Energy Proceedings
Vol 41, 2024
ISSN 2004-2965

Comparative study of machine learning algorithms for predicting steam-assisted gravity drainage (SAGD) production performance

Bin Wang [1], Shijun Huang [1*], Fenglan Zhao [1], Yue Chen [1], Xinhan Fan [1]

1 College of Petroleum Engineering, China University of Petroleum (Beijing)

(*Corresponding Author: hshj@cup.edu.cn)

ABSTRACT

Steam-assisted gravity drainage (SAGD) is one effective and well-established technology for recovering heavy oil and bitumen resources. Extensive research has been conducted on data-driven models to evaluate the production performance of the SAGD process. The artificial neural network (ANN) is a commonly used machine learning method. However, it is crucial to explore other machine learning methods such as Symbolic Regression (SR), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) using field data. In this study, firstly, a data set consisting of thirteen input/output attributes describing production-related properties and production characteristics was extracted from Long Lake field data. Secondly, three different machine learning methods, including Neural Networks (ANN), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Symbolic Regression (SR), were employed to establish a relationship between the input and output parameters in the different data sets. Subsequently, a range of models were created, evaluated, and compared. Furthermore, the impact of two feature scaling methods, namely standardization and normalization, on the accuracy of a series of prediction models was explored. Lastly, the sensitivity of the input parameters was analyzed. Analysis of the forecasting results obtained from different models leads to the following conclusions. The study found that standardization and normalization significantly enhance the performance of the artificial neural network model, with standardization being more effective. However, the impact of data scaling on integrated learning models (random forest and extreme gradient boosting tree) is minimal. Interestingly, for models based on symbolic regression algorithms, not using data scaling yields the best results. Both artificial neural network and symbolic regression algorithms demonstrate significant advantages and are suitable for predicting SAGD production. However, the symbolic regression model can derive analytical expressions that describe the input-output relationship, which are easy to interpret and apply. This suggests that symbolic regression algorithms may be preferable to artificial neural network algorithms. The top five factors with the greatest impact on cumulative oil production at the end of the main production stage in SAGD, ranked in descending order of importance, are CSI2 (cumulative steam injection volume at the end of the main production stage), LE (effective length of the horizontal well section), MSIP (mean steam injection pressure during the main production stage), HPVH (height of the hydrocarbon pore volume), and H (effective thickness). The methods proposed in this article are helpful for establishing an intelligent energy management system for SAGD (Steam-Assisted Gravity Drainage) development in heavy oil fields. The system can support decision-making processes by providing accurate forecasting and predictive analytics.

Keywords: Steam-Assisted Gravity Drainage (SAGD), Machine learning, Artificial Neural Networks, Symbolic Regression, Random Forest, Extreme Gradient Boosting

NONMENCLATURE

Abbreviations	
SAGD	Steam-Assisted Gravity Drainage
Symbols	
L_{hidden}	a hidden layer
w_{ij}	the weight of the connection between neuron j of the previous L_{hidden} and neuron i of the current L_{hidden}
x_j	the input from the previous L_{hidden}
y_i	the output at neuron i of the current L_{hidden}

This is a paper for 15th International Conference on Applied Energy (ICAE2023), Dec. 3-7, 2023, Doha, Qatar.

