Fuel 350 (2023) 128623

Contents lists available at ScienceDirect

Fuel

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

Full length article

Development of hybrid computational data-intelligence model for flowing bottom-hole pressure of oil wells: New strategy for oil reservoir management and monitoring

Leonardo Goliatt [a,*], Reem Sabah Mohammad [b], Sani I. Abba [d], Zaher Mundher Yaseen [c,d,*]

[a] Department of Applied and Computational Mechanics, Federal University of Juiz de Fora, Juiz de, Fora, 36036-900, Brazil
[b] University of Misan, Kahla Road, Misan, Iraq
[c] Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[d] Interdisciplinary Research Centre for Membranes and Water Security, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

ARTICLE INFO

Keywords:
Hybrid machine learning models
Flowing bottom-hole pressure
Feature selection
Oil and gas well management

ABSTRACT

Among several metric parameters concerning the assessment of oil and gas well production, the flowing bottom-hole pressure (FBHP) is considered essential. Accurate prediction of FBHP is crucial for petroleum engineering and management. Several related parameters are associated with the FBHP magnitude influence; thus, proper inspection of those parameters is another vital concern. This research proposes a hybrid modeling framework based on the hybridization of machine learning (ML) models (i.e., Extreme Learning Machine (ELM), Support Vector Machine Regressor (SVR), Extreme Gradient Boosting (XGB), and Multivariate Adaptive Regression Spline (MARS)) and nature-inspired Differential Evolutional (DE) optimization for FBHP prediction. The adjustment of the internal parameters of the ML-based models and the input feature selection is formulated as an incremental learning problem that is solved by the evolutionary algorithm. Problem-specific samples were collected from the open-source literature for this investigation. Modeling results are adaptable, automatically determining the most relevant variables for the context of the ML model. The adaptive polynomial structure of hybridized MARS model attained the best average performance for the FBHP modeling with correlation (R = 0.94) and minimum root mean square (RMSE = 97.88). The proposed modeling framework produces an alternative efficient computer aid model for FBHP prediction, resulting in reliable automated technology to assist oil and gas well management.

1. Introduction

As the global primary fuel sources, oils and gas industries have become the world's dominant energy industries and have a significant role in today's global economy [1]. Despite the adverse impacts of oils and gas sectors, such as climate change, biodiversity, and environmental footprint, these sectors have a significant potential to contribute to sustainable development goals (SDGs) [2]. According to Millikan and Sidwell [3], a precise understanding of pressure, especially at the bottom of an oil well, is the most critical in petroleum engineering. Hence, it is essential to identify the practical methods and system recovery. FBHP is the fundamental parameter in reservoir analysis and oil and gas production processes. It is an indicator for evaluating the performance of the economic design of the wells [4]. A widely engaging problem in the petroleum industry is accurately quantifying bottom pressure during the real-time multi-phase flow of liquids [5–8].

The literature reports several approaches to determine FBHP depending on the type and specific functions of the well [9]. The installation of pressure gauges at the bottom of the well is the most straightforward approach to measuring FBHP. However, several associated weaknesses include calibration, risk, and frequent maintenance [10]. Several researchers have proposed conventional and empirical approaches to predict bottom-hole pressure in exploration and exploitation wells, including correlation-based models. However, these methods are mainly formulated on a laboratory scale and are subject to error, uncertainty, and low precision [5,8]. Furthermore, the complex nature of oil and gas production may not satisfy the assumptions of physics-based equations for many reasons [11], such as non-ideal fluid behavior, chaotic fluid pattern, heterogeneity, and anisotropy. Consequently, conventional, mechanical, and other empirical techniques prove insufficient in effectively managing these complex conditions.

* Corresponding authors.
E-mail addresses: leonardo.goliatt@ufjf.br (L. Goliatt), reem-sabah@uomisan.edu.iq (R.S. Mohammad), sani.abba@kfupm.edu.sa (S.I. Abba), z.yaseen@kfupm.edu.sa (Z.M. Yaseen).

https://doi.org/10.1016/j.fuel.2023.128623
Received 18 October 2022; Received in revised form 30 April 2023; Accepted 3 May 2023

Machine Learning (ML) models have confirmed the potential to simulate historical data observations of petroleum engineering [12–15]. ML models have evidenced their practical application in almost every industry with enormous possibilities for growth and innovation [16–18], causing ML to rise to prominence in the scientific spheres [19,20]. The attributes of ML are its capacity to acquire knowledge from raw data, handle nonlinear tasks, accommodate defective data in a fault-tolerant manner, and provide generalizations efficiently and accurate predictions [21]. Modern technology includes various soft computing models that have evidenced their potential in solving nonlinear prediction related issues, such as artificial neural networks (ANN) [5,22–25], support vector regression (SVR) [20,26–29], fuzzy logic (FL) [8,30], and adaptive neuro-fuzzy logic (ANFIS) [28,31–33].

Previous literature indicated that empirical and ML-based models were employed to predict FHBP in different scenarios. Nevertheless, the former and latter were attributed to low accuracy, generalization, or real-expert implementation. Tariq et al. [34] reported that different scholars had proposed different ML models to mitigate problems with data and modeling. Another unique challenge in modeling FHBP is the input variable combination and selection approach; this is generally a problem when dealing with a large set of input variables in the field of data modeling and simulation [35–39]. It is evident that the correlation-based techniques mainly proposed for calculating the FBHP [40–45], involved several mathematical assumptions, and numerals variables that need a considerable amount of information and perhaps debated about the failure to provide accurate results [46]. It is worth mentioning that the recent technical literature has criticized correlation-based input variable selection [47–51]. Hadi et al. [35] reported that feature selection and input identification are essential steps for any intelligent data algorithm. Evolutionary data mining techniques are innovative approaches for detecting patterns, anomalies, and linkages between complicated processes in massive databases, which can be used to predict future trends. Because interactions among parameters follow a complex process, developing new models with high accuracy is critical. As a result, these strategies are ideally suited to exploiting the vast volumes of real-time, multivariate data being generated for hydrocarbon exploration systems.

Depending on their intended use, hybrid approaches could undergo prediction or optimization [52,53]. Consequently, it is justifiable that hybrid strategies consist of several or integrated single methodologies and optimization algorithms that have proven to be more dependable and capable of exceeding single models in terms of modeling precision [35,39,54–58], hybrid learning has proven to be not only beneficial and superior to single models but also covers a wide range of issues that are connected with single procedures [59].

Hence, the primary contribution of this research is to propose a new hybrid computer aid model based on the hybridization of ML models and an evolutionary algorithm for predicting the FHBP of oil wells. The merit of the proposed modeling framework is the feasibility of being automated in configuring the essential input parameters for the prediction matrix. The expected outcome of the proposed hybrid model is to reduce the complexity of the learning process and attain a more reliable predictive model that can be satisfactorily implied for oil well monitoring and management. The remainder of this research on evolutionary machine learning with feature selection is organized as follows. First, the dataset statistical properties are presented. Then the computational framework is then described in depth, and feature selection and machine learning model parameter search are shown as optimization problems. Next, computational experiments are presented and discussed based on performance metrics and error and uncertainty analysis, highlighting the advantages and limits of the proposed approach. Finally, further research and conclusion are given.

2. Material and methods

This section provides context for the evolutionary feature and model selection approaches proposed in this paper. The details on the ML models are described in Appendix

2.1. Flowing bottom hole flowing pressure dataset

A multi-phase flow dataset from vertical wells consisting of 206 samples was obtained from open resources [60]. The wells where the data was collected flowed with artificial lifting processes. Downhole flow pressure is recorded during measurements using downhole pressure gauges just above the boreholes. The target variable to be predicted is the downhole flow pressure (FBHP), represented by the dependent variable Pwf (psia), modeled using nine production-related variables: oil flow (Qo (bbl/day)), inside diameter of the production pipeline (ID (inches)), gas flow (Qg (Mscf/ day)), water flow (Qw (bbl/day)), oil density (Api), well-drilling depth (Depth (ft)), downhole temperature (Bt (F)), temperature of the surface (ST (F)), and wellhead pressure (Pwh (psia)). The output is the flowing bottom-hole pressure (Pwf (psia)). The main motivation for adopting a computer aid model for FBHP prediction is to provide reliable alternative technology for oil and gas wells management and operation.

The dataset was divided into a training set consisting of 165 samples and a test set of 41 samples. Tables 1 and 2 show the basic statistics for the training and test sets. The datasets are available at https://github.com/LGoliatt/fbhp_hybrid.

Fig. 1 shows the correlation coefficients among input variables and the relationship between the input variables and the floating bottom hole pressure (Pwf). The coefficient ranges between +1 and −1, where +1 represents a direct correlation between the variables and −1 an indirect relationship between the two variables. In Fig. 1, one can observe a strong correlation between the oil flow (Qo) and the gas flow (Qg). This strong positive correlation (0.92) is expected because the flow is multiphase. In addition, a strong positive correlation is observed between oil density (Api) and surface temperature (St) because the density of drilling fluids, particularly oil-based fluids, varies with pressure and temperature [61]. As seen in the last line of the correlation matrix in Fig. 1, the drilling depth and water flow coefficients are moderate, 0.62 and 0.5, respectively. Besides, a low positive correlation of FBHP with wellhead pressure (0.37) and oil density (0.33) is observed. The remaining input variables have a weak correlation with the FBHP value.

2.2. Computational model for the hybrid approach

The Differential Evolution (DE) algorithm [62] was used to find the best parameters for the ELM, SVM, XGB, and MARS predictors. DE is a population-based stochastic evolutionary algorithm for solving global optimization problems. Using the difference between two randomly generated vectors, the technique causes a disturbance in the solutions [63]. In this approach, each individual in the population encodes a candidate solution (an ML model) to model FBHP.

