example prediction time	0.011478 (s)

Table 2
Control parameters for the SVR to predict PP.

Control parameters	Status
Kernel function	RBF
ε range	0.2
C range	105000
Cross-validation	Not applied
γ range (RBF)	0.045

cause overfitting and decreased level of accuracy (Terzaghi, 1943; Biot and Willis, 1957).

In this study, the DT module of scikit-learn is codded by Python, where the "gini" criterion is employed to determine the feature's importance and "best" splitter is used for making a decision on which features and threshold value to in making each split (see Table 1). The developed DT regression model is applied for predicting pore pressure.

2.2. Support vector regression algorithm

The Support Vector Machine (SVM) algorithm was proposed by Cortes and Vapnik based on statistical learning theory in 1995 (Moosavi et al., 2021). This algorithm is one of the most widely utilized algorithms in various fields to solve classification, regression, and time-series prediction problems (Ahmad et al., 2020). For regression performing, this algorithm first employs core functions for mapping nonlinear vectors to higher dimensions, then builds a hyperplane in the feature space. Finally, it divides the data into two classes to create support vectors that maximize the distance between the hyperplane and classes for the subset used for training (Rui et al., 2019). In the present study, an SVR algorithm is proposed to predict PP.

In the SVR, the parameters $[x_i] \in X = R^h$ and $y_i \in X = R$ are used to specify the input and output variables, respectively, while $i = 1, 2, 3, \ldots, h$. The predictions for this ML model are obtained from a regression function $y = f(x)$, and the output values are approximated by an objective learning function (Eq. (7)).

$$f(x, w) = wZ(x) + b \tag{7}$$

Where; $f(x, w)$ = objective learning function of SVR, $Z(x) =$ Feature mapping to high-dimensional space, $w \in R =$ vector of weight, $b \in R =$ bias (threshold). To keep the presented work concise, the potential readers can refer to previous publications to read more about the SVR model theory (Barjouei et al., 2021; Shao et al., 2020; Smola and Schölkopf, 2004).

Overcoming the complexity of computation involving high-dimensional space, a suitable Kernel function is required to be defined. Any function which is able to satisfy Mercer's condition can be applied as the Kernel function. Generally, polynomial, sigmoid, radial basis function (RBF), and linear are the four most commonly used Kernel functions in SVR (Vapnik, 2013). It should be mentioned that the kernel functions can affect the prediction performance of the SVR. In this paper, the radial basis function is applied for the SVR developed to predict PP. This kernel function is used because of its key feature that helps avoid noises in the training data (Hashemitaheri et al., 2020). Table 2 lists the control parameters for the SVR developed to predict pore pressure.

2.3. Random forest algorithm

The random forest algorithm can be considered an extended version of the DT algorithm since this algorithm builds multiple decision trees to examine (Hidayat and Astsauri, 2021). This algorithm is a supervised learning approach that is appropriate

Table 3
Control parameters values for the regression RF algorithm developed for PP prediction.

Control parameters	Value
Maximum depth	1050
Random state	0
Number of decision trees	1050
Objective function	Mean squared error
Example prediction time	5.42314 (s)

for both applications, regression and classification, based on the two subsets of testing and training data points for which the dependent and independent variables are known (Zhou et al., 2020; Grape et al., 2020). This algorithm constructs multiple decision parallelly trees, each of which uses relatively few layers and nodes. Similar to the DT algorithm, the likelihood of overfitting in the random forest method is slight, as it works on the basis of individual decision trees. Moreover, without compromising the accuracy of the decision, this algorithm reduces prediction results' variance and bias by evaluating predictions made by all the decision trees in a collective manner (Ahmad et al., 2018).

Random forest ML models are trained by bootstrapping the subsets of data points from the entire dataset. The subsets with bootstrapped data points can be applied for developing an unpruned regression or classification decision tree. This model does not use all input variables (M) to construct each tree as splitting candidates; rather, it randomly selects a small, randomly chosen number of the input variables available for each tree and then utilizes them in splitting. Multiple trees are iteratively constructed in this approach until a defined number of trees (M) are reached. For solving the regression problem, output variable predictions are obtained by aggregating the predictions (i.e., bagging) from all the single regression trees constructed. The processing of bagging decreases the individual trees' complexity and likely reduces the overall likelihood of the model to overfit the training subset. Eq. (8) defines the prediction function of the random forest algorithm.

$$\hat{f}_{RF}^{M}(x) = \frac{1}{M}\sum_{k=1}^{M} R_i(x) \tag{8}$$

Where $F(X)$ represents random forest prediction function, M stand for the number of independent regression trees, x represents the vector of input variable, and $R_i(x)$ is prediction made by a single tree for the ith data point.

To determine the error associated with the random forest model, an out-of-bag (OOB) error analysis is conducted progressively as the forest containing individual trees is built. OOB is obtained by unchosen data points (i.e., OOB subset) as a test to the Mth tree once it gets trained through the bagging process. More details on the estimation prediction accuracy of the relative importance dependent variables can be found in previous studies (Abidin, 2014; Ahmad et al., 2018). In this research, the Scikit Learn Random Forest Regressor is applied to establish the regression RF algorithm to predict PP. The control parameters of the RF algorithm developed are listed in Table 3.

2.4. Artificial neural network algorithm

ANN is one of the most widely used intelligent techniques in diverse areas of science and engineering that helps with solving complex classification and regression problems (Shahbaz et al., 2019). This algorithm is highly reputed of the black box system because of its hidden layer of regression-like computation. Multi-hidden-layer perceptron (MLP) and feed-forward neural network that contains only one hidden layer are the two most commonly used types of ANN. In this study, an ANN is constructed for PP prediction, which has a single hidden layer (Belhaj et al., 2021).

In ANN, the information is adjusted by the bias and weight vectors and sent from one layer's neurons to the next layer's neurons. The processing of information is conducted in the hidden layer's neurons, and the signal processed is adjusted using an activation function and sent forward to the out layer. The activator function for the ANN is given in Eq. (9).

$$f(x) = f\left(\sum_{i=1} W_{ij}x_i + b_j\right) \tag{9}$$

Where; $f(x)$ resents activator function, b_j stands for bias in the hidden layer, x_i signifies the ith input variable, w_i represents the weights of the connection between the jth neuron and the ith input.

To enhance the prediction performance of neural network, a backpropagation algorithm is typically used in the training process of the network to adjust the values of weights and bias assigned to the hidden layer. It is provided through the minimization of the mean squared error (MSE) between the predicted and measured values collectively for all the records of data within the subset used for training, as given in Eq. (10).

$$Error_{MSE} = \frac{1}{S}\sum_{i}^{m}(\hat{y}_i - y_i)^2 \tag{10}$$

Where; S represents the total number of available data records in the subset used for training, y_i stands for the predicted value for input variable for the data record i, and $\hat{y}_i$ denotes the measured value for the input variable for the data record i.

In ANN training, applying alternative optimization for backpropagation can improve the network's convergence efficiency and prediction performance. For this purpose, a variety of optimizers such as Momentum, RMSprop, Adagrad, Adam, Nesterov Accelerated Gradient, and Adadelta have been applied so far. In the present study, RMSprop, an enhanced gradient-based algorithm, is employed as an optimizer. In this optimizer, the estimated gradient is steadily divided by the rolling average of gradient values recently obtained. The relationships given in Eqs. (11) and (12) are used to update the RMSprop's initial learning rate for different weights (Kartal and Özveren, 2020).

$$E\left[gr^2\right]_p = 0.9E\left[gr^2\right]_{p-1} + 0.1gr_p^2 \tag{11}$$

$$\tau_{p+1} = \tau_n - \frac{\delta}{\sqrt{\left[gr^2\right]_p + \epsilon}} gr_n \tag{12}$$

Where; $E\left[gr^2\right]_p$ represents mean gradient at iteration p, gr_p stands for objective function's gradient at iteration n, τ_p denotes the objective function at iteration p, δ represents the rate of learning, and ϵ signifies smoothing term.