f	the activation function

1. INTRODUCTION

With conventional petroleum resources depleting, unconventional oil and gas resources are gaining significance as an alternative. Among these resources, oil sands or bitumen reserves have demonstrated substantial potential within the existing economic and technological framework. Steam Assisted Gravity Drainage (SAGD) is a prominent in-situ bitumen recovery technique that involves the utilization of two parallel horizontal wellbores, typically ranging from 500 m to 1,000 m in length[1]. Detailed analysis using numerical flow simulation models for the SAGD process is a time-consuming and costly endeavor. The quantitative operational conditions and uncertain reservoir properties play a crucial role in determining production and development strategies for oil sands operations. The presence of numerous parameters complicates decision-making and accurate predictions of future production performance in SAGD processes. This complexity further increases when considering reservoir heterogeneities and other enhanced oil recovery techniques. Consequently, engineers require decision-making tools and analyses that can guide them in the presence of vast SAGD information. This has prompted engineers to utilize data-driven models to predict SAGD production performance as an alternative to simulation processes[2]. Data-driven modeling is a modern approach that involves a comprehensive analysis of available data to characterize the system of interest. It entails the construction of models that describe the behavior of physical processes using different machine learning techniques[3]. The relationships within collected data are evaluated in a data-driven model without prior knowledge of the data[4]. One of the significant advantages of data-driven models is that they do not require users to have a thorough understanding of the underlying problem. Instead, they rely heavily on data, rather than domain-specific human expertise[5]. By combining big data and domain knowledge, data-driven models can help predict outcomes in complex systems. Data-driven models have gained popularity and have been more fully developed over the past few years. Previous studies in the petroleum industry have extensively explored the use of data-driven models for various applications. These studies include estimating corrosion rates in pipelines[6], predicting cumulative oil production in unconventional reservoirs, accelerating reservoir simulations[7]. Additionally, in the specific context of SAGD, considerable research has been conducted on data-driven models for predicting production performance[8], history matching[9], clustering[10], and optimizing SAGD processes[11]. These studies have significantly enhanced prediction efficiency and expanded the application range of data-driven models. Previous studies on data-driven models in the context of SAGD have employed various machine learning techniques, including decision trees[12], Artificial Neural Network[3], and K-means clustering[13]. Among these methods, Neural Network (NN) stands out as one of the most commonly used approaches in the literature.

However, in the field of petroleum engineering, particularly in predicting SAGD production performance, previous studies have predominantly relied on synthetic datasets, lacking research that explores the predictive capabilities of different algorithms using real field data. Additionally, there is a dearth of studies evaluating the effectiveness of these algorithms using real field data, as many previous studies have been based on synthetic datasets. This serves as the motivation for this research. The novelty of this study lies in the investigation of the predictive capabilities of three typical machine learning algorithms for SAGD production forecasting, based on real field datasets, and the examination of the impact of data scaling methods on the accuracy of prediction models.

2. DATASETS AND METHODS

2.1 Research Process

The specific process of this research is shown in Figure 1, which is mainly composed of four main stages, including data collection and preprocessing, model establishment and optimization based on different algorithms, model accuracy evaluation, and analysis of model results.

2.2 Datasets sources

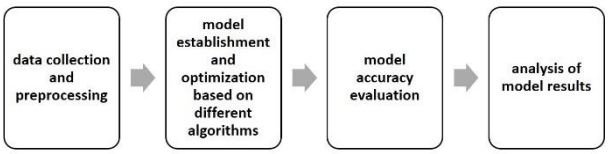

Fig. 1. The specific process of this research

This study collected the cumulative oil production and SAGD production-related parameters of 105 horizontal well pairs in the Long Lake heavy oil field at the end of the primary production phase. The cumulative

oil production at the end of the major production stage (CO) was selected as the prediction target.

Six geological/reservoir parameters that affect production were extracted as input variables, including porosity (POR), water saturation (SW), effective thickness of the reservoir (H), height of the hydrocarbon pore volume (HPVH), net-to-gross ratio (NTG), and heterogeneity index (HI). Water saturation and porosity are averaged values within the total thickness of the producing layer. HPVH is the product of effective thickness, porosity, and oil saturation. NTG represents the ratio of net thickness to total thickness of the producing layer, where the net thickness refers to the oil-saturated sand interval and the total thickness refers to the entire production interval. The presence of shale interlayers typically affects the development of steam chambers, thereby influencing production dynamics. Therefore, a heterogeneity index is formulated to capture the impact of heterogeneity factors such as shale interlayers. The heterogeneity index is a normalized indicator defined as the ratio of non-net thickness to total thickness (1-NTG). A higher HI value indicates thicker shale interlayers within the production interval, which may hinder the development of steam chambers and have a negative impact on development effectiveness.