After initializing the population, the selection operator chooses the parents according to a mix of criteria combining their fitness and a random component. Then, new candidate solutions are generated using operators such as mutation, recombination, and fitness evaluation are carried out sequentially. At the end of each generation, the fittest individuals are selected for the next generation. Following is a description of the DE algorithm in depth. First, an initial population of candidate solutions $\{\theta_{i,G} \mid i = 1, 2, \ldots, NP\}$ is created at random. Then, the DE iteratively performs the operations [64]:

- Selection: for each θ_i, select three parents r_1, r_2, and r_3 at random in the population.
- Mutation: generate a new candidate mutated solution $v_{i,G+1} = \theta_{r1,G} + F(\theta_{r2,G} - \theta_{r3,G})$ where $F \in (0, 2)$ is a scaling parameter.
- Crossover: generates a vector

$$\mu_{ji,G+1} = \begin{cases} v_{ji,G+1}, & \text{if } rand\ b(j) \le CR \text{ or } j = rnbr(i), \\ x_{ji,G}, & \text{if } rand\ b(j) > CR \text{ or } j \ne rnbr(i), \end{cases}$$

where $rand\ b(j) \in [0, 1]\ \forall j$. CR $\in [0, 1]$ is the crossover probability, and $rnbr(i) \in 1, 2, \ldots, D$ ensures $\mu_{i,G+1}$ is a different candidate solution from $v_{i,G+1}$.

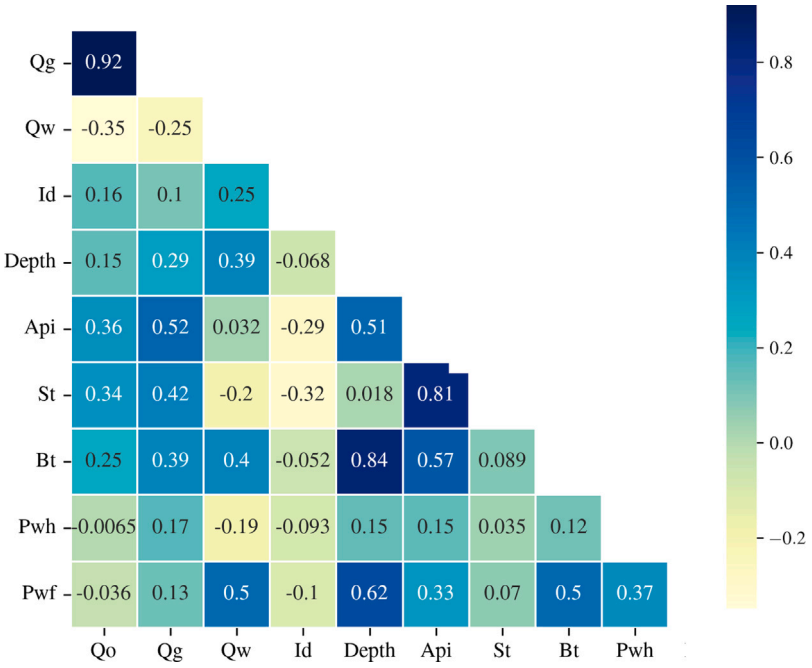

Fig. 1. Correlation coefficient among input variables and floating bottom hole pressure (Pwf).

Table 1
Training set (165 samples) basic statistics. The first column identifies all variables in the prediction problem. The target variable is the flowing bottom-hole pressure, indicated in the last line (Pwf (psia)). The second column (mean) shows the average value of the variable, the third column (std) is the associated standard deviation, and the fourth column presents its minimum observed value (min). Interquartile values are shown in the fifth, sixth, and seventh columns, respectively. The upper bound of the first interquartile range is identified in column 25%, the value of the second interquartile range, or median, is shown in column 50%, and the upper bound of the third interquartile range is shown in column 75%. The last column shows the maximum observed value for the respective variable in the dataset..

Variable	Mean	Std	Min	25%	50%	75%	Max
Qo (bbl/day)	6257.36	5031.25	280.00	2350.00	4700.00	9600.00	19618.00
Qg (mscf/day)	3364.90	3049.08	33.60	1012.30	2448.75	4949.02	13562.20
Qw (bbl/day)	2621.88	2744.54	0.00	1.00	1794.00	4600.00	11000.00
Id (inches)	3.81	0.43	2.00	3.81	3.96	3.96	4.00
Depth (ft)	6382.23	541.47	4550.00	6317.00	6518.00	6709.00	7100.00
Api	33.88	2.31	30.00	32.60	32.60	36.50	37.00
St (f)	118.79	31.15	76.00	90.00	97.00	155.00	160.00
Bt (f)	204.02	16.40	157.00	208.00	212.00	212.00	215.00
Pwh (psia)	320.81	153.62	80.00	210.00	280.00	390.00	960.00
Pwf (psia)	2500.98	305.17	1227.00	2296.00	2505.00	2710.00	3217.00

Table 2
Test set (41 samples) basic statistics. The first column identifies all variables in the prediction problem. The target variable is the flowing bottom-hole pressure, indicated in the last line (Pwf (psia)). The second column (mean) shows the average value of the variable, the third column (std) is the associated standard deviation, and the fourth column presents its minimum observed value (min). Interquartile values are shown in the fifth, sixth, and seventh columns, respectively. The upper bound of the first interquartile range is identified in column 25%, the value of the second interquartile range, or median, is shown in column 50%, and the upper bound of the third interquartile range is shown in column 75%. The last column shows the maximum observed value for the respective variable in the dataset..

Variable	Mean	Std	Min	25%	50%	75%	Max
Qo (bbl/day)	6579.71	3993.18	840.00	3700.00	5600.00	9293.00	14800.00
Qg (mscf/day)	3622.02	3175.20	75.20	1413.09	2904.80	4730.14	12580.00
Qw (bbl/day)	3014.41	2995.32	0.00	10.00	2288.00	5496.00	10500.00
Id (inches)	3.93	0.07	3.81	3.81	3.96	3.96	4.00
Depth (ft)	6269.88	656.49	4550.00	6245.00	6406.00	6722.00	7079.00
Api	33.34	2.35	30.00	32.60	32.60	36.50	37.00
St (f)	113.46	29.28	90.00	90.00	90.00	153.00	159.00
Bt (f)	202.12	19.17	161.00	208.00	212.00	212.00	215.00
Pwh (psia)	322.15	155.24	130.00	210.00	275.00	400.00	800.00
Pwf (psia)	2440.98	288.36	1906.00	2190.00	2444.00	2679.00	2984.00

- Replacement: if the new generated candidate solution $\mu_{i,G+1}$ is fittest than $\theta_{i,G}$, then $\mu_{i,G+1}$ replaces $\theta_{i,G+1}$ and $\theta_{i,G}$ is kept otherwise.

Searching for the optimal parameters includes selecting the model's performance-enhancing characteristics. This phase was implemented by enabling and disabling binary representation features. Each candidate

Table 3
Encoding of internal parameters of candidate solutions. The description of the mathematical model of the ML models can be found in Appendix.

Model	θ^{MS}	Description	Range/Set
ELM	θ_1	Regularization parameter, C	[1, 10000]
	θ_2	Regularization strategy, γ	0: L_1, 1: L_2
SVR	θ_1	Regularization parameter, C	[1, 10000]
	θ_2	Bandwidth parameter, γ	[10^{-5}, 1000]
	θ_3	Kernel, ϕ	0: linear, 1: RBF, 2: sigmoid
XGB	θ_1	No. predictors	[10, 300]
	θ_2	Learning rate	[10^{-6}, 1]
	θ_3	L_2 weight regularization (λ)	[0, 1]
	θ_4	L_1 weight regularization (α)	[0, 1]
MARS	θ_1	Degree of polynomials, q	[0,3]
	θ_2	Penalty factor, γ	[1, 9]
	θ_3	Number of terms, M	[1, 500]

solution $\theta = (\theta^{FS}, \theta^{MS})$ encodes an ML model and a subset of features, as shown in Table 3. The available ML models are described in Appendix. Let the θ^{FS} vector be the set of features as shown in Tables 1 and 2. The binary representation consists of each entry being set to 0 or 1, referring respectively to the inactive and activated features. As an example, a vector of features

$$[\text{Qo (bbl/day)}, \text{Qg (bbl/day)}, \text{Api}, \text{Bt (f)}]$$

is represented as $\theta^{FS} = [1, 1, 0, 0, 0, 1, 0, 1, 0]$.

The vector θ^{FS} corresponds to the internal model parameters, which are specified as follows:

- Extreme Learning Machine (ELM): $\theta^{MS} = (\theta_1, \theta_2)$, where θ_1 corresponds to the regularization parameter C and θ_2 sets the regularization approach (L_1 or L_2).
- Support Vector Machine Regression (SVR): $\theta^{MS} = (\theta_1, \theta_2, \theta_3)$. The first parameter corresponds to the regularization parameter C, the second parameter corresponds to the bandwidth parameters γ, and the last one represents the kernel identification.
- Extreme Gradient Boosting (XGB): $\theta^{MS} = (\theta_1, \theta_2, \theta_3, \theta_4)$, where θ_1 symbolizes the number of predictors, θ_2 is the learning rate, θ_3 and θ_4 are the L_1 and L_2 penalty regularization term, respectively
- Multivariate Adaptive Regression Spline (MARS): $\theta^{MS} = (\theta_1, \theta_2, \theta_3)$, where θ_1 is the maximum degree of the piecewise polynomials, θ_2 encodes the penalty factor and (θ_3) corresponds the maximum number of terms in the polynomial model.

The objective of the Differential Evolution (DE) strategy lies in the fine-tuning of internal parameters within the predictor and the identification of a subset of features that yield computed outcomes consistent with the actual results derived from the training data. A simplified diagram of the approach presented in this paper is displayed in Fig. 2.

3. Application results and analysis

This section presents the results obtained of the proposed modeling framework for gathered dataset from the available literature. The performance results of the hybrid model without the feature selection process and the comparison with others with similar strategies are presented in Section 3.1. The evolutionary feature selection strategy is evaluated in Section 3.2. The following packages were used to implement the framework: pandas [65], NumPy [66], statsmodels [67], scikit-learn framework [68], adaptations of the source codes provided by Friedman [69] and Virtanen et al. [70]. The experiments were conducted on a computer with the following specifications: CPU Intel i7-9700F Opteron (4500 Mhz, eight cores of 12.00 GHz, and cache memory of 2 MB), RAM of 32 GB under operating system Linux Ubuntu 18.04.1.