The structure of the ANN developed consists of an input layer with the same number of neurons as in the hidden layer, a single hidden layer that has 600 neurons, an output layer that contains a neuron to predict the target variable. Table 4 lists the value of control parameters for the developed ANN.

2.5. Feature selection method

The prediction performance and computation time for ML models can be improved by using the most effective input variables in pp modeling (Abad et al., 2022). The involvement of a large number of potential input variables in training ML models dramatically affects the speed and accuracy of the models. As a result, To overcome this issue, a feature selection analysis is recommended to be conducted (Wahab et al., 2015). Filtering,

Table 4
Control parameters of the developed ANN for predicting PP.

Control parameters	Status
Number of hidden layers	1
Number of neurons in the hidden layer	600
Activation function used input to hidden layer	SELU (Scaled Exponential Linear Unit)
Objective function minimized for training subset	MSE
Activation function used hidden to output layer	SELU
Optimization algorithm	RMSprop
Patience (number of iterations)	30
Minimum delta	0.02%
Number of iterations	240
Learning rate	0.015

packing, and embedding are three common methods for feature selection analysis, among which the wrapping method is regarded to be more effective and precise (Jain and Zongker, 1997; Shah et al., 2020; Fu et al., 2020). This method uses a hybrid genetic algorithm (GA) with simple multilayer perceptron (MLP) (MLP-GA) to predict PP, reducing the cost function of mean square error (MSE) or root mean square error (RMSE) (Farsi et al., 2021a; Salehi et al., 2020; Jotheeswaran and Koteeswaran, 2020). For a dataset including N initial input features, the warping method, first, categorizes all the initial input features into sets with one, two, …, n features. Then, the target parameter is predicted by the model used for feature selection for each set of features, and the RMSE value associated with the target parameter prediction is estimated and reported for each feature set. Finally, the set of features that provides the most accurate prediction for the target parameter is selected as the most efficient set of input features to be used in predictive model training (Table 6).

2.6. Error parameters

To evaluate and compare the algorithms' performance in PP prediction, the following statistical error indicators, namely relative error (RE), average relative error (ARE), absolute average relative error (AARE), mean squared error (MSE), coefficient of determination (R^2), root mean squared error (RMSE); the objective function of the ML models), and standard deviation (STD) are applied in this study (see Eqs. (13)–(20))

Relative error (RE):

$$\mathrm{RE} = \frac{\mathrm{PP}_{(Measured)} - \mathrm{PP}_{(Predicted)}}{\mathrm{PP}_{(Measured)}} \times 100 \tag{13}$$

Average relative error (ARE):

$$\mathrm{ARE} = \frac{\sum_{i=1}^{n} \mathrm{RE}_i}{n} \tag{14}$$

Absolute average relative error (AARE):

$$\mathrm{AARE} = \frac{\sum_{i=1}^{n} |\mathrm{RE}_i|}{n} \tag{15}$$

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\mathrm{PP}_{Predicted\,i} - \mathrm{PP}_{Measured\,i})^2}{\sum_{i=1}^{N} (\mathrm{PP}_{Predicted\,i} - \frac{\sum_{i=1}^{n} \mathrm{PP}_{Measured\,i}}{n})^2} \tag{16}$$

Mean Square Error (MSE):

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\mathrm{PP}_{Measured\,i} - \mathrm{PP}_{Predicted\,i})^2 \tag{17}$$

Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{18}$$

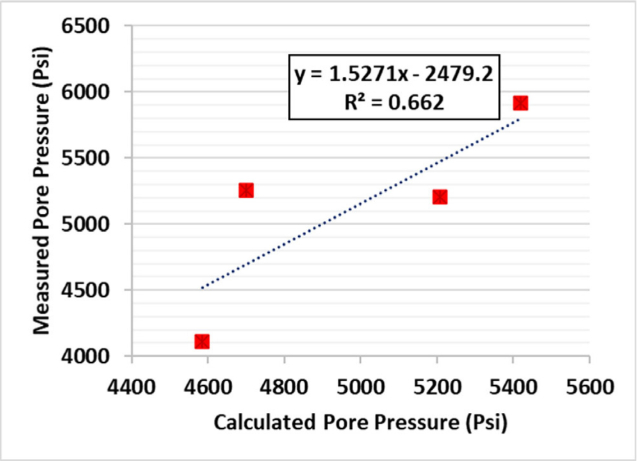

Fig. 2. Calculated PP measured PP Vs. data by RFT tool.

Standard Deviation (STD):

$$\mathrm{STD} = \sqrt{\frac{\sum_{i=1}^{n} (D_i - Dimean)^2}{n-1}} \tag{19}$$

$$Dimean = \frac{1}{n} \sum_{i=1}^{n} (\mathrm{PP}_{Measured\,i} - \mathrm{PP}_{Predicted\,i}) \tag{20}$$

3. Data collection, feature selection, data description

3.1. Data collection

This paper uses data collected from three wells A, B, and C, drilled in one of the oil fields in the Middle East. 988 data records are collected for Well A at depth interval from 3257 to 3454 m, 905 data records are collected from Well B at the depth interval from 3194 to 3375 m, 934 data records are collected from Well C at the depth interval from 3204 and 3390 m. It should be noted that the data recording distance for all three wells A, B, and C is 0.2 m. The initial dataset provided contains data records for one output variable (PP) and 12 input variables. The Initial input variables used for PP modeling are laterolog shallow (LLS), corrected gamma ray (CGR), shear-wave velocity (Vs), uncorrected spectral gamma-ray (SGR), sonic compression transit time (DT), bulk density (RHOB), the photoelectric absorption factor (PEF), neutron porosity (NPHI), deep resistivity (ILD), caliper (CALL), hole size (HS), and compression-wave velocity (Vp). In this article, pore pressure was calculated by Eaton's formula, which appeared to be realistic based on the limited repeat formation tester (RFT) data (see Fig. 2).

3.2. Feature selection

After the MLP-GA algorithm was constructed to be used in feature selection analysis, all the data records of Well A and Well B were unified on a data set (1893 data records). Then, the whole data records were provided to the MLP-GA algorithm. After that, dividing the input variables into different combinations of variables as well as considering evolution criteria (minimum value of RMSE), PP was estimated by MLP-GA algorithm, and the best combination of input variables (in terms of performance accuracy) with most effective parameters was detected and selected to be used in PP modeling.

Table 5 shows the indicators for all the input variables. As displayed in Fig. 3, the set of input variables containing nine

Table 5
Indicators of each feature to predict PP.

Features	Label
NPHI	H1
HS	H2
ILD	H3
RHOB	H4
CGR	H5
LLS	H6
CALL	H7
SGR	H8
PEF	H9
Vs	H10
DT	H11
Vp	H12

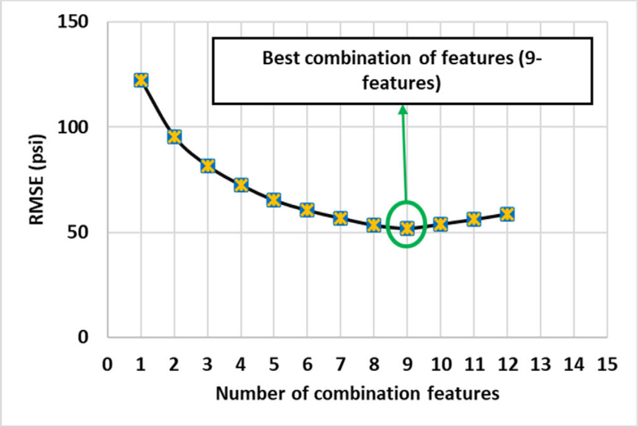

Fig. 3. RMSE values obtained by each set of features (the dataset of two wells A and B is considered).

features (H1, H3, H4, H5, H6, H8, H9, H10, and H12) is the most efficient combination of input variables among all 12 sets of features analyzed in terms of prediction performance accuracy. Consequently, this 9-variable set is selected as input to be further involved in the development of ML models for PP prediction in this study.