Six horizontal well/operation parameters that affect production were extracted as inputs for the model, including effective length of the horizontal well (LE), ineffective length ratio along the wellbore (LI), cumulative preheating time (TP), cumulative steam injection during preheating period (CSI1), cumulative steam injection at the end of the major production stage (CSI2), and average steam injection pressure at the major production stage (MSIP). The effective length of the horizontal well refers to the total length of the well segments with clay content less than 30% along the wellbore. The ineffective length ratio along the wellbore is defined as the ratio of the total length of well segments with clay content greater than 30% to the original length of the horizontal well.

The dataset was preprocessed using techniques such as missing value imputation, outlier detection, and filling. Four statistical indicators were calculated for 13 variables based on the processed data, as shown in Table 1.

Table 1 Dataset descriptive statistics (for five randomly selected variables)

Statistical variables	LI (-)	TP (mon)	CSI1 (m^3)	CSI2 (m^3)	MSIP (KPa)	CO (m^3)
Average value	0.18	7.27	5584	208002	1829	49403
Standard deviation	0.11	6.76	25100	215573	160	54093
Minimum value	0.02	0	0	2931	1444	1944
Maximum value	0.54	59	145007	1006943	2119	304877

2.3 Algorithm types

In this section, we will provide a brief introduction to the algorithms involved in the machine learning model developed in this study.

2.3.1 Artificial neural network (ANN)

The artificial neural network (ANN) is a statistical learning model that draws inspiration from the biological nervous system[14]. The typical structure of an ANN model includes an input layer (L_{input}), a hidden layer (L_{hidden}) with multiple interconnected neurons, and an output layer (L_{output})[15].

The mathematical formula for the *i-th* neuron in the ANN model can be expressed as follows:

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

The mathematical expression involves various parameters, including w_{ij}, which represents the weight of the connection between neuron *j* of the previous L_{hidden} and neuron *i* of the current L_{hidden}; b_i, which denotes the deviation at neuron *i* of the current Lhidden; x_j, which represents the input from the previous L_{hidden}; y_i, which denotes the output at neuron *i* of the current L_{hidden}; *f*, which is the activation function; and *k*, which is the number of neurons in the previous L_{hidden}.

If the ANN model does not incorporate any activation function, it reduces to a simple linear model, as shown in the brackets of equation (1). Activation functions such as linear, sigmoid, hyperbolic tangent (tanh), etc., are crucial for enabling ANN models to learn the non-linear complex functional mapping relationship among L_{input}, L_{hidden}, and L_{output}.

2.3.2 Random forest (RF)

The random forest is a classification and prediction algorithm based on decision trees[16]. The model generates multiple independent decision trees, each of which can be expressed by equation (2).

$$h_N(x) = N_{DT} \sum_{i=1}^{N} h_i(x) \quad (2)$$

In the equation, $h_i(x)$ represents a decision tree, and N_{DT} denotes the total number of decision trees.

Within the random forest, the observed data in each branch of the decision tree is divided into left and right paths based on a threshold of model input variables. In the case of a regression tree, the dataset is split using an

error metric minimization approach to obtain the predicted values at the leaf nodes. By aggregating numerous decision trees, the random forest model can approximate highly complex non-linear surfaces, making it a robust tool for addressing intricate non-linear regression and classification problems.

2.3.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm that is widely used for regression and classification tasks. It is an enhanced version of the gradient boosting algorithm that combines the strengths of both gradient boosting and random forest techniques[17]. XGBoost is known for its high accuracy and efficiency in handling large datasets. It uses a gradient boosting framework, where weak prediction models, typically decision trees, are sequentially trained to correct the errors made by the previous models. The algorithm optimizes an objective function by minimizing the loss function and adding regularization terms to prevent overfitting. One key feature of XGBoost is its ability to handle missing values in the dataset. It automatically learns the best direction to take when encountering missing values during training, reducing the need for data preprocessing. XGBoost also provides various hyperparameters that can be tuned to improve model performance, such as the learning rate, maximum tree depth, and number of boosting rounds. Additionally, it supports parallel processing, making it suitable for large-scale applications. Overall, XGBoost is a powerful and versatile algorithm that has gained popularity in various domains, including finance, healthcare, and natural language processing, due to its exceptional performance and flexibility.