Table 4
Performance metrics.

Name	Expression
R^2	$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{N}(y_i - \bar{y})$
MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}\rvert$
RMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$
MAPE	$\frac{100}{N}\sum_{i=1}^{N}\left\lvert\frac{y_i - \hat{y}}{y_i}\right\rvert$

In ML-based model development, statistical assessment is usually adopted to determine the performance of models. The assessment can be done with the help of metrics that allow you to quantify and compare performance. Table 4 describes the metrics used to assess the model's performance. These metrics were chosen because they can capture different characteristics of the models' behavior and allow comprehensive comparisons.

3.1. Performance of the hybrid approach without model selection

According to Table 5, the proposal presented in this study yielded consistent results across all criteria. The metrics showed that the modeling performance was significantly enhanced by utilizing feature selection. It is also clear from comparing the MARS model's best-averaged values (in boldface) to the averages obtained by the other models that the MARS model consistently showed accurate results across the separate runs.

The XGB model is compatible with the MARS model in terms of performance, considering all metrics. In contrast, ELM and SVM had poor performance and great variability in the final predictions, as evidenced by their standard deviations (shown in parenthesis). Considering the comparison with the results in the literature, the available results only present the correlation coefficient value shown in the first column. Orkiszewski [71] compares several empirical models while Sami and Ibrahim [11] implemented a the neural network to predict FBHP. The results present the best models reported by the authors. It can be observed that the models developed in this research outperform (in average performance) the best models found in the literature for the same problem.

The models presented in Table 5 use all input variables collected in the field. The input variables are described in Tables 1 and 2. These models' attributes are critical in cases where sensor failures can occur, in addition to helping construct simplified models that can potentially be less susceptible to noise when reading the data. A practical research point is finding models with fewer variables, less complexity, and similar predictive power to those with all available input variables. The evolutionary process model is presented in the next section to accomplish this task.

3.2. Evolutionary model and feature selection approach

The automatic feature determination process was evaluated to identify a subset of pertinent features for inclusion in model construction. Evolutionary algorithms are powerful tools that can identify a few features most useful in predicting the target variable. The advantage of using evolutionary algorithms for feature selection is that they can handle a large number of features and can find a suitable solution even if the search space is highly nonlinear.

Table 6 shows the models' performance metrics with the evolutionary feature selection framework. The suffix -FS indicates that the feature selection procedure was used to execute the models. The values in bold indicate the model that obtained the best performance in average values in the analyzed metrics. The results demonstrate that

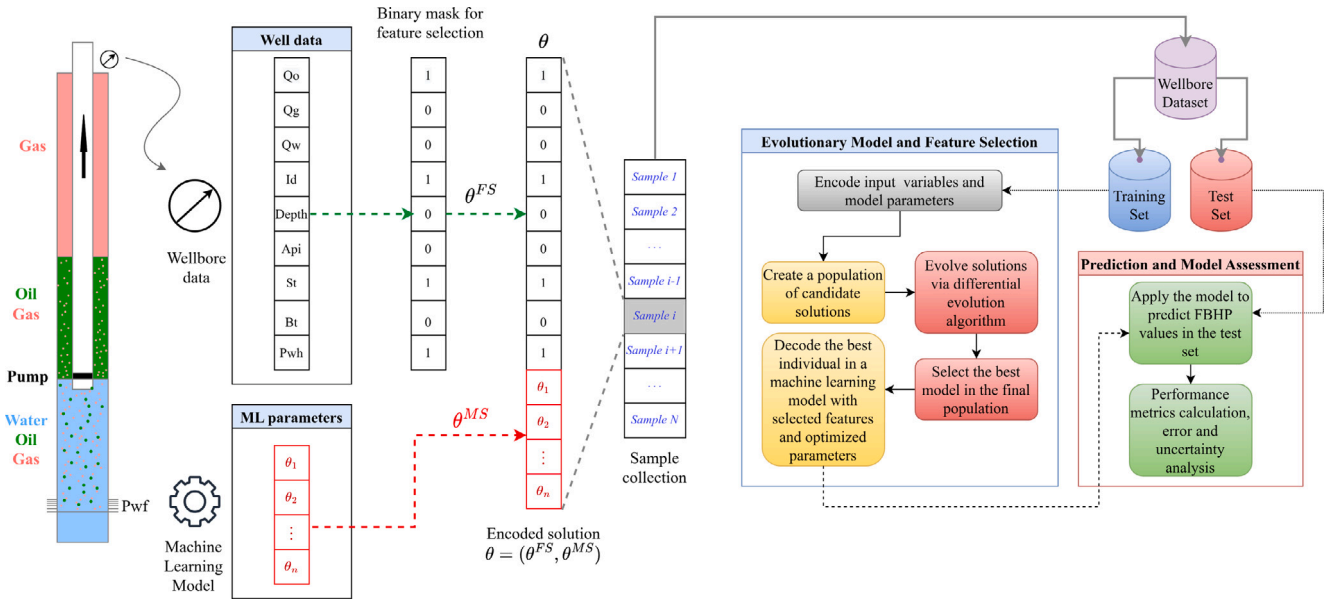

Fig. 2. Schematic representation of the evolutionary feature and model selection framework. The data collected from the wellbore composes the input parameters θ^{FS} while the ML internal parameters are encoded into the vector θ^{MS}. A binary mask is used to select the input variables: an entry equal to 1 indicates the input variable is chosen to build the ML model, and 0 otherwise. The vector $\theta = (\theta^{FS}, \theta^{MS})$ encodes a candidate solution for the problem. The problem is formulated as an incremental learning problem (where what must be learned are the input variables and the internal parameters of the ML model). An intelligent search algorithm then solves the learning problem. The differential of the proposal is to determine the most relevant variables as the fittest solution in an evolutionary process, where the best solutions result in the most accurate models.

Table 5
Averaged results for performance metrics. In this experiment, the nine variables are frozen, and the models use all of them to predict the FBHP. The values between parentheses indicate the standard deviation in 50 independent runs. Entries in boldface indicate the best averaged values.

ML Model	R	R^2	RMSE (psia)	MAE (psia)	MAPE (%)
ELM	0.883 (0.136)	0.588 (1.35)	141.53 (115.67)	92.13 (21.63)	3.88 (1.01)
MARS	**0.940** (0.016)	**0.879** (0.031)	**98.21** (12.19)	**75.26** (10.19)	**3.13** (0.428)
SVR	0.768 (0.044)	0.197 (0.432)	247.43 (62.89)	205.28 (58.57)	8.66 (2.41)
XGB	0.925 (0.010)	0.852 (0.019)	109.22 (6.80)	84.29 (6.39)	3.53 (0.257)
N/A[a]	0.902 (–)	–	–	–	< 10
RF[b]	0.83 (–)	–	–	–	–
KNN[b]	0.86 (–)	–	–	–	–
ANN[b]	0.93 (–)	–	–	–	–

[a]Reported by Orkiszewski [71]
[b]Reported by Sami and Ibrahim [11]

Table 6
Averaged results for the evolutionary feature selection framework to predict the FBHP values in the test set. The values within parentheses indicate the standard deviation in 50 independent runs. Entries highlighted in bold indicate the best averaged values.

ML Model	R	R^2	RMSE (psia)	MAE (psia)	MAPE (%)
ELM-FS	0.913 (0.021)	0.828 (0.042)	117.50 (13.46)	88.97 (11.03)	3.74 (0.462)
MARS-FS	**0.931 (0.019)**	**0.862 (0.039)**	**104.87 (14.31)**	**80.29 (11.85)**	**3.35 (0.493)**
SVR-FS	0.805 (0.019)	0.377 (0.323)	218.85 (51.23)	184.09 (48.02)	7.81 (1.96)
XGB-FS	0.922 (0.018)	0.847 (0.032)	110.74 (11.15)	85.09 (9.46)	3.57 (0.393)

the MARS-FS is superior to other models across all criteria. MARS-FS showed more prominent predictability due to its modeling flexibility, as described in Appendix A.4. Its multivariate adaptive spline modeling flexibility allowed for achieving the best performances for all metrics, even with an internal evolutionary process that reduces the number of input variables.

Table 7 presents the computing time (CPU time) obtained for the proposed hybrid approach. The SVR model (without feature selection) produced the lowest averaged CPU times, followed by ELM, XGB, and MARS. The mathematical formulation of the SVR using the kernel to express the internal products in large dimensions saves computational resources resulting in a shorter CPU execution time. On the other hand, the CPU time of the ELM model depends on the number of neurons in the hidden layer, which requires more operations to perform the training when compared to SVR. The XGB uses an additive modeling formulation using decision trees, while the MARS model employs an additive mathematical formulation with high-order polynomials that results in higher CPU processing. The feature selection procedure implemented in ELM-FS, SVR-FS, XGB-FS and MARS-FS does not significantly affect CPU time. Although the proposed approach allows training the model with fewer variables, there were no significant variations in the CPU times, as observed in Table 7.

Comparing the results presented in Tables 5 and 6, we also observed that the performance was slightly lower than the model that used all the input variables. This degradation in performance is expected because less information was used to build the models. However, comparing the MARS-FS model and the MARS model, this average performance decay was 0.96% for the R metric and 1.98% for the R^2 metric, 6.35%

Table 7
Averaged CPU time in seconds with standard deviations (calculated on 50 independent runs). Computer specifications: Intel i7-9700F Opteron CPU (4500 Mhz, eight 12.00 GHz cores and 2 MB cache memory), 32 GB RAM under Linux Ubuntu 18.04.1 operating system.

ML Model	CPU Time (s)
ELM	6.406 ± 0.549
ELM-FS	6.630 ± 0.768
MARS	9.036 ± 1.477
MARS-FS	9.258 ± 1.376
SVR	6.056 ± 0.620
SVR-FS	6.277 ± 0.811
XGB	8.142 ± 0.619
XGB-FS	8.062 ± 0.644

for the RMSE, 6.26% for the MAE, while a 6.75% percentage decrease was computed for the mean error percentage (MAPE). For comparison purposes, it is noted that the performance decrease was smaller for the XGB model (−0.33% for R, −0.59% for R2, −1.37% for RMSE, −0.94% and −1.12% for MAPE). However, the model achieved a worse performance compared to the MARS model. Therefore, a slight drop in the percentual difference is expected as the performance improves.