3.3. Data description

Performing the feature selection analysis, nine variables, namely corrected gamma ray, neutron porosity, photoelectric index, deep resistivity, shear-wave velocity, Laterolog shallow, spectral gamma-ray, and bulk density, were employed as input for PP modeling. The data collected from two wells, A and B (carbonate reservoir), are applied to construct the four ML models. The statistical parameters corresponding to the input and output variables involved in PP modeling for all the datasets (wells A, B, C, and the total dataset) are provided in Tables 7 and 8, respectively. It should be highlighted that the PP data are verified applying well test data.

4. Result and discussion

For developing the four ML algorithms (RF, SVR, ANN, and DT), data records gathered from well A and well B was unified in a dataset, where 70% of data records was used for algorithms training, 15% of the data points was used for testing the algorithms. The remaining 15% was used for validation. Ensuring that all the training and validation records are involved in the evaluation, a 8-fold cross-validation is performed on the validation and training subsets. Tables 9–12 list the values of the statistical errors delivered in predicting PP by each of the four algorithms developed for the training, validation, testing, total subsets, respectively.

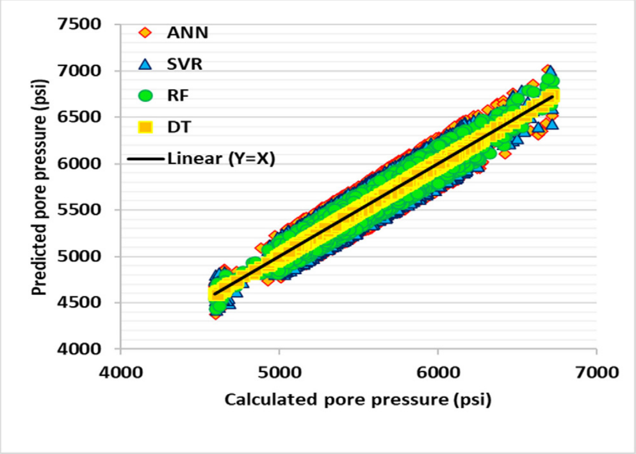

Fig. 4. Predicted values versus calculated PP (using Eaton's formula) for the four ML algorithms developed for the total subset (the entire 1893 data records collected from wells A and B).

Based on the results listed in Tables 9 to 12, it is evident that the prediction accuracy delivered by the DT algorithm is much higher than those of the other three algorithms (SVR, RF, and ANN). The excellent performance of DT in prediction PP could be attributed to its outstanding feature that eliminates the outrange data. From the results presented in Tables 9–12, it can be seen that DT performs high accuracy predictions, where the RMSE value was equal to 14.33, 16.42, 14.53, 16.46 psi for the training, testing, validation, and total subsets, respectively. Outperforming the other three algorithms in terms of the accuracy of prediction, the DT algorithm was recognized as the best for PP prediction in the present study.

Fig. 4 illustrates the values of the PP predicted by the four ML algorithms versus the measured PP values for the total subset (the entire data records collected from wells A and B). It is clear from Fig. 4 that the DT algorithm archives higher PP prediction accuracy in comparison to the SVR, ANN, and RF algorithms.

To visually compare the evaluated algorithms' performance in PP prediction, the cross plot of the PP predicted and measured values for each algorithm is displayed in Fig. 5. Looking closely at Fig. 5, it is evident that the most accurate predictions for PP are achieved by the DT algorithm, and the accuracy of the predictions delivered by the algorithms with respect to R^2 values can be ordered as DT > RF > SVR > ANN.

Fig. 6 shows the relative error of the PP predicted values obtained by the four ML algorithms (ANN, SVR, RF, and DT). The results displayed in Fig. 6 suggest that the relative error achieved by the DT algorithm is much lower than those of the other algorithms evaluated ($0.49 < RE\% < 0.49$). It needs to be highlighted that the improved accuracy obtained by the DT algorithm is because of its ability to consider all the possible outcomes of a decision and conclude each path. This creates a comprehensive analysis of the consequences along each branch and node and identifies decisions that need further analysis.

Fig. 7 displays the RMSE values versus 100 in iterations for the PP predictions obtained by the four ML algorithms (ANN, SVR, RF, and DT). As it can be seen from Fig. 7, the DT algorithm presents a high degree of convergency from the outset, where, at iteration 5, it shows a promising accuracy (RMSE = 14.50 psi) from the outset. However, the RMSE values for SVR, ANN, and RF, in the beginning, are very high (low prediction accuracy), and then they

Table 6
Results of feature selection method performed using data records of two wells A and B (1893 data records).

Number of combination features	Features	RMSE (psi)
1	H10	122.1249
2	H10, H12	95.3524
3	H4, H12, H10	81.5487
4	H12, H3, H10, H4	72.3655
5	H6, H4, H10, H12, H3	65.1785
6	H8, H3, H6, H4, H10, H12	60.3248
7	H5, H10, H8, H12, H4, H3, H6	56.6555
8	H9, H8, H12, H6, H3, H4, H10, H5	53.3476
9	**H1, H6, H3, H8, H4, H10, H12, H5, H9 (Best combination)**	**51.9215**
10	H11, H9, H4, H10, H12, H5, H8, H3, H6, H1	53.7845
11	H2, H4, H1, H5, H10, H8, H9, H6, H11, H12, H3	56.0024
12	H7, H11, H12, H1, H3, H2, H8, H9, H5, H10, H4, H6	58.6315