2.3.4 Symbolic Regression (SR)

Symbolic Regression is a machine learning technique that aims to discover mathematical expressions that best fit a given dataset. It automatically searches for the optimal mathematical equation that represents the underlying relationship between the input variables and the target variable, without relying on predefined functional forms[18]. The algorithm starts with a population of randomly generated mathematical expressions, which are evaluated and assigned fitness scores based on how well they fit the training data. The fittest individuals are selected for reproduction, and genetic operators such as crossover and mutation are applied to create new offspring. This process continues iteratively, with the population evolving and improving over generations. Symbolic Regression has been successfully applied in various domains, including physics, biology, finance, and engineering, where the underlying relationships are often complex and not easily captured by traditional regression models. The discovered symbolic models can provide valuable insights into the underlying mechanisms and relationships within the data, enabling better understanding, prediction, and decision-making.

3. RESULTS AND DISCUSSION

3.1 Calculation methods for feature scaling and model performance indicators

Feature scaling of the input feature is an essential step to eliminate differences between variables of different orders of magnitude. Standardization scales the data by means and standard deviations, as shown in equation (3). Normalization scales the data by maximum and minimum values, as shown in equation (4).

$$x_{i,1} = \frac{x_i - \mu}{\sigma} \tag{3}$$

$$x_{i,2} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

In the equations above, x_i denotes the original data corresponding to row i of variable x in the dataset, μ denotes the mean value of variable x, σ denotes the standard deviation of variable x, $x_{i,1}$ is the data after standard deviation normalisation of row i of variable x, x_{min} denotes the minimum value of variable x, x_{max} denotes the maximum value of variable x, and $x_{i,2}$ is the normalised data of row i of variable x.

In this paper, the SAGD performance forecasting model uses the root mean square error (E_{rmse}) to train the model, as in equation (5). In addition, the coefficient of determination (R^2) is used to evaluate the model as shown in equation (6).

$$E_{rmse} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2}{\sum_{i=1}^{m}(\bar{y} - y^{(i)})^2} \tag{6}$$

In the equations above, $\hat{y}^{(i)}$ is the predicted value of the model, $y^{(i)}$ is the actual value, m is the number of samples in the dataset that need to be evaluated and measured, and $\bar{y}$ denotes the average value of the actual value $y^{(i)}$.

Table 2 Values of hyperparameters used for ANN model

Number	Symbols	Description
1	Input Layer	The model starts with an input layer that expects input data with 12 dimensions.
2	Dense Layer 1	This layer is a fully connected layer with 30 units/neurons. It uses the ReLU activation function.
3	Dropout Layer 1	This layer applies dropout regularization with a rate of 0.1, meaning 10% of the inputs are randomly set to 0 during training to prevent overfitting.
4	Dense Layer 2	This layer is another fully connected layer with 40 units/neurons. It also uses the ReLU activation function.
5	Dropout Layer 2	This layer applies dropout regularization with a rate of 0.1.
6	Output Layer	This layer is the final layer of the model. It has 1 unit/neuron and uses a linear activation function. This indicates that the model is performing regression.
7	Optimizer	The model uses the Adam optimizer with a learning rate of 0.001 and a decay rate of 0.00001.
8	The loss function used for training	Mean squared error (MSE)
9	The model architecture	The model architecture consists of two fully connected layers with dropout regularization applied between them, and a final output layer for regression.