Fig. 3 shows the distribution of variables active in the final models after 50 independent runs. The input variable occurrences strongly indicate the importance of the variable in the context of the model within the evolutionary feature selection process. It can be seen that the set of input variables (Depth, Pwh, Qw) appears in all models (EN, SVR, XGB, and MARS) in all independent runs. This result indicates that, regardless of the ML model, the set, as mentioned earlier, of variables is fundamental in the modeling process. Specifically for the MARS-FS model, the variable Id is essential in predicting the FBHP. The pipeline's inside diameter (Id) is an important design parameter that affects the fluid's flow rate and bottom pressure. Increasing the pipeline Id reduces friction losses and increases the flow rate. The Id is decreased to increase the bottom pressure. Additionally, for the MARS-FS model, the input variables Qw proved to be equally important, and the variable Qw is also representative since they were selected in 46 out of 50 runs.

Similarly, based on previous outcomes, the model MARS-FS was chosen for further analysis of the distribution of features. After model evolution, the final models' results were evaluated, and the variables selected by the evolutionary procedure were stored and presented in Table 8. The purpose of the table is to show the distribution of the occurrence of the groups of variables, the number of input variables in the final models, and the number of times they occurred during independent runs. The first column shows the input variables used to model the FBHP output, while the second column displays the count of input variables in the model. The third column indicates the number of occurrences of the input variables set in 50 independent runs. The remaining columns present the averaged performance metrics associated with each model. The standard deviation values indicated with (–) were not displayed as they occurred only once and did not allow computation. The table rows were arranged from the highest average performing model to the lowest average performing model.

Considering the number of input variables in the models, the most effective models have between seven and eight variables. It can be noticed that among the most accurate models, the variables Api and St do not appear among the selected ones. Fig. 3 supports this observation, showing that these two variables were the least likely to be chosen in different runs of the evolutionary feature selection model. Interestingly, the model with nine variables occurred more times in 50 runs (6 out of 50), indicating that the feature selection process led to the original model without feature selection. On the other hand, this 9-variable model was not the best performer, indicating that a reduction in input variables can improve FBHP predictions.

Fig. 4 displays the Taylor diagram for the models MARS, XGB, MARS-FS and XGB-FS. The results of all 50 final models are displayed. The Taylor diagram was constructed based on predicted and measured FBHP values in the test set. Taylor's diagram visually depicts how well predictions and measured values match based on the correlation coefficient (R), the centered root mean square deviation (RMSD), and the standard deviation. The diagram shows that models XGB and MARS generated results close to the observed data, with higher R and lower RMSD values. In contrast, the agreement quality with the observed data decreases for XGB-FS and MARS-FS.

4. Discussion

The knowledge acquired using the feature selection is used in Section 4.1 to freeze some variables and propose simpler models. An error analysis is performed in the 4.2 section, while in 4.3 section, an uncertainty analysis on the developed models is performed. Finally, Section 4.4 discusses the strengths and limitations of the model.

4.1. Performance of hybrid approach freezing input variables

The proposed approach allows the freezing of some variables during the search process. This approach works as the insertion of domain-specific knowledge by the specialist, ensuring that a group of variables are always present in the prediction model [72]. In contrast, specific variables are deliberately discarded from the modeling process. This flexibility of the computational framework allows searching for a specific group of variables while some remain unsearchable [59].

An advantage of the evolutionary feature selection process is that it allows learning about the most important input variables for the model and data being analyzed. This knowledge can build simpler models with similar or superior precision to previously determined models. In freezing variables, it is important to determine which ones will be inserted and which ones will be included in the modeling process. A group of specialists can determine this criterion by considering the production conditions or the costs involved. Alternatively, in this study, we chose the six most frequent variables of the MARS model. This choice is based on Fig. 3, where the six most frequent variables are (Qw, Id, Depth, Pwh, Qo, Bt).

Table 9 shows the results obtained with three different groups of variables. The V4 group is formed by the six most frequent variables in Fig. 3, the V3 group by the five most frequent variables, and the V1 group by the four most frequent variables. The objective is to evaluate how the MARS model behaves in predictive capacity with a gradual reduction of information available in the input variables. The average results in Table 9 show a decrease in performance with the reduction in the number of input variables. This result is expected because the model produces worse predictions as relevant input variables are excluded. On the other hand, when comparing the results in Tables 5 and 9, it is noted that the MARS-V4 model effectively fitted to the data, allowing an improvement in the average performance for the analyzed metrics.

Fig. 5 shows the scatter plot of the best MARS-V4 model with the variable set (Qw, Id, Depth, Pwh, Qo, Bt). This result shows that including the most relevant variables benefits the FBHP modeling, allowing the model to use the predictive capacity to adjust to the most relevant information. Furthermore, including more variables may not increase the prediction performance because the model has to deal with more potentially irrelevant information. We emphasize that the final model results from a search process, so non-informative variables can include noise in the search and make the evolutionary algorithm more challenging to produce reasonable solutions. This problem can be overcome by increasing the number of generations of the evolutionary algorithms, but with it also comes an increase in computational cost and complexity to find a solution.

Table 8
Summary of results for MARS model simulations with evolutionary feature selection over 50 independent runs. Column Set represents the variables present in the final model. Column No. Var. shows the number of input variables, and column Occur. displays the number of occurrences of the variable set out of 50 runs. The remaining columns display the performance metrics R, R^2, RMSE, MAE, and MAPE, respectively.

Set	No. Var.	Occur.	R	R^2	RMSE (psia)	MAE (psia)	MAPE (%)
Api, Bt, Depth, Id, Pwh, Qg, Qw	7	1	0.951 (–)	0.900 (–)	90.12 (–)	72.39 (–)	2.98 (–)
Api, Bt, Depth, Id, Pwh, Qg, Qw, St	8	1	0.951 (–)	0.905 (–)	87.99 (–)	70.70 (–)	2.93 (–)
Api, Bt, Depth, Id, Pwh, Qo, Qw, St	8	2	0.949 (0.00)	0.895 (0.003)	92.40 (1.48)	66.33 (3.45)	2.79 (0.155)
Bt, Depth, Id, Pwh, Qg, Qo, Qw, St	8	5	0.948 (0.013)	0.893 (0.025)	92.48 (10.59)	73.44 (8.10)	3.07 (0.338)
Bt, Depth, Id, Pwh, Qg, Qo, Qw	7	1	0.946 (–)	0.896 (–)	92.04 (–)	64.50 (–)	2.74 (–)
Bt, Depth, Id, Pwh, Qo, Qw, St	7	1	0.946 (–)	0.894 (–)	92.95 (–)	63.72 (–)	2.69 (–)
Api, Bt, Depth, Id, Pwh, Qo, Qw	7	4	0.938 (0.024)	0.877 (0.046)	98.69 (17.65)	78.18 (13.90)	3.21 (0.610)
Bt, Depth, Id, Pwh, Qg, Qw, St	7	1	0.936 (–)	0.876 (–)	100.41 (–)	78.57 (–)	3.30 (–)
Api, Bt, Depth, Id, Pwh, Qg, Qo, Qw	8	4	0.935 (0.009)	0.873 (0.016)	101.26 (6.32)	75.10 (6.79)	3.12 (0.268)
Api, Depth, Id, Pwh, Qo, Qw	6	2	0.930 (0.003)	0.860 (0.005)	106.45 (1.86)	79.12 (1.25)	3.38 (0.054)
Api, Bt, Depth, Id, Pwh, Qg, Qo, Qw, St	9	6	0.928 (0.028)	0.856 (0.059)	106.49 (20.35)	78.28 (14.75)	3.26 (0.609)
Bt, Depth, Id, Pwh, Qo, Qw	6	6	0.927 (0.019)	0.855 (0.039)	107.47 (14.83)	83.16 (12.95)	3.47 (0.542)
Api, Depth, Id, Pwh, Qg, Qw	6	3	0.927 (0.015)	0.854 (0.034)	108.49 (12.91)	87.79 (9.68)	3.68 (0.435)
Depth, Id, Pwh, Qg, Qo, Qw	6	1	0.924 (–)	0.850 (–)	110.20 (–)	85.85 (–)	3.60 (–)
Depth, Id, Pwh, Qg, Qo, Qw, St	7	3	0.922 (0.017)	0.842 (0.032)	112.69 (11.22)	81.78 (12.13)	3.41 (0.461)
Depth, Id, Pwh, Qo, Qw	5	2	0.917 (0.033)	0.838 (0.064)	113.62 (22.79)	89.86 (17.04)	3.79 (0.732)
Api, Depth, Id, Pwh, Qo, Qw, St	7	1	0.916 (–)	0.836 (–)	115.30 (–)	92.07 (–)	3.87 (–)
Api, Depth, Id, Pwh, Qg, Qo, Qw, St	8	5	0.914 (0.012)	0.828 (0.024)	117.79 (7.80)	91.16 (7.11)	3.77 (0.302)
Api, Depth, Id, Pwh, Qg, Qo, Qw	7	1	0.894 (–)	0.788 (–)	131.26 (–)	102.51 (–)	4.24 (–)

Table 9
Results for freezing variables strategy during the search process.

ML Model	Var. Set	No. Var.	R	R^2	RMSE (psia)	MAE (psia)	MAPE (%)
MARS	V1	4	0.849 (0.007)	0.691 (0.015)	158.20 (3.77)	122.07 (2.32)	5.21 (0.100)
MARS	V3	5	0.913 (0.022)	0.825 (0.050)	118.16 (15.89)	89.90 (9.82)	3.77 (0.402)
MARS	V4	6	**0.940** (0.016)	**0.880** (0.032)	**97.88** (12.43)	**74.69** (10.52)	**3.12** (0.442)

Variable sets:
V1: (Qw, Id, Depth, Pwh)
V3: (Qw, Id, Depth, Pwh, Qo)
V4: (Qw, Id, Depth, Pwh, Qo, Bt).