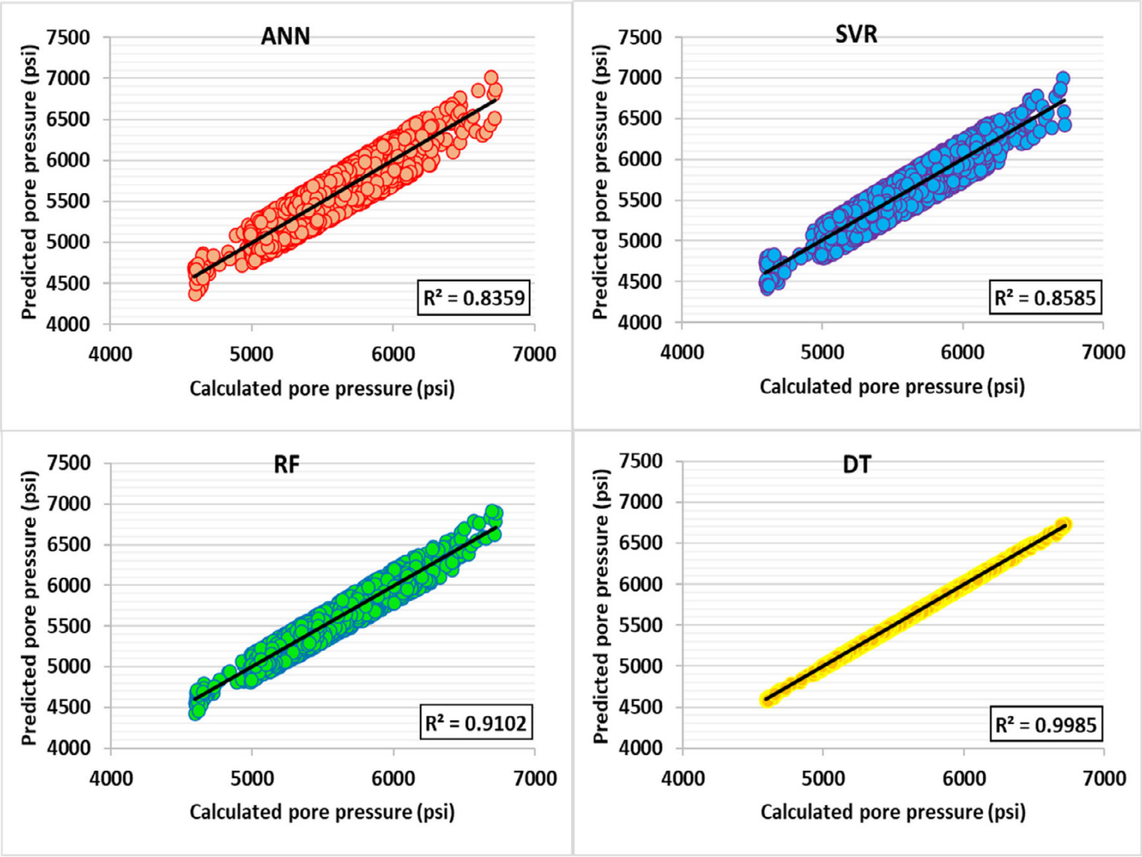

Fig. 5. Cross plot of the calculated PP (using Eaton's formula) versus predicted values of PP for the four artificial intelligence algorithms (ANN, SVR, RF, and DT) for all records of data collected from wells A and B (1893 data).

converge rapidly, and the performance accuracy of the algorithms improves. Comparing the convergence speed of the algorithms evaluated shows that the DT algorithm provides a better high convergence speed in finding a solution than the SVM, RF, and ANN algorithms.

4.1. Generalization of the DT algorithm in PP prediction

The results discussed in the previous section display training, testing, and validation for the four evaluated algorithms in respect of the data points gathered from well A and Well B. In this work, an additional set of data containing 934 data records collected from another well (Well C), located at the same field as wells A and B, is considered to evaluate the capability of the DT algorithm in making precise PP predictions for general application in the field under evaluation. Table 12 lists the statistical measures of accuracy obtained by the best-performing algorithm (DT) applying the Well C dataset. Comparing the results reported in Table 13 with those presented in Tables 9–12 corroborates the substantial ability of the developed DT algorithm to predict PP when used for another well accurately (Well C) in the field under evaluation. Fig. 8 demonstrates the measured versus predicted values of PP obtained by the DT algorithm trained with Wells A and B applied to the dataset collected from Well C. The results presented in Fig. 8 also confirm the DT algorithm's credibility in predicting PP in other wells drilled throughout the field under evaluation. It is also worth noting that the proposed algorithm can be modified and optimized by other researchers to be applied further to predict PP in other fields.

Table 7
Statistical parameters of selected input variables for total dataset (sum of data records in wells A, B, and C; 2827 records of data).

Wells	Variables	Corrected gamma ray	Neutron porosity	Compressional-wave velocity	Photoelectric index	Deep resistivity	Shear-wave velocity	Laterolog shallow	Uncorrected spectral gamma-ray	Bulk density	Pore pressure
	Symbol	CGR	NPHI	vp	PEF	ILD	vs	LLS	SGR	RHOB	PP
	Units	GAPI	PU	km/s	Barn/cm^3	mmho/m	km/s	mmho/m	GAPI	g/cm^3	psi
All wells A, B, and C (2827 data points)	Mean	23.45	13.03	53.64	3.03	1226.05	246.57	209.35	45.00	2.98	5718.30
	Std. Dev.	19.51	5.50	2.96	1.48	4421.68	241.40	1631.94	20.92	0.54	401.51
	Variance	380.69	30.22	8.75	2.20	19544359.31	58251.17	2662282.39	437.54	0.29	161151.80
	Minimum	1.06	−1.55	45.72	−0.45	0.42	57.55	0.48	12.21	1.20	4592.54
	Maximum	124.27	46.67	82.91	6.33	20012.34	738.98	20003.12	146.30	3.93	6690.12
	Skewness	2.05	0.92	2.09	−0.42	3.85	0.92	11.24	1.45	0.21	−0.07
	Kurtosis	4.78	4.84	12.90	−0.54	13.17	−1.09	130.26	2.94	−1.12	−0.53

Table 8
Statistical parameters of the selected input variables for the datasets collected from wells A, B, and C separately.

Wells	Variables	Corrected gamma ray	Neutron porosity	Compressional-wave velocity	The photoelectric index	Deep resistivity	Shear-wave velocity	Laterolog shallow	Uncorrected spectral gamma-ray	Bulk density	Pore pressure
	Symbol	CGR	NPHI	vp	PEF	ILD	vs	LLS	SGR	RHOB	PP
	Units	GAPI	PU	km/s	Barn/cm^3	mmho/m	km/s	mmho/m	GAPI	g/cm^3	psi
Well A	Mean	19.70	12.35	54.13	4.22	967.65	99.14	109.98	44.34	2.61	5318.17
(988 data point)	Std. Dev.	19.37	5.24	2.76	0.72	3779.54	6.29	937.14	18.92	0.11	252.51
	Variance	374.72	27.38	7.59	0.52	14270606.99	39.49	877345.44	357.43	0.01	63699.61
	Minimum	1.06	1.34	47.05	2.12	0.42	79.58	0.48	14.61	2.29	4592.54
	Maximum	107.90	28.24	67.30	5.63	20000.00	187.09	20000.00	143.60	2.87	6690.12
	Skewness	2.51	0.11	0.30	−0.42	4.51	4.00	16.74	1.90	−0.92	0.35
	Kurtosis	6.74	−0.20	0.54	0.38	19.20	46.29	304.39	5.80	1.05	1.74
Well B	Mean	24.59	14.84	52.75	3.33	1563.19	369.43	270.68	38.69	2.71	5789.94
(905 data point)	Std. Dev.	18.48	5.38	3.73	1.03	4997.27	264.19	1918.88	20.19	0.44	226.39
	Variance	341.08	28.91	13.91	1.05	24945075.87	69717.49	3678039.10	407.34	0.20	51197.85
	Minimum	3.31	1.92	45.72	1.26	0.45	82.48	0.53	12.21	1.20	5234.94
	Maximum	110.20	46.67	82.91	6.33	20012.34	738.98	20000.00	119.30	3.49	6325.86
	Skewness	1.57	1.52	3.41	0.41	3.30	−0.02	9.71	1.26	0.01	−0.13
	Kurtosis	2.67	8.32	17.87	−0.45	9.14	−1.92	96.13	1.30	−0.06	−0.76
Well C	Mean	26.34	12.01	53.97	1.48	1175.50	285.05	256.09	51.81	3.62	6076.42
(934 data points)	Std. Dev.	20.03	5.46	1.96	1.05	4447.88	264.73	1890.17	21.62	0.21	260.00
	Variance	400.84	29.80	3.84	1.10	19762416.72	70007.72	3568929.33	467.04	0.05	67527.30
	Minimum	6.18	−1.55	48.06	−0.45	2.55	57.55	3.65	25.41	2.33	5300.99
	Maximum	124.27	43.20	65.54	3.96	20002.09	712.52	20003.12	146.30	3.93	6679.28
	Skewness	2.13	1.30	0.93	−0.16	3.91	0.44	9.87	1.50	−2.78	0.13
	Kurtosis	5.26	6.92	2.57	−1.11	13.50	−1.76	99.36	2.58	12.85	−0.22