Table 3 Values of hyperparameters used for RF model

Symbols	Functionality of Hyperparameters	Optimized values
bootstrap	It specifies whether bootstrap samples are used when building trees.	TRUE
max_depth	It sets the maximum depth of each decision tree.	202
max_features	It determines the number of features to consider when looking for the best split.	'auto'
min_samples_leaf	It sets the minimum number of samples required to be at a leaf node.	1
min_samples_split	It sets the minimum number of samples required to split an internal node.	3
n_estimators	It specifies the number of trees to be built in the random forest ensemble.	832

Table 4 Values of hyperparameters used for XGBoost model

Symbols	Functionality of Hyperparameters	Optimized values
colsample_bytree	It specifies the fraction of columns to be randomly sampled for each tree.	0.8
gamma	It controls the minimum loss reduction required to make a further partition on a leaf node of the tree.	0
learning_rate	It determines the step size at each boosting iteration.	0.03
max_depth	It sets the maximum depth of each decision tree.	3
min_child_weight	It defines the minimum sum of instance weight needed in a child.	2
n_estimators	It specifies the number of boosting rounds (trees) to be built.	2000
reg_alpha	It is the L1 regularization term on the weights.	0.001
reg_lambda	It is the L2 regularization term on the weights.	1
subsample	It specifies the fraction of the training instances to be randomly sampled for each tree.	0.9

Table 5 Values of hyperparameters used for SR model

Symbols	Functionality of Hyperparameters	Optimized values
population_size	The number of individuals in each generation of the genetic programming algorithm.	10000
generations	The number of iterations or generations for the genetic programming algorithm.	100
stopping_criteria	The fitness value at which the genetic programming algorithm stops evolving.	0.01
p_crossover	The probability of performing crossover operation during evolution.	0.7
p_subtree_mutation	The probability of performing subtree mutation operation during evolution.	0.1
p_hoist_mutation	The probability of performing hoist mutation operation during evolution.	0.05
p_point_mutation	The probability of performing point mutation operation during evolution.	0.1
max_samples	The proportion of samples used for building each generation of individuals.	0.9
verbose	The level of verbosity for the output.	1
parsimony_coefficient	The coefficient used to balance the complexity and fitness of the model.	0.01
random_state	The seed used for random number generation.	0

3.2 *Impact of two data scaling methods on model performance*

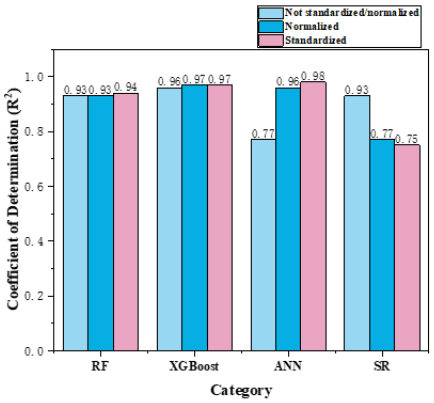

Fig. 2. The Impact of two data scaling methods (indicators: coefficient of determination)

The grid search method was employed to find the optimized hyperparameter values for four algorithm models. The hyperparameter optimization values for the four machine learning algorithms used in this study are presented in Tables 2 to 5, respectively. To quantitatively assess the significance of two data scaling methods, the model's prediction performance is compared under three scenarios: dataset standardization, normalization, and no transformation. It is important to note that the performance prediction models, employing different algorithms, are trained with optimized hyperparameters. Figure 2 and Figure 3 evaluate the model performance using the root mean square error and coefficient of determination metrics, which are calculated on the entire dataset.

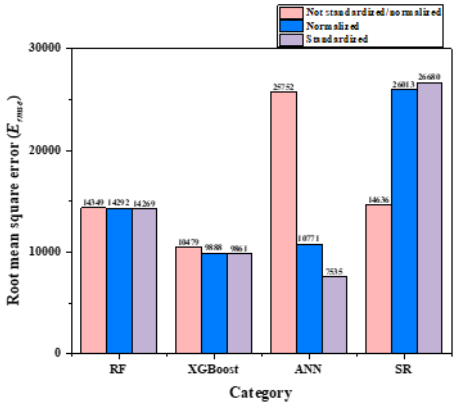

Fig. 3. The Impact of two data scaling methods on Model (indicators: Root mean square error)

The results indicate that both standardization and normalization significantly improve the performance of the artificial neural network model, with standardization being more effective. However, for the integrated learning models examined in this study (random forest and extreme gradient boosting tree), the impact of data scaling methods on performance improvement is minimal. Interestingly, for models based on symbolic regression algorithms, not applying data scaling actually yields the best results.