Table 10
Error analysis for MARS, XGB, MARS-FS, XGB-FS machine learning models. MARS freezing variables were also included in the comparison.

ML Model	Input variable set	Mean prediction error (psia)	Width of uncertainty band (psia)	95% confidence prediction error interval (psia)
MARS	All	−0.00	±0.014	0.939 to 1.07
XGB	All	+0.002	±0.018	0.918 to 1.08
MARS-FS	–	+0.002	±0.015	0.930 to 1.07
XGB-FS	–	+0.002	±0.017	0.921 to 1.07
MARS	V1	+0.009	±0.027	0.867 to 1.11
MARS	V3	+0.002	±0.017	0.922 to 1.08
MARS	V4	+0.00	±0.013	0.942 to 1.06

4.2. Error analysis

An error analysis was conducted to assess the models' ability to predict FBHP outcomes. The error e_j for the sample j is the difference between the FBHP value measured and the FBHP predicted by the ML models. The 95% confidence band around the predicted FBHP values is given by $(\bar{e} - 1.96S_e,\ \bar{e} + 1.96S_e)$, where $\bar{e} = \sum_{j=1}^{N} e_j$ is the mean of the prediction and $S_e = \sqrt{\sum_{k=1}^{N}(\bar{e} - e_k)^2/(N-1)}$ is the standard deviation.

The error analysis is displayed in Table 10. The results show that the models produced similar mean prediction errors (in the 3rd column), indicating that the evolutionary feature selection does not change the optimized XGB and MARS behavior. In addition, one can observe that except for model MARS-V1, all models produced similar uncertainty widths, as seen in the last column. MARS-V1 uses four input variables, which are reflected negatively in the model error analysis.

4.3. Uncertainty analysis

The uncertainty analysis is a method that can quantify the degree to which the output varies as a result of changes to the input. The calculation of statistical measures such as the median, mean, and population quantiles are typically used to carry out the analysis [73].

Uniform distributions were used to model the variations of input parameters. The maximum and minimum values for each input parameter are shown in Table 3. The Mean Absolute Deviation (MAD)

$$\mathrm{MAD} = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} |\mathrm{FBHP}_{p_i} - median(\mathrm{FBHP}_p)| \tag{1}$$

is used to provide an output's uncertainty that is calculated as

$$\text{Uncertainty } \% = \frac{100 \times \mathrm{MAD}}{median(\mathrm{FBHP}_p)} \tag{2}$$

where $N_{MC} = 250000$ and FBHP_{p_i} is the flow bottom-hole pressure predicted for the ith sample. The study by Sattar and Gharabaghi [74] provides more information on the Monte Carlo approach.

Uncertainty analysis for FBHP modeling is shown in Table 11. The results demonstrate that the XGB model exhibited the lowest uncertainty when utilizing all input variables, while the MARS model yielded the highest uncertainty. This disparity may suggest the potential overfitting of the MARS model. Considering models that use evolutionary feature selection, the uncertainty of the XGB-FS model is smaller than the MARS-FS model. The XGB is an ensemble model that combines multiple simple models (weak learners) with error control strategies, resulting in a more robust model with reduced uncertainty. Conversely, the MARS model, which combines polynomials of varying degrees, displayed higher susceptibility to input data variations. Notably, the MARS-V1, MARS-V3, and MARS-V4 models, which selectively froze certain variables and utilized only the most informative ones, exhibited lower uncertainty values, reinforcing the importance of properly selecting input variables.

Fig. 6 illustrates the graphical comparison concerning RMSE and uncertainty. The figure reveals that the employment of MARS models resulted in higher uncertainty and lower RMSE values, while XGB models exhibited lower uncertainty but higher RMSE values. Specifically, the

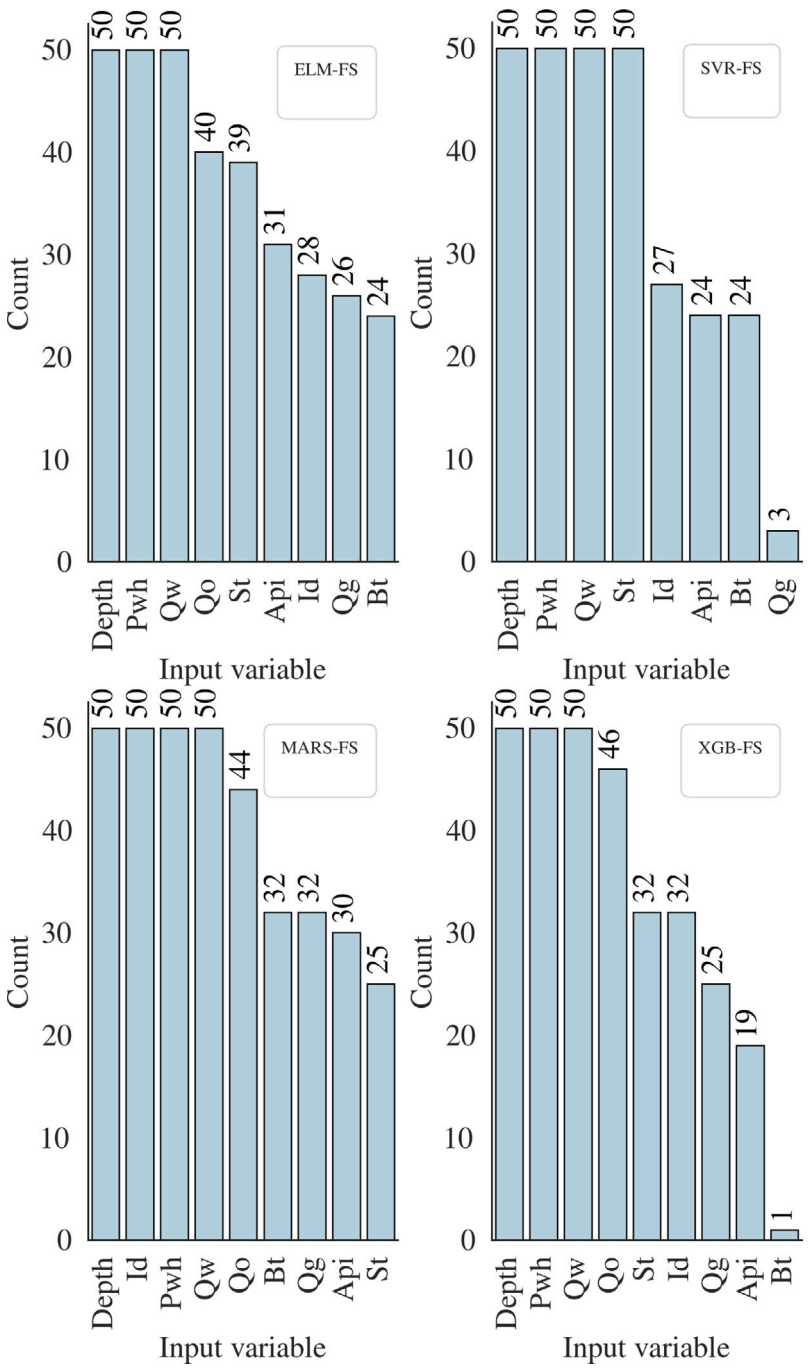

Fig. 3. Evolutionary feature selection: distribution of the input variables active in the final models after 50 independent runs.

Table 11
Uncertainty predicts for MARS, XGB, MARS-FS, XGB-FS machine learning models. MARS freezing variables were also included in the comparison.

ML Model	Var. Set.	No. features	Median (psia)	MAD (psia)	Uncertainty (%)	RMSE (psia)
MARS	All	9	3344.291	1741.008	52.059	78.002
XGB	All	9	2491.751	134.891	5.413	95.931
MARS-FS	–	8	2650.413	275.819	10.407	81.262
XGB-FS	–	8	2483.031	151.678	6.109	93.358
MARS	V1	4	2616.948	212.913	8.136	153.485
MARS	V3	5	2734.173	268.818	9.832	91.973
MARS	V4	6	2377.138	368.138	15.487	72.572

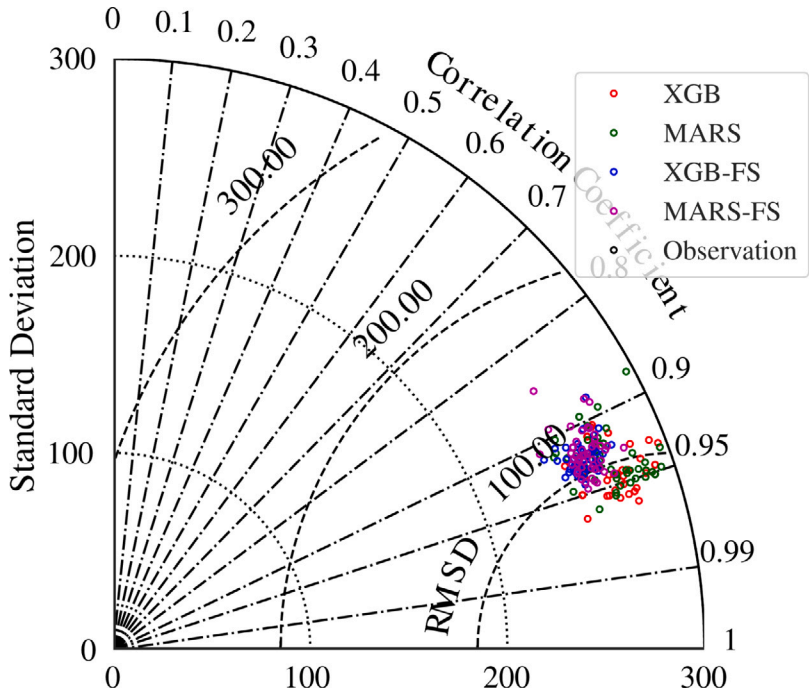

Fig. 4. Taylor diagram for MARS and XGB, including the evolutionary feature selection models.