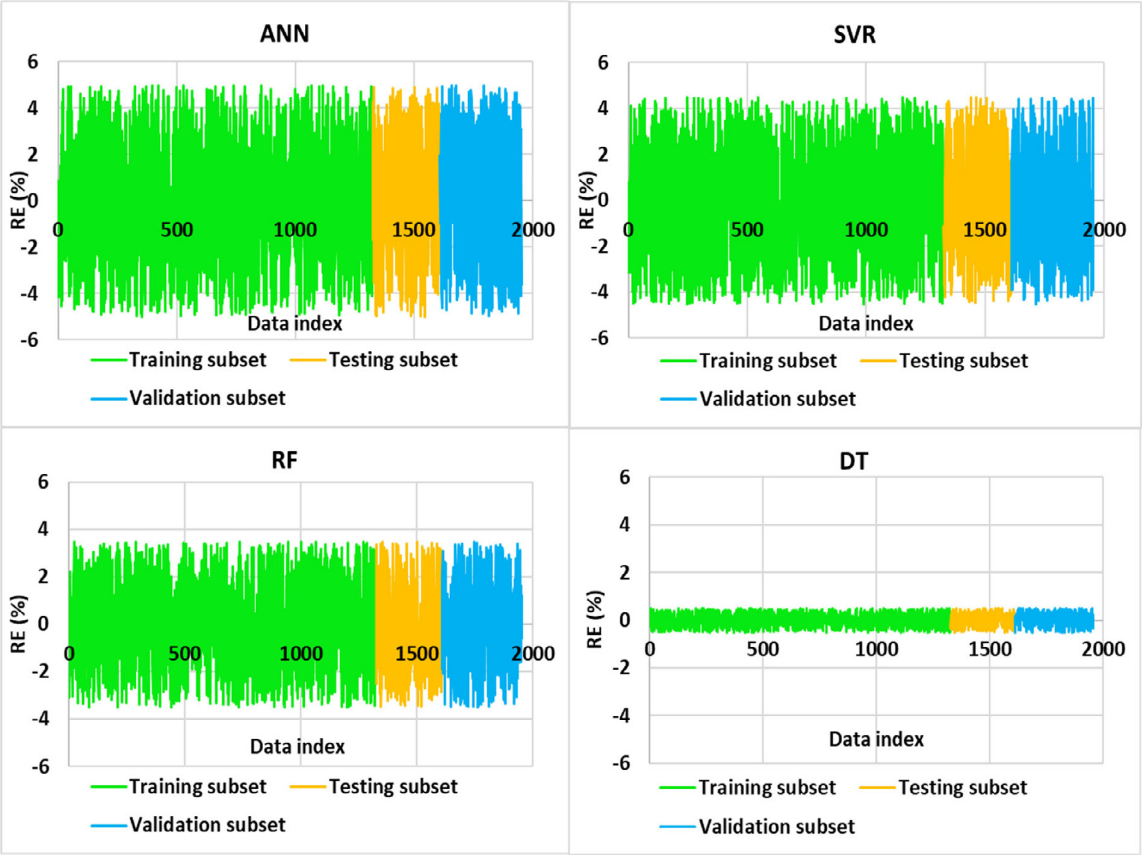

Fig. 6. RE valued involved in PP predictions obtained by the four ML algorithms (all 1893 data records collected for wells A and B are considered).

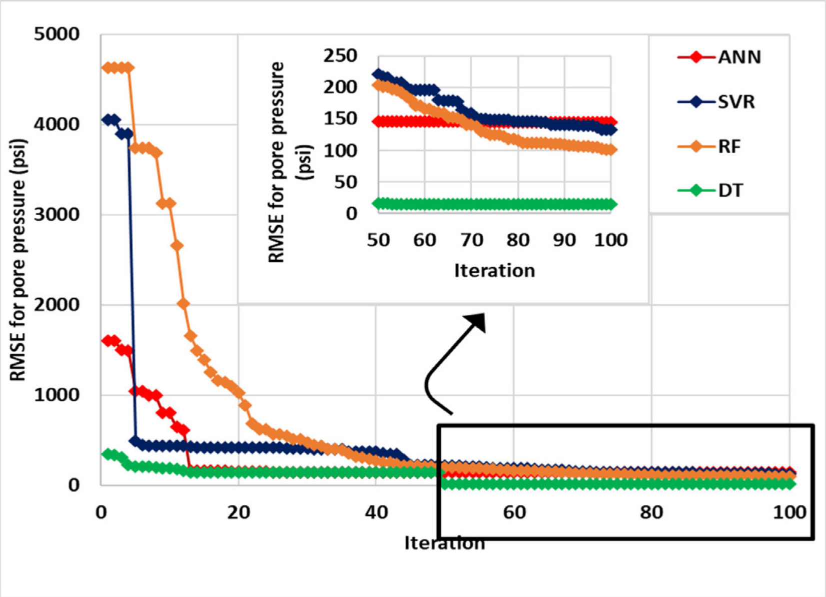

Fig. 7. Comparison of RMSE values reached after each iteration for PP predictions obtained by the four evaluated algorithms (ANN, SVR, RF, and DT).

Table 9
Prediction accuracy for PP predictions by the four algorithms developed based on training subset (1325 data records; 70% of all data collected from wells A and B).

Models	ARE	AARE	STD	MSE	RMSE	R^2
Units	(%)	(%)	(psi)	(psi)	(psi)	–
ANN	−0.057	1.908	143.541	2.058E+04	143.4727	0.8069
SVR	−0.096	1.801	133.042	1.772E+04	133.0981	0.8312
RF	0.045	1.351	101.311	1.027E+04	101.3535	0.8951
DT	−0.002	0.191	14.336	2.055E+02	14.3365	0.9980

Table 10
Prediction accuracy for PP predictions by the four algorithms developed based on a testing subset (284 data records; 15% of all data collected from wells A and B).

Models	ARE	AARE	STD	MSE	RMSE	R^2
Units	(%)	(%)	(psi)	(psi)	(psi)	–
ANN	0.115	2.870	164.232	2.701E+04	164.3352	0.7167
SVR	−0.161	2.634	149.232	2.234E+04	149.4746	0.7747
RF	−0.153	1.974	114.449	1.316E+04	114.7117	0.8482
DT	0.007	0.282	16.428	2.699E+02	16.4298	0.9982

Table 11
Prediction accuracy for PP predictions by the four algorithms developed based on validation subset (284 data records; 15% of all data collected from wells A and B).

Models	ARE	AARE	STD	MSE	RMSE	R^2
Units	(%)	(%)	(psi)	(psi)	(psi)	–
ANN	0.086	3.058	171.530	2.944E+04	171.5848	0.7535
SVR	−0.082	2.647	148.133	2.196E+04	148.1807	0.8016
RF	−0.047	2.037	116.222	1.351E+04	116.2371	0.8921
DT	0.060	0.293	16.271	2.731E+02	16.5256	0.9954

Table 12
Prediction accuracy for PP predictions by the four algorithms developed based on the total subset (1893 data records; 100% of all data collected from wells A and B).

Models	ARE	AARE	STD	MSE	RMSE	R^2
Units	(%)	(%)	(psi)	(psi)	(psi)	–
ANN	−0.041	2.039	145.611	2.114E+04	145.3917	0.8359
SVR	−0.068	1.886	132.973	1.770E+04	133.0242	0.8585
RF	0.032	1.420	101.657	1.033E+04	101.6480	0.9102
DT	−0.001	0.202	14.470	2.093E+02	14.4669	0.9985

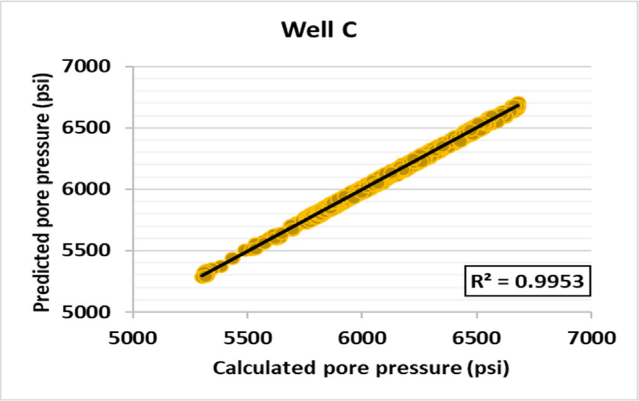

Fig. 8. Cross plot of calculated PP (using Eaton's formula) versus predicted values of PP obtained by the DT algorithm trained using wells A and B dataset applying the whole dataset for well C in the field under study.