3.3 *Comparing production performance prediction models based on different machine learning algorithms*

It is worth mentioning that the yield prediction models based on different algorithms are trained with optimized hyperparameters and utilize the best data scaling methods. The results of the root mean square error (E_{rmse}) and the coefficient of determination (R^2) on the training, the test and overall dataset are shown in Table 6.

Table 6 Comparison of results from different production performance prediction models after applying the optimal scaling method and optimizing parameters

Performance forecasting model	Data set	E_{rmse}	R^2
Random forest	Training set	12071	0.96
	Test set	19488	0.74
	Overall	14269	0.94
Extreme Gradient Boosting	Training set	10443	0.97
	Test set	19447	0.74
	Overall	9861	0.97
Artificial neural network	Training set	4802	0.99
	Test set	8279	0.89
	Overall	7535	0.98
Symbolic Regression	Training set	15615	0.93
	Test set	11342	0.91
	Overall	14636	0.93

The different production prediction models all achieved good prediction accuracy in overall dataset (R^2>0.93), which indicates that the production prediction models can fit the non-linear relationship between the SAGD production-related data well. The difference between the training set Ermse and the test set Ermse shows that the random forest model and the extreme gradient boosting tree model have more severe overfitting problems than the neural network model in this paper. The overall prediction results show that both the artificial neural network algorithm and symbolic regression algorithm have significant advantages and are more suitable for the SAGD production prediction problem in this paper.

3.4 *Analysis of factors influencing yield*

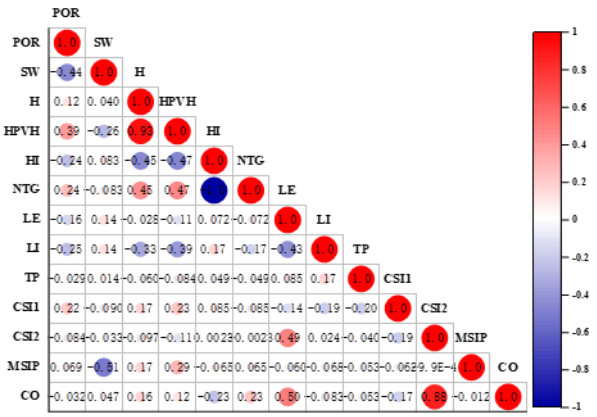

Fig. 4. Impact assessment graph of production-related factors (Correlation plot for the SAGD combined dataset)

Based on the dataset established in the previous section, a correlation plot between 13 variables is generated by using the Pearson correlation coefficients between variables, as shown in Figure 4. The values and colours on the squares in the plot reflect the magnitude of the correlation between individual pairs of variables, with high positive values in the plot indicating a strong positive correlation between two parameters and low negative values indicating a strong negative correlation between two parameters. A strong negative linear correlation is observed between POR (porosity) and SW (water saturation), and a strong positive linear correlation is observed between CO (cumulative oil production at the end of the main production period) and CSI2 (cumulative steam injection at the end of the main production period).