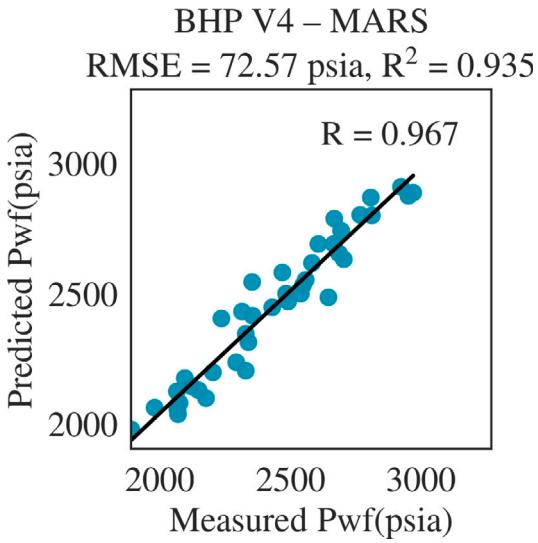

Fig. 5. Predictions of the best MARS-V4 model with the variable set (Qw, Id, Depth, Pwh, Qo, Bt).

MARS-V4 model outperformed other models regarding RMSE, whereas XGB models demonstrated superior performance regarding uncertainty. In the region where the models presented the best performance, a trade-off is observed between the RMSE values and uncertainties of gradient boosting-based models (XGB) and spline-based models (MARS). However, model MARS-V4 achieved similar uncertainty and lower RMSE with reduced complexity considering the number of variables.

4.4. Model strengths, limitations, and future developments

Recently, it has become common practice to combine meta-heuristics and machine learning models [75–78]. Although there are additional well-established feature selection techniques, they are typically used offline, either before or after the ML model is constructed. This approach can be interpreted as an incremental learning process that blends learning with an evolutionary algorithm [79]. A procedure for improving potential solutions is incorporated into the evolutionary model's framework. While internal parameters are being changed, the most important characteristics are chosen simultaneously, resulting in continuous model improvement throughout the evolutionary process.

The primary strength of the model lies in its capability to automatically identify the most relevant variables through the feature selection process and to determine the internal parameters of ML models. Instead of focusing on the model's parameters or deciding which variables to employ to model the input–output relationship, specialists can focus on the decision-making process. [80].

The current proposal also offers the flexibility to interchange ML models and optimization techniques. Different evolutionary algorithms can be integrated with island-based approaches [81], which can make use of the potential of diverse algorithms to produce valuable solutions. Since the model formulation and the internal parameters the evolutionary algorithm needs to explore are already known, integrating additional ML models into the framework is straightforward.

Over the past two decades, the applications of ML models have been adopted noticeably for the domain of oil and gas field related issues [82]. However, ML models developed for particular prediction/prediction situations can be implemented in online learning machine technology [83]. The development of modern technology for predicting oil wells FHBP using cutting-edge federated learning technology should be the focus of the next generation. FHBP detection "quantification" can be functionalized based on the viability and potential of the recently found federated learning technology. The key building blocks of the anticipated system are a distributed network architecture, smart Internet of Things (IoT) sensors, edge servers, and a centralized federated learning aggregation hub, as shown in Fig. 7. The primary phase of this intelligence system is field data, or "sensors", where the related data are collected. The second phase consists of the edge servers, in which all those related data from the first data are transmitted to the predictive model where this research was developed. The third phase accommodates the federated learning process that is communicated, and the learning process results in a centralized cloud. At the final stage, the updated cloud delivers the final prediction

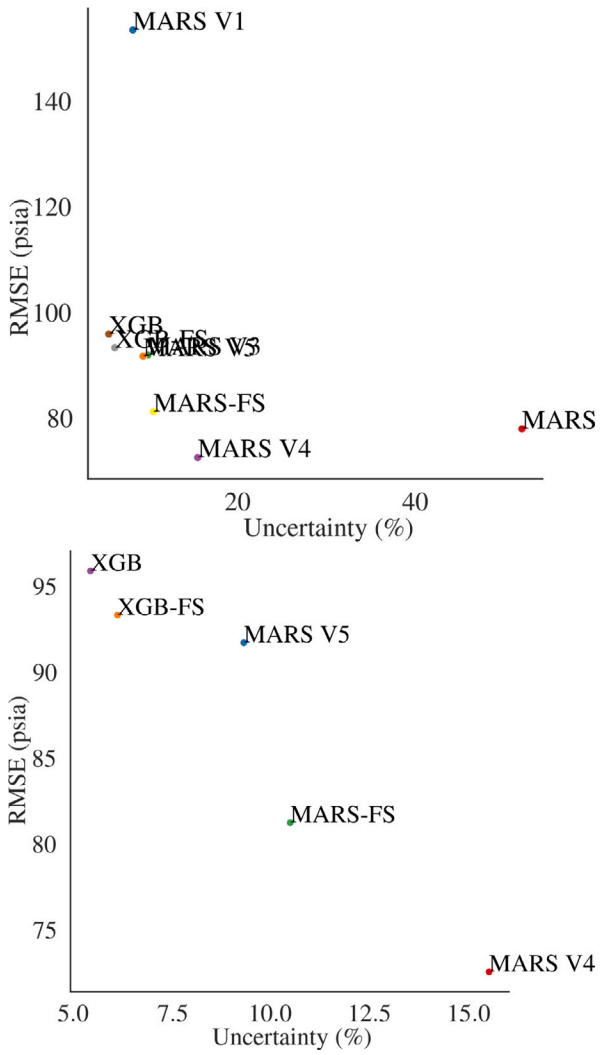

Fig. 6. Top: Graphical comparison concerning RMSE and uncertainty for all models shown in Table 11. Bottom: Detailed view of best-performing models.

results of the FHBP. The planned federated learning technology can be practically implemented, where decision-makers can benefit from this methodology for oil and gas management and sustainability.

5. Conclusion

This research contributes to developing a hybrid model integrating a feature selection algorithm with an ML-based tool for FBHP prediction. The computational framework is adaptive to different input data, allowing for automatic feature selection and internal ML model parameters. Four ML models were hybridized with nature-inspired differential evolution algorithms to model the flow bottom pressure. To demonstrate the effectiveness of the proposed modeling framework, standard ML models were developed. To ensure the stability and robustness of the proposal, 30 independent runs were performed, and error performance and uncertainty analyses were adopted. The modeling results indicated that MARS and XGB models produced the best results without a feature selection algorithm. The evolutionary MARS and XGB without feature selection outperformed ELM and SVR, as well as empirical and mathematical models from the literature. The evolutionary feature selection algorithm reduced the number of variables in the MARS model from 9 to 6. In addition, the MARS model produced the best averaged performance metrics (R = 0.94, R^2 = 0.88, RMSE = 97.88, MAE = 74.69 e MAPE = 3.12%). Including evolutionary feature selection, MARS, and XGB models become less complex since the number of input variables has been reduced, maintaining their error level. The proposed approach has the advantage of generating several alternative models with different sets of input variables, which allows for their practical use when some information from the wellbore cannot be retrieved due to sensor failure or communication malfunction.

CRediT authorship contribution statement

Leonardo Goliatt: Data curation, Methodology, Software, Formal analysis, Validation, Visualization, Writing – original draft. **Reem Sabah Mohammad:** Data curation, Formal analysis, Validation, Visualization, Writing – review & editing. **Sani I. Abba:** Data curation, Investigation, Validation, Writing – review & editing. **Zaher Mundher Yaseen:** Conceptualization, Data curation, Validation, 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

Data will be made available on request

Acknowledgments

The first author acknowledges the funding agencies CNPq (401796/2021-3, 307688/2022-4, and 409433/2022-5) and FAPEMIG (APQ-00334/18) for their financial support.

Appendix. Machine learning models

A.1. Extreme learning machine

The Extreme Learning Machine is a single-layer neural network with randomly chosen hidden input connection weights [84]. Its structure is simple, with fast convergence. Using a least squares formulation, the weights in the output layer are calculated by the inverse of the generalized Moore–Penrose inverse matrix multiplied by the outputs of activation functions in the output layer. The ELM output is written as

$$\hat{y} = \sum_{i=1}^{L} \beta_i G(\alpha_i \mathbf{x} + b_i) \tag{A.1}$$

where $\{(x_i, y_i), x_i \in R^n, y_i \in R^1, i = 1, 2, \ldots, N\}$ are the training samples, $\{\alpha_i,\ i = 1, 2, \ldots, N\}$ is the weight vector, $\{\beta_i,\ i = 1, 2, \ldots, N\}$ output weights, L the number of neurons, b_i is the bias for the ith hidden node, $G(\cdot, \cdot, \cdot)$ the neuron's activation function, and $\hat{y}$ is the model output.

Eq. (A.1) can be formulated as $\mathbf{H}\beta = \mathbf{T}$ where $H_{ij} = G(\boldsymbol{\alpha}_j, b_j, \mathbf{x}_i)$ and $T_i = y_i$. The least-squares method is used to minimize the objective quadratic function $\sum_{i=1}^{N} \|\hat{y}_i - y_i\| = 0$ and determine output weight vector $\beta = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T}$.