Table 13
Statistical accuracy metrics for PP predictions achieved by the DT algorithm trained using wells A and B dataset applied the whole dataset for well C in the field under study.

Models	ARE	AARE	STD	MSE	RMSE	R^2
Units	(%)	(%)	(psi)	(psi)	(psi)	–
DT	−0.001	0.237	17.157	2.944E+02	17.1573	0.9953

5. Conclusion

Determination of pore pressure (PP), a key reservoir parameter that is handy for evaluating geomechanical parameters in reservoir and drilling, is so important in oil and gas fields development. This study presents four robust ML algorithms constructed for predicting PP using petrophysical data. The four ML models developed include RF, SVR, ANN, and DT, which were trained, tested, and validated with the dataset collected from two wells located (Well A and Well B) in an oil field in the Middle East. Conducting feature selection, the best combination of variables to be used as input for predictive is recognized, that contains nine variables, including laterolog shallow (LLS), corrected gamma ray (CGR), shear-wave velocity (Vs), spectral gamma-ray (SGR), bulk density (RHOB), photoelectric absorption factor (PEF), neutron porosity (NPHI), deep resistivity (ILD), and compression-wave velocity (Vp). Comparing the prediction performance accuracy achieved by each algorithm evaluated, it was found that the DT algorithm outperforms the other three predictive models in terms of performance prediction accuracy ($R^2 = 0.9985$ and RMSE $=$ 14.460 psi). Finally, the generalizability of the best-performing model, DT, is assessed by applying the DT model to an additional dataset collected from another (Well C) in the same field for PP prediction. The results undoubtedly proved that the DT model can be used throughout the field under study to predict PP since it presented a high accuracy in making PP predictions for the Well C dataset. For future studies, the PP modeling can be integrated with geological modeling to insight into geological parameters' effect on PP.

CRediT authorship contribution statement

Guodao Zhang: Conceptualization, Data curation, Investigation, Resources, Validation, Writing – review & editing. **Shadfar Davoodi:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft. **Shahab S. Band:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft. **Hamzeh Ghorbani:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft. **Amir Mosavi:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft. **Massoud Moslehpour:** Conceptualization, Data curation, Visualization, Writing - review & editing.

Declaration of competing interest

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

Data availability

Since the data used in the study are confidential, the authors cannot share them publicly.

Acknowledgment

This research was supported by the Tomsk Polytechnic University, Russia development program.

Nomenclature		
ANN	=	Artificial neural network algorithm
ARE	=	Average relative error
AARE	=	Absolute average relative error
b	=	Bias
DT	=	Decision tree
DTCO	=	Compressional wave transit times
DTSM	=	Shear wave transit times
ET	=	Extra trees
FP	=	Fracture pressure
F(X)	=	Random forest prediction function
GR	=	Gamma-ray
LR	=	Linear regression
M	=	Number of independent regresssion trees
ML	=	Machine learning
MSE	=	Mean squared error
OOB	=	Out-of-bag
PHIT	=	Total porosity
PP	=	Pore pressure
RBF	=	Radial basis function kernel
RF	=	Random forest
RHOB	=	Density
RMSE	=	Root mean squared error
STD	=	Standard deviation
SVR	=	Support vector regression
w	=	Weight vector
x	=	Data variable value range
X	=	Value of variable x in a specific data record
Z(x)	=	Feature mapping to high-dimensional space
ϑ	=	Poisson's ratio
ϑp	=	Compressional velocity
ϑs	=	Shear wave velocity

Appendix

The results for two alternative splits of data (80% training, 10% validation, and testing 10%; 60% training 20% validation, and 20% testing) employed for all data records are presented in Tables A.1 and A.2. These results should be compared with those presented in Table 12 for the split of data records in 70% training, 15% validation, and 15% testing.

Table A.1
Prediction accuracy for PP predictions by the four algorithms developed based on the total subset (for the split data record of 80% training, 10% validation, and testing 10%).

Models	ARE	AARE	STD	MSE	RMSE	R^2
Units	(%)	(%)	(psi)	(psi)	(psi)	–
ANN	−0.038	3.639	138.241	1.883E+04	137.213	0.8567
SVR	−0.054	2.441	129.890	1.666E+04	129.070	0.8994
RF	0.060	1.027	98.164	9.910E+03	99.547	0.9563
DT	−0.019	0.165	12.224	1.531E+02	12.375	0.9987

Table A.2
Prediction accuracy for PP predictions by the four algorithms developed based on the total subset (for the split data record of 60% training, 20% validation, and 20% testing).

Models	ARE	AARE	STD	MSE	RMSE	R^2
Units	(%)	(%)	(psi)	(psi)	(psi)	–
ANN	−0.087	2.574	128.123	1.650E+04	128.447	0.8811
SVR	−0.076	2.129	119.518	1.416E+04	119.001	0.9004
RF	0.052	1.954	107.451	1.151E+04	107.269	0.9462
DT	−0.014	0.131	11.372	1.217E+02	11.0334	0.9993

References

Abad, A.R.B., Ghorbani, H., Mohamadian, N., Davoodi, S., Mehrad, M., Aghdam, S.K.-y., et al., 2022. Robust hybrid machine learning algorithms for gas flow rates prediction through wellhead chokes in gas condensate fields. Fuel 308, 121872. http://dx.doi.org/10.1016/j.fuel.2021.121872.

Abad, A.R.B., Mousavi, S., Mohamadian, N., Wood, D.A., Ghorbani, H., Davoodi, S., et al., 2021a. Hybrid machine learning algorithms to predict condensate viscosity in the near wellbore regions of gas condensate reservoirs. J. Natural Gas Sci. Eng. 95, 104210. http://dx.doi.org/10.1016/j.jngse.2021.104210.

Abad, A.R.B., Tehrani, P.S., Naveshki, M., Ghorbani, H., Mohamadian, N., Davoodi, S., et al., 2021b. Predicting oil flow rate through orifice plate with robust machine learning algorithms. Flow Meas. Instrum. 102047. http://dx.doi.org/10.1016/j.flowmeasinst.2021.102047.

Abdelaal, A., Elkatatny, S., Abdulraheem, A., 2021. Data-driven modeling approach for pore pressure gradient prediction while drilling from drilling parameters. ACS Omega http://dx.doi.org/10.1021/acsomega.1c01340.

Abidin, M.H., 2014. Pore pressure estimation using artificial neural network. http://utpedia.utp.edu.my/id/eprint/14317.

Ahedor, M.K.-N., Anumah, P., Sarkodie-Kyeremeh, J., 2020. Post-drill pore pressure and fracture gradient analyses of Y-field in the offshore tano basin of Ghana. OnePetro http://dx.doi.org/10.2118/203659-MS.

Ahmad, M.S., Adnan, S.M., 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. http://dx.doi.org/10.1016/j.conbuildmat.2020.118475.

Ahmad, M.W., 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. Prod. 203, 810–821. http://dx.doi.org/10.1016/j.jclepro.2018.08.207.

Bahmaei, Z., Hosseini, E., 2020. Pore pressure prediction using seismic velocity modeling: Case study, Sefid–Zakhor gas field in Southern Iran. J. Pet. Explor. Prod. Technol. 10 (3), 1051–1062, https://link.springer.com/article/10.1007/s13202--019-00818-y.

Barjouei, H.S., Ghorbani, H., Mohamadian, N., Wood, D.A., Davoodi, S., Moghadasi, J., et al., 2021. Prediction performance advantages of deep machine learning algorithms for two-phase flow rates through wellhead chokes. J. Pet. Explor. Prod. Technol. 11 (3), 1233–1261.

Belhaj, A.F., Elraies, K.A., Alnarabiji, M.S., Kareem, F.A.A., Shuhli, J.A., Mahmood, S.M., et al., 2021. Experimental investigation binary modelling and artificial neural network prediction of surfactant adsorption for enhanced oil recovery application. Chem. Eng. J. 406, 127081. http://dx.doi.org/10.1016/j.cej.2020.127081.

Biot, M.A., Willis, D.G., 1957. The Elastic Coefficients of the Theory of Consolidation. http://dx.doi.org/10.1115/1.4011606.

Bowers, G.L., 1995. Pore pressure estimation from velocity data: Accounting for overpressure mechanisms besides undercompaction. SPE Drill. Complet. 10 (02), 89–95. http://dx.doi.org/10.2118/27488-PA.

Choubineh, A., Ghorbani, H., Wood, D.A., Moosavi, S.R., Khalafi, E., Sadatshojaei, E., 2017. Improved predictions of wellhead choke liquid critical-flow rates: Modelling based on hybrid neural network training learning based optimization. Fuel 207, 547–560. http://dx.doi.org/10.1016/j.fuel.2017.06.131.

Darvishpour, A., Seifabad, M.C., Wood, D.A., Ghorbani, H., 2019. Wellbore stability analysis to determine the safe mud weight window for sandstone layers. Pet. Explor. Dev. 46 (5), 1031–1038. http://dx.doi.org/10.1016/S1876-3804(19)60260-0.

Eaton, B.A., 1975. The equation for geopressure prediction from well logs. OnePetro http://dx.doi.org/10.2118/5544-MS.

Farsi, M., Barjouei, H.S., Wood, D.A., Ghorbani, H., Mohamadian, N., Davoodi, S., et al., 2021b. Prediction of oil flow rate through orifice flow meters: Optimized machine-learning techniques. Measurement 174, 108943. http://dx.doi.org/10.1016/j.measurement.2020.108943.

Farsi, M., Mohamadian, N., Ghorbani, H., Wood, D.A., Davoodi, S., Moghadasi, J., et al., 2021a. Predicting formation pore-pressure from well-log data with hybrid machine-learning optimization algorithms. Nat. Resour. Res. 1–27, https://link.springer.com/article/10.1007/s11053-021-09852-2.

Fu, G.-H., Wu, Y.-J., Zong, M.-J., Pan, J., 2020. Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data. BMC bioinformatics 21 (1), 1–14, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3411-3.

Ghorbani, H., Moghadasi, J., Wood, D.A., 2017. Prediction of gas flow rates from gas condensate reservoirs through wellhead chokes using a firefly optimization algorithm. J. Natural Gas Sci. Eng. 45, 256–271. http://dx.doi.org/10.1016/j.jngse.2017.04.034.

Ghorbani, H., Wood, D.A., Choubineh, A., Mohamadian, N., Tatar, A., Farhangian, H., et al., 2020. Performance comparison of bubble point pressure from oil PVT data: Several neurocomputing techniques compared. Exp. Comput. Multiph. Flow 2 (4), 225–246. http://dx.doi.org/10.1007/s42757-019-0047-5.

Ghorbani, H., Wood, D.A., Moghadasi, J., Choubineh, A., Abdizadeh, P., Mohamadian, N., 2019. Predicting liquid flow-rate performance through wellhead chokes with genetic and solver optimizers: An oil field case study. J. Pet. Explor. Prod. Technol. 9 (2), 1355–1373, https://link.springer.com/article/10.1007/s13202-018-0532-6.

Grape, S., Branger, E., Elter, Z., Balkeståhl, L.P., 2020. Determination of spent nuclear fuel parameters using modelled signatures from non-destructive assay and random forest regression. Nucl. Instrum. Methods Phys. Res. A 969, 163979. http://dx.doi.org/10.1016/j.nima.2020.163979.

Haris, A., Sitorus, R.J., Riyanto, A., 2017. Pore pressure prediction using probabilistic neural network: Case study of South Sumatra basin. IOP Conf. Ser. Earth Environ. Sci. 62, 012021. http://dx.doi.org/10.1088/1755-1315/62/1/012021.

Hashemitaheri, M., Mekarthy, S.M.R., Cherukuri, H., 2020. Prediction of specific cutting forces and maximum tool temperatures in orthogonal machining by support vector and Gaussian process regression methods. Procedia Manuf. 48, 1000–1008. http://dx.doi.org/10.1016/j.promfg.2020.05.139.

Hassanpouryouzb, A., Joonaki, E., Edlmann, K., Haszeldine, R.S., 2021. Offshore geological storage of hydrogen: Is this our best option to achieve net-zero? ACS Energy Lett. 6, 2181–2186. http://dx.doi.org/10.1021/acsenergylett.1c00845.

Hazbeh, O., Aghdam, SK-y, Ghorbani, H., Mohamadian, N., Alvar, M.A., Moghadasi, J., 2021b. Comparison of accuracy and computational performance between the machine learning algorithms for rate of penetration in directional drilling well. Pet. Res. http://dx.doi.org/10.1016/j.ptlrs.2021.02.004.

Hazbeh, O., Ahmadi Alvar, M., Aghdam, K-y, Ghorbani, H., Mohamadian, N., Moghadasi, J., 2021a. Hybrid computing models to predict oil formation volume factor using multilayer perceptron algorithm. J. Pet. Min. Eng. 14–27. http://dx.doi.org/10.21608/JPME.2021.52149.1062.

Hidayat, F., Astsauri, TMS., 2021. Applied random forest for parameter sensitivity of low salinity water injection (LSWI) implementation on carbonate reservoir. Alex. Eng. J. http://dx.doi.org/10.1016/j.aej.2021.06.096.

Hottmann, C.E., Johnson, R.K., 1965. Estimation of formation pressures from log-derived shale properties. J. Pet. Technol. 17 (06), 717–722. http://dx.doi.org/10.2118/1110-PA.

Hu, L., Deng, J., Zhu, H., Lin, H., Chen, Z., Deng, F., et al., 2013. A new pore pressure prediction method-back propagation artificial neural network. Electron J. Geotech. Eng. 18, 4093–4107, www.ejge.com/2013/Ppr2013.371mlr.pdf.

Jain, A., Zongker, D., 1997. Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19 (2), 153–158. http://dx.doi.org/10.1109/34.574797.

Jotheeswaran, J., Koteeswaran, S., 2020. Sentiment polarity classification using conjure of genetic algorithm and differential evolution methods for optimized feature selection. Recent Adv. Comput. Sci. Commun. 13 (6), 1284–1291. http://dx.doi.org/10.2174/2213275911666180904110105.

Karmakar, M., Maiti, S., 2019. Short term memory efficient pore pressure prediction via Bayesian neural networks at bering sea slope of IODP expedition 323. Measurement 135, 852–868, https://onepetro.org/ARMAUSRMS/proceedings-abstract/ARMA18/All-ARMA18/ARMA-2018-1098/124075.

Kartal, F., Özveren, U., 2020. A deep learning approach for prediction of syngas lower heating value from CFB gasifier in Aspen plus®. Energy 209, 118457. http://dx.doi.org/10.1016/j.energy.2020.118457.

Larestani, A., Mousavi, S.P., Hadavimoghaddam, F., Hemmati-Sarapardeh, A., 2022. Predicting formation damage of oil fields due to mineral scaling during water-flooding operations: Gradient boosting decision tree and cascade-forward back-propagation network. J. Pet. Sci. Eng. 208, 109315. http://dx.doi.org/10.1016/j.petrol.2021.109315.

Legg, S., Hutter, M., 2007. A collection of definitions of intelligence. Front. Artif. Intell. Appl. 157, 17.

Li, S., George, J., Purdy, C., 2012. Pore-pressure and wellbore-stability prediction to increase drilling efficiency. J. Pet. Technol. 64 (02), 98–101. http://dx.doi.org/10.2118/144717-JPT.

Lieder, F., Griffiths, T.L., 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behav. Brain Sci. 43. http://dx.doi.org/10.1017/S0140525X1900061X.