A.2. Support vector machines

Support Vector Regression (SVR) [85] is a regression model that uses linear or nonlinear using kernel functions to approximate the samples in the dataset. Given a dataset $(x_1, y_1), \ldots, (x_l, y_l)$, the SVR solution is obtained by solving the optimization problem

$$\min \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T K(\mathbf{x}_i, \mathbf{x}_j)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \sum_{i=1}^{l}(y_i + \varepsilon)(\alpha_i - \alpha_i^*) \tag{A.2}$$

subject to

$$e^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0,$$

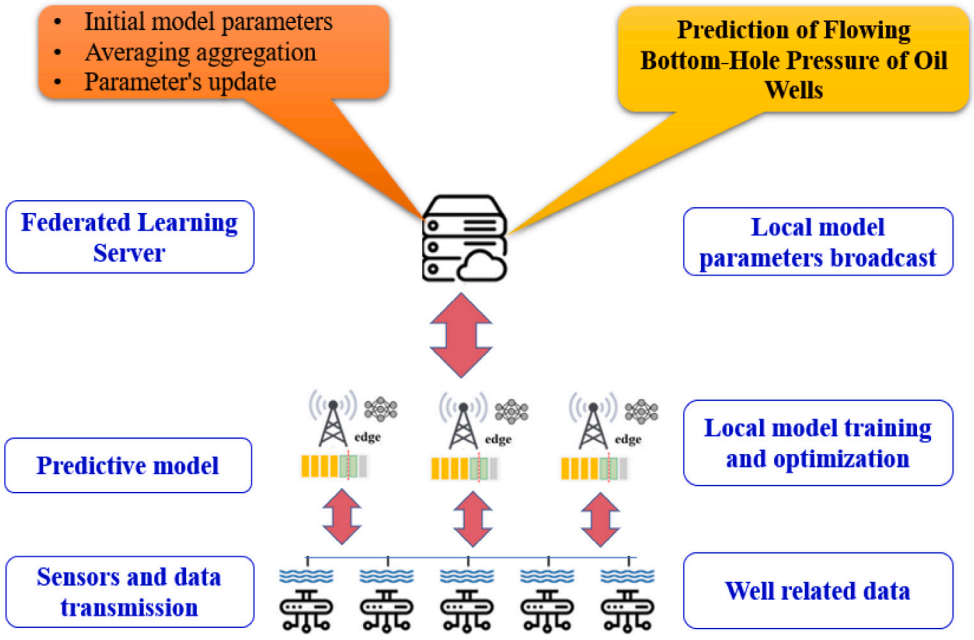

Fig. 7. The proposed intelligence system based on online federated learning technology for FHBP prediction.

$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, l.$

where $x_i \in R^n$ is the input vector (data samples), $y_i \in R^1$ is the output vector, $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$, $\phi(\cdot)$ is the kernel function. The model parameters are $\epsilon > 0$ and $C > 0$. Eq. (A.2) can be solved to obtain the parameters for constructing the SVR approximation. The following equation provides a prediction based on

$$\hat{y}_i = \epsilon \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(\mathbf{x}, \mathbf{x}_i) + b.$$

A.3. Extreme gradient boosting

Extreme Gradient Boosting (XGB) is a version of gradient boosting that is more effective in supervised learning. Ibrahem Ahmed Osman et al. [86]. The following steps are performed. Adopting a dataset with m features and a n number of samples $(x_1, y_1), \ldots, (x_n, y_n)$, the predicted output of XGB is

$$\phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathbb{F} \tag{A.3}$$

where $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $i = 0, \ldots, n$, K indicates the number of decision trees, f_k is the model of the kth decision tree, and the depth of f_k is represented by m_{depth}. To generate the XGB approximation is required to minimize the regularized loss function

$$L(\phi) = \sum_i l(y_i, \phi(x_i)) + \frac{1}{2}\|w\|^2, \quad l = \|\hat{y}_i - y_i\| \tag{A.4}$$

where $\hat{y}_i$ and y_i are the predicted and true output, respectively, and w is the weight of the leaf.

A.4. Multivariate adaptive regression spline

First proposed by [69], Multivariate Adaptive Regression Splines (MARS) is a piecewise polynomial model [87]

$$\hat{y}(x) = F_m(x) = c_0 + \sum_{m=1}^{M} c_m B_m^K(x) \tag{A.5}$$

where $B_m^K(x)$ is the basis function (BF), M is the number of BF, c_0 is a constant, c_m the coefficients of the mth basis function $B_m^K(x)$ given by

$$B_m^K(x) = \prod_{k=1}^{K} [\pm(x - t)]_+^q, \tag{A.6}$$

where K number of piecewise polynomials and q is the polynomial's degree.

The piecewise polynomial is represented by

$$\begin{aligned} [-(x-t)]_+^q &= \begin{cases} (t-x)^q & \text{if } x < t \\ 0 & \text{otherwise} \end{cases}, \\ [+(x-t)]_+^q &= \begin{cases} (x-t)^q & \text{if } x \geq t \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{A.7}$$

where t are the polynomial knots (points connected by the polynomial). The Generalized Cross-Validation (GCV) [88] is minimized and its solution gives the set of coefficients

$$\text{GCV} = \frac{\sum_{i=1}^{N} (\hat{y}_i(x) - y_i)^2}{N\left(1 - \frac{(M+1)+\gamma M}{N}\right)^2}, \tag{A.8}$$

where γ is a penalization coefficient.

References

[1] Cherepovitsyn A, Rutenko E, Solovyova V. Sustainable development of oil and gas resources: A system of environmental, socio-economic, and innovation indicators. J Mar Sci Eng 2021;9:1307.
[2] IPIECA U. IFC. Mapping the oil and gas industry to the sustainable development goals: An atlas. London: IPIECA, United Nations development programme. Int Finance Corp 2017.
[3] Millikan CV, Sidwell V. Bottom-hole pressures in oil wells. Trans AIME 1931;92:194–205.
[4] Ahmadi MA, Chen Z. Machine learning models to predict bottom hole pressure in multi-phase flow in vertical oil production wells. Can J Chem Eng 2019;97:2928–40.
[5] Awadalla M, Yousef H. Neural networks for flow bottom hole pressure prediction.. Int J Electr Comput Eng (2088-8708) 2016;6.
[6] El-Saghier RM, Abu El Ela M, El-Banbi A. A model for calculating bottom-hole pressure from simple surface data in pumped wells. J Pet Explor Prod Technol 2020;10:2069–77.
[7] Nait Amar M, Zeraibi N. A combined support vector regression with firefly algorithm for prediction of bottom hole pressure. SN Appl Sci 2020;2:1–12.

[8] Amar MN, Zeraibi N, Redouane K. Bottom hole pressure estimation using hybridization neural networks and grey wolves optimization. Petroleum 2018;4:419–29.

[9] Rathnayake S, Rajora A, Firouzi M. A machine learning-based predictive model for real-time monitoring of flowing bottom-hole pressure of gas wells. Fuel 2022;317:123524.

[10] Afanaskin I, Kryganov P, Volpin S, Kolevatov A, Glushakov A, Yalov P. Multi-well deconvolution issue solving for producing well with increasing water-cut through CRM-model application. J Pet Sci Eng 2022;110679.

[11] Sami NA, Ibrahim DS. Forecasting multiphase flowing bottom-hole pressure of vertical oil wells using three machine learning techniques. Pet Res 2021.

[12] Salem AM, Yakoot MS, Mahmoud O. Addressing diverse petroleum industry problems using machine learning techniques: Literary methodology - spotlight on predicting well integrity failures. ACS Omega 2022;7:2504–19.

[13] Khozani ZS, Khosravi K, Pham BT, Kløve B, Wan Mohtar WHM, Yaseen ZM. Determination of compound channel apparent shear stress: Application of novel data mining models. J Hydroinform 2019;21:798–811.

[14] Afan HA, Allawi MF, El-Shafie A, Yaseen ZM, Ahmed AN, Malek MA, et al. Input attributes optimization using the feasibility of genetic nature inspired algorithm: Application of river flow forecasting. Sci Rep 2020;10:1–15.

[15] Cui J, Sang Q, Li Y, Yin C, Li Y, Dong M. Liquid permeability of organic nanopores in shale: Calculation and analysis. Fuel 2017;202:426–34.

[16] Bughin J, Seong J, Manyika J, Chui M, Joshi R. Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute; 2018.

[17] Koroteev D, Tekic Z. Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. Energy AI 2021;3:100041.

[18] Organization W-WIP. WIPO technology trends 2019-artificial intelligence. 2019.

[19] Haenlein M, Kaplan A. Guest editorial to the special issue, a brief history of AI: On the past, present, and future of artificial intelligence. Calif Manage Rev 2019;61:5–14.

[20] Li H, Yu H, Cao N, Tian H, Cheng S. Applications of artificial intelligence in oil and gas development. Arch Comput Methods Eng 2021;28:937–49.

[21] Bello O, Holzmann J, Yaqoob T, Teodoriu C. Application of artificial intelligence methods in drilling system design and operations: A review of the state of the art. J Artif Intell and Soft Comput Res 2015;5:121–39.

[22] Ashena R, Moghadasi J. Bottom hole pressure estimation using evolved neural networks by real coded ant colony optimization and genetic algorithm. J Pet Sci Eng 2011;77:375–85.

[23] Jahanandish Ie, Salimifard B, Jalalifar H. Predicting bottomhole pressure in vertical multiphase flowing wells using artificial neural networks. J Pet Sci Eng 2011;75:336–42.

[24] Memon PQ, Yong S-P, Pao W, Pau JS. Dynamic well bottom-hole flowing pressure prediction based on radial basis neural network. In: Science and information conference. Springer; 2014, p. 279–92.

[25] Nwanwe CC, Duru UI, Anyadiegwu C, Ekejuba AIB. An artificial neural network visible mathematical model for real-time prediction of multiphase flowing bottom-hole pressure in wellbores. Pet Res 2022.

[26] Chen W, Di Q, Ye F, Zhang J, Wang W. Flowing bottomhole pressure prediction for gas wells based on support vector machine and random samples selection. Int J Hydrogen Energy 2017;42:18333–42.

[27] Liang H, Liu G, Zou J, Bai J, Jiang Y. Research on calculation model of bottom of the well pressure based on machine learning. Future Gener Comput Syst 2021;124:80–90.

[28] Marfo SA, Asante-Okyere S, Ziggah YY. A new flowing bottom hole pressure prediction model using M5 prime decision tree approach. Model Earth Syst Environ 2021;1–9.

[29] Zhang B, Wang J-l, Zhang N-s. New method for flow rate and bottom-hole pressure prediction based on support vector regression. In: International field exploration and development conference. Springer; 2019, p. 3812–29.

[30] Ozbayoglu EM, Miska SZ, Reed T, Takach N. Analysis of bed height in horizontal and highly-inclined wellbores by using artificial neuraletworks. In: SPE international thermal operations and heavy oil symposium and international horizontal well technology conference. OnePetro; 2002.