Lorena, A.C., de Carvalho, A.C., 2007. Protein cellular localization prediction with support vector machines and decision trees. Comput. Biol. Med. 37 (2), 115–125. http://dx.doi.org/10.1016/j.compbiomed.2006.01.003.

Maddahi, I., Moradzadeh, A., Nejati Kalateh, A., 2020. Comparison of pore pressure prediction using conventional seismic velocity and acoustic impedance-based methods. J. Pet. Res. 29 (109), 96–107. http://dx.doi.org/10.22078/PR.2019.3771.2719.

Mahetaji, M., Brahma, J., Sircar, A., 2020. Pre-drill pore pressure prediction and safe well design on the top of Tulamura anticline, Tripura, India: A comparative study. J. Pet. Explor. Prod. Technol. 10 (3), 1021–1049, https://link.springer.com/article/10.1007/s13202-019-00816-0.

Moosavi, S.R., Vaferi, B., Wood, D.A., 2021. Auto-characterization of naturally fractured reservoirs drilled by horizontal well using multi-output least squares support vector regression. Arab. J. Geosci. 14 (7), 1–12, https://link.springer.com/article/10.1007/s12517-021-06559-9.

Mousavipour, F., Riahi, M.A., Moghanloo, H.G., 2020. Prediction of in situ stresses, mud window and overpressure zone using well logs in south pars field. J. Pet. Explor. Prod. Technol. 10 (5), 1869–1879. http://dx.doi.org/10.2118/189665-PA.

Naveshki, M., Naghiei, A., Soltani Tehrani, P., Ahmadi Alvar, M., Ghorbani, H., Mohamadian, N., et al., 2021. Prediction of bubble point pressure using new hybrid computationail intelligence models. J. Chem. Pet. Eng. http://dx.doi.org/10.22059/JCHPE.2021.314719.1341.

Oloruntobi, O., Falugba, O., Ekanem-Attah, O., Awa, C., Butt, S., 2020. The Niger delta basin fracture pressure prediction. Environ. Earth Sci. 79 (13), 1–11, https://link.springer.com/article/10.1007/s12665-020-09081-5.

Osei-Bryson, K.-M., 2004. Evaluation of decision trees: A multi-criteria approach. Comput. Oper. Res. 31 (11), 1933–1945. http://dx.doi.org/10.1016/S0305-0548(03)00156-4.

Poole, D., Mackworth, A., Goebel, R., 1998. Computational intelligence.

Rajabi, M., Beheshtian, S., Davoodi, S., Ghorbani, H., Mohamadian, N., Radwan, A.E., et al., 2021. Novel hybrid machine learning optimizer algorithms to prediction of fracture density by petrophysical data. J. Pet. Explor. Prod. Technol. 1–23, https://link.springer.com/article/10.1007/s13202-021-01321-z.

Ranaee, E., Ghorbani, H., Keshavarzian, S., Abarghoei, P.G., Riva, M., Inzoli, F., et al., 2021. Analysis of the performance of a crude-oil desalting system based on historical data. Fuel 291, 120046. http://dx.doi.org/10.1016/j.fuel.2020.120046.

Rashidi, M., Asadi, A., 2018. An artificial intelligence approach in estimation of formation pore pressure by critical drilling data. OnePetro.

Richards, G., Roberts, D., Bere, A., Martinez, S., Tilita, M., Harrold, T., 2020. Pore Pressure Prediction Based on the Full Effective Stress (FES) Method, first ed. European Association of Geoscientists & Engineers, pp. 1–5. http://dx.doi.org/10.3997/2214-4609.202038004.

Rui, J., Zhang, H., Zhang, D., Han, F., Guo, Q., 2019. Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. J. Pet. Sci. Eng. 180, 699–706. http://dx.doi.org/10.1016/j.petrol.2019.06.014.

Russell, S., Norvig, P., 2002. Artificial intelligence: A modern approach. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/27702.pdf.

Salehi, M., Farhadi, S., Moieni, A., Safaie, N., Ahmadi, H, 2020. Mathematical modeling of growth and paclitaxel biosynthesis in corylus avellana cell culture responding to fungal elicitors using multilayer perceptron-genetic algorithm. Front. Plant Sci. 11. http://dx.doi.org/10.3389/fpls.2020.01148.

Shah, S.M.S., Shah, F.A., Hussain, S.A., Batool, S., 2020. Support vector machines-based heart disease diagnosis using feature subset wrapping selection and extraction methods. Comput. Electr. Eng. 84, 106628. http://dx.doi.org/10.1016/j.compeleceng.2020.106628.

Shahbaz, M., Taqvi, S.A., Loy, A.C.M., Inayat, A., Uddin, F., Bokhari, A., et al., 2019. Artificial neural network approach for the steam gasification of palm oil waste using bottom ash and CaO. Renew. Energy 132, 243–254. http://dx.doi.org/10.1016/j.renene.2018.07.142.

Shamshirb, S., Fathi, M., Chronopoulos, A.T., Montieri, A., Palumbo, F., Pescapè, A., 2020. Computational intelligence intrusion detection techniques in mobile cloud computing environments: Review, taxonomy, and open research issues. J. Inform. Secur. Appl. 55, 102582. http://dx.doi.org/10.1016/j.jisa.2020.102582.

Shamshirb, S., Fathi, M., Dehzangi, A., Chronopoulos, A.T., Alinejad-Rokny, H., 2021. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. J. Biomed. Inform. 113, 103627. http://dx.doi.org/10.1016/j.jbi.2020.103627.

Shamshirb, S., Rabczuk, T., Chau, K.-W., 2019. A survey of deep learning techniques: Application in wind and solar energy resources. IEEE Access 7, 164650-66. http://dx.doi.org/10.1109/ACCESS.2019.2951750.

Shao, M., Wang, X., Bu, Z., Chen, X., Wang, Y., 2020. Prediction of energy consumption in hotel buildings via support vector machines. Sustainable Cities Soc. 57, 102128.

Shen, Y., Luan, G., Zhang, H., Liu, Q., Zhang, J., Ge, H., 2017. Novel method for calculating the effective stress coefficient in a tight sandstone reservoir. KSCE J. Civ. Eng. 21 (6), 2467. http://dx.doi.org/10.1007/s12205-016-0514-5.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.

Terzaghi, K., 1943. Theoretical Soil Mechanics. John Wiley & Sons, New York.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer science & business media.

Wahab, M.N.A., Nefti-Meziani, S., Atyabi, A., 2015. A comprehensive review of swarm optimization algorithms. PLoS One 10 (5), 1–36. http://dx.doi.org/10.1371/journal.pone.0122827.

Wang, Y., Ma, H., Fu, W., 2010. Formation pressure prediction based on hybrid genetic algorithm. IEEE 2535–2538. http://dx.doi.org/10.1109/ICOSP.2010.5656925.

Yoshida, C., Ikeda, S., Eaton, B.A., 1996. An investigative study of recent technologies used for prediction, detection, and evaluation of abnormal formation pressure and fracture pressure in North and South America. OnePetro http://dx.doi.org/10.2118/36381-MS.

Yu, H., Chen, G., Gu, H., 2020. A machine learning methodology for multivariate pore-pressure prediction. Comput. Geosci. 143, 104548. http://dx.doi.org/10.1016/j.cageo.2020.104548.

Zhang, Y., Lv, D., Wang, Y., Liu, H., Song, G., Gao, J., 2020. Geological characteristics and abnormal pore pressure prediction in shale oil formations of the Dongying depression, China. Energy Sci. Eng. 8 (6), 1962–1979. http://dx.doi.org/10.1002/ese3.641.

Zhou, X., Lu, P., Zheng, Z., Tolliver, D., Keramati, A., 2020. Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. Reliab. Eng. Syst. Saf. 200, 106931. http://dx.doi.org/10.1016/j.ress.2020.106931.