[31] Ali A, Guo L. Neuro-adaptive learning approach for predicting production performance and pressure dynamics of gas condensation reservoir. IFAC-PapersOnLine 2019;52:122–7.

[32] Awadalla MHA. Radial basis function neural network for predicting flow bottom hole pressure. Organization 2019;10.

[33] Nwanwe CC, Ilozurike.Duru U. An adaptive neuro-fuzzy inference system white-box model for real-time multiphase flowing bottom-hole pressure prediction in wellbores. Petroleum 2023.

[34] Tariq Z, Mahmoud M, Abdulraheem A. Real-time prognosis of flowing bottom-hole pressure in a vertical well for a multiphase flow using computational intelligence techniques. J Pet Explor Prod Technol 2020;10:1411–28.

[35] Hadi SJ, Abba SI, Sammen SS, Salih SQ, Al-Ansari N, Yaseen ZM. Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. IEEE Access 2019;7:141533–48.

[36] Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A. Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia. Neural Comput Appl 2017;28:893–905.

[37] Yaseen ZM, Allawi MF, Yousif AA, Jaafar O, Hamzah FM, El-Shafie A. Non-tuned machine learning approach for hydrological time series forecasting. Neural Comput Appl 2018;30:1479–91.

[38] Yaseen ZM. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. Chemosphere 2021;277:130126.

[39] Yaseen ZM, Ehteram M, Sharafati A, Shahid S, Al-Ansari N, El-Shafie A. The integration of nature-inspired algorithms with least square support vector regression models: Application to modeling river dissolved oxygen concentration. Water 2018;10:1124.

[40] Ansari A, Sylvester N, Sarica C, Shoham O, Brill J. A comprehensive mechanistic model for upward two-phase flow in wellbores. SPE Prod Facil 1994;9:143–51.

[41] Asheim H. MONA, an accurate two-phase well flow model based on phase slippage. SPE Prod Eng 1986;1:221–30.

[42] Aziz K, Govier GW. Pressure drop in wells producing oil and gas. J Can Pet Technol 1972;11.

[43] Gomez L, Shoham O, Schmidt Z, Chokshi R, Northug T. Unified mechanistic model for steady-state two-phase flow: Horizontal to vertical upward flow. SPE J 2000;5:339–50.

[44] Pucknell JK, Mason JNE, Vervest EG. An evaluation of recent mechanistic models of multiphase flow for predicting pressure drops in oil and gas wells. In: SPE offshore europe. OnePetro; 1993.

[45] Govier GW, Fogarasi M. Pressure drop in wells producing gas and condensate. J Can Pet Technol 1975;14.

[46] Ahmadi MA, Galedarzadeh M, Shadizadeh SR. Low parameter model to monitor bottom hole pressure in vertical multiphase flow in oil production wells. Petroleum 2016;2:258–66.

[47] Nourani V, Elkiran G, Abba SI. Wastewater treatment plant performance analysis using artificial intelligence–an ensemble approach. Water Sci Technol 2018;78:2064–76.

[48] Martinho AD, Saporetti CM, Goliatt L. Approaches for the short-term prediction of natural daily streamflows using hybrid machine learning enhanced with grey wolf optimization. Hydrol Sci J 2022;1–18.

[49] Elkiran G, Nourani V, Abba SI. Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. J Hydrol 2019;577:123962.

[50] Pham QB, Abba SI, Usman AG, Linh NTT, Gupta V, Malik A, et al. Potential of hybrid data-intelligence algorithms for multi-station modelling of rainfall. Water Resour Manag 2019;33:5067–87.

[51] Pham QB, Sammen SS, Abba SI, Mohammadi B, Shahid S, Abdulkadir RA. A new hybrid model based on relevance vector machine with flower pollination algorithm for phycocyanin pigment concentration estimation. Environ Sci Pollut Res 2021;28:32564–79.

[52] Saporetti CM, Fonseca DL, Oliveira LC, Pereira E, Goliatt L. Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields. Mar Pet Geol 2022;105783.

[53] Basílio SdCA, Putti FF, Cunha AC, Goliatt L. An evolutionary-assisted machine learning model for global solar radiation prediction in minas gerais region, southeastern Brazil. Earth Sci Inform 2023.

[54] Souza DP, Martinho AD, Rocha CC, Christo EdS, Goliatt L. Group method of data handling to forecast the daily water flow at the cahora bassa dam. Acta Geophys 2022;1–13.

[55] Ikram RMA, Goliatt L, Kisi O, Trajkovic S, Shahid S. Covariance matrix adaptation evolution strategy for improving machine learning approaches in streamflow prediction. Mathematics 2022;10.

[56] Heddam S, Yaseen ZM, Falah MW, Goliatt L, Tan ML, Sa'adi Z, et al. Cyanobacteria blue-green algae prediction enhancement using hybrid machine learning–based gamma test variable selection and empirical wavelet transform. Environ Sci Pollut Res 2022.

[57] Abba SI, Pham QB, Saini G, Linh NTT, Ahmed AN, Mohajane M, et al. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. Environ Sci Pollut Res 2020;27:41524–39.

[58] Boratto TH, Saporetti CM, Basilio SC, Cury AA, Goliatt L. Data-driven cymbal bronze alloy identification via evolutionary machine learning with automatic feature selection. J Intell Manuf 2022;1–17.

[59] Franco VR, Hott MC, Andrade RG, Goliatt L. Hybrid machine learning methods combined with computer vision approaches to estimate biophysical parameters of pastures. Evol Intell 2022;1–14.

[60] Ayoub MA. Development and testing of an artificial neural network model for predicting bottomhole pressure in vertical multiphase flow (Ph.D. thesis), King Fahd University of Petroleum and Minerals; 2005.

[61] Islam MR, Hossain ME. Advances in managed pressure drilling technologies. In: Islam M, Hossain M, editors. Drilling engineering. Sustainable oil and gas development series, Gulf Professional Publishing; 2021, p. 383–453.

[62] Storn R, Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 1997;11:341–59.

[63] Xavier CR, Silva JGR, Duarte GR, Carvalho IA, Vieira VdF, Goliatt L. An island-based hybrid evolutionary algorithm for caloric-restricted diets. Evol Intell 2023;16:553–64. http://dx.doi.org/10.1007/s12065-021-00680-0.

[64] Saporetti CM, da Fonseca LG, Pereira E. A lithology identification approach based on machine learning with evolutionary parameter tuning. IEEE Geosci Remote Sens Lett 2019;16:1819–23.
[65] Wes McKinney. Data structures for statistical computing in Python. In: Stéfan van der Walt, Jarrod M, editors. Proceedings of the 9th python in science conference. 2010, p. 56–61. http://dx.doi.org/10.25080/Majora-92bf1922-00a.
[66] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585:357–62.
[67] Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: 9th python in science conference. 2010.
[68] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[69] Friedman JH. Multivariate adaptive regression splines. Ann Statist 1991;1–67.
[70] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 2020;17:261–72.
[71] Orkiszewski J. Predicting two-phase pressure drops in vertical pipe. J Pet Technol 1967;19:829–38.
[72] Goliatt L, Yaseen ZM. Development of a hybrid computational intelligent model for daily global solar radiation prediction. Expert Syst Appl 2023;212:118295.
[73] Geffray C, Gerschenfeld A, Kudinov P, Mickus I, Jeltsov M, Kööp K, et al. 8 - verification and validation and uncertainty quantification. In: Roelofs F, editor. Thermal hydraulics aspects of liquid metal cooled nuclear reactors. Woodhead Publishing; 2019, p. 383–405.
[74] Sattar AMA, Gharabaghi B. Gene expression models for prediction of longitudinal dispersion coefficient in streams. J Hydrol 2015;524:587–96.
[75] Basilio SdCA, Saporetti CM, Goliatt L. An interdependent evolutionary machine learning model applied to global horizontal irradiance modeling. Neural Comput Appl 2023.
[76] Basilio SA, Goliatt L. Gradient boosting hybridized with exponential natural evolution strategies for estimating the strength of geopolymer self-compacting concrete. Knowl Based Eng Sci 2022;3:1–16.
[77] Boratto T, Cury A, Goliatt L. A fuzzy approach to drum cymbals classification. IEEE Lat Am Trans 2022;20:2172–80.
[78] Boratto THA, Cury AA, Goliatt L. Machine learning-based classification of bronze alloy cymbals from microphone captured data enhanced with feature selection approaches. Expert Syst Appl 2023;215:119378.
[79] Goliatt L, Sulaiman SO, Khedher KM, Farooque AA, Yaseen ZM. Estimation of natural streams longitudinal dispersion coefficient using hybrid evolutionary machine learning model. Eng Appl Comput Fluid Mech 2021;15:1298–320.
[80] Basílio SCA, Saporetti CM, Yaseen ZM, Goliatt L. Global horizontal irradiance modeling from environmental inputs using machine learning with automatic model selection. Environ Dev 2022;44:100766.
[81] Duarte GR, Castro Lemonge ACd, Fonseca LGd, Lima BSLPd. An Island model based on stigmergy to solve optimization problems. Nat Comput 2020.
[82] Cui J, Cheng L. A theoretical study of the occurrence state of shale oil based on the pore sizes of mixed Gaussian distribution. Fuel 2017;206:564–71.
[83] Yaseen ZM. The next generation of soil and water bodies heavy metals prediction and detection: New expert system based edge cloud server and federated learning technology. Environ Pollut 2022;313:120081.
[84] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: A new learning scheme of feedforward neural networks. In: 2004 IEEE international joint conference on neural networks (IEEE Cat. No.04CH37541), vol. 2. 2004, p. 985–90. http://dx.doi.org/10.1109/IJCNN.2004.1380068.
[85] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222.
[86] Ibrahem Ahmed Osman A, Najah Ahmed A, Chow MF, Feng Huang Y, El-Shafie A. Extreme gradient boosting (XGBoost) model to predict the groundwater levels in selangor Malaysia. Ain Shams Eng J 2021;12:1545–56.
[87] Cheng M-Y, Cao M-T. Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. Appl Soft Comput 2014;22:178–88.
[88] Hastie T, Tibshirani R, Friedman J. Elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer Science+Business Media; 2009.