It is worth noting that the Pearson correlation coefficient only reflects the degree of linear negative or positive association between a factor and production performance, and the degree of association is not exactly equivalent to the degree of importance. Therefore, in order to reasonably utilize the machine learning models based on various algorithms established in this paper to explore the importance of various production influencing factors, this section proposes a method for calculating the importance of yield influencing factors based on production performance forecasting models. This method is referred to as Evaluation of Root Mean Square Error Change (EOCRMSE) in this paper. In the EOCRMSE method, the contribution of a feature variable to the model is evaluated by observing the change in the model's root mean square error (RMSE) when the feature variable is removed. If the removal of a feature variable leads to a significant increase in the model's RMSE, it can be considered as an important variable for the model's performance. Conversely, if the removal of a feature variable results in minimal change in the model's RMSE, it can be inferred that the variable has a minor contribution to the model's performance. The EOCRMSE method enables machine learning researchers to identify and select feature variables that have a significant impact on model performance, thereby enhancing the accuracy and interpretability of the model. The results of the importance of input variables obtained for the three models are transformed into percentage form. In addition, the average importance is calculated based on the results of the importance of the input variables for different models, and the final results are shown in Figure 5. As can be seen from the figure, for the dataset used in this paper, the top five factors in descending order of importance based on the average importance of the input variables are CSI2 (cumulative steam injection at the end of the main production period), LE (effective length of the horizontal well section), average steam injection pressure of the main production period (MSIP), hydrocarbon pore volume height (HPVH), and effective thickness (H).

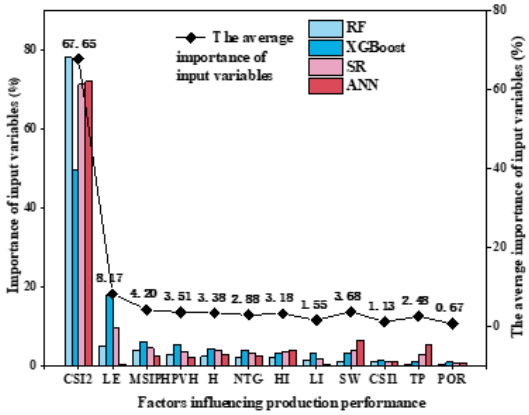

Fig. 5. Impact assessment graph of production-related factors (Importance of production-related factors for different models obtained using the CRMSE method)

4. CONCLUSIONS

Analysis of the forecasting results obtained from different models leads to the following conclusions.

(1) The study found that standardization and normalization significantly enhance the performance of the artificial neural network model, with standardization being more effective. However, the impact of data scaling on integrated learning models (random forest and extreme gradient boosting tree) is minimal. Interestingly, for models based on symbolic regression algorithms, not using data scaling yields the best results.

(2) Both artificial neural network and symbolic regression algorithms demonstrate significant advantages and are suitable for predicting SAGD

production. However, the symbolic regression model can derive analytical expressions that describe the input-output relationship, which are easy to interpret and apply. This suggests that symbolic regression algorithms may be preferable to artificial neural network algorithms.

(3) The top five factors with the greatest impact on cumulative oil production at the end of the main production stage in SAGD, ranked in descending order of importance, are CSI2 (cumulative steam injection volume at the end of the main production stage), LE (effective length of the horizontal well section), MSIP (mean steam injection pressure during the main production stage), HPVH (height of the hydrocarbon pore volume), and H (effective thickness).

ACKNOWLEDGEMENT

This work was supported by the National Science and Technology Major Project (Grant No. U1762102).

DECLARATION OF INTEREST STATEMENT

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. All authors read and approved the final manuscript.

REFERENCE

[1] Cui G, Liu T, Xie J, Rong G, Yang L. A review of SAGD technology development and its possible application potential on thin-layer super-heavy oil reservoirs. Geoscience Frontiers 2022;13:101382.

[2] Akbilgic O, Zhu D, Gates ID, Bergerson JA. Prediction of steam-assisted gravity drainage steam to oil ratio from reservoir characteristics. Energy 2015;93:1663–70.

[3] Ma Z, Leung JY. Integration of data-driven modeling techniques for lean zone and shale barrier characterization in SAGD reservoirs. Journal of Petroleum Science and Engineering 2019;176:716–34.

[4] Kumar A, Hassanzadeh H. Impact of shale barriers on performance of SAGD and ES-SAGD — A review. Fuel 2021;289:119850.

[5] Huang Z, Yang M, Chen Z. Integration of machine learning and data analysis for the SAGD production performance with infill wells. Can J Chem Eng 2023;101:6928–43.

[6] Shaik NB, Pedapati SR, Taqvi SAA, Othman AR, Dzubir FAA. A Feed-Forward Back Propagation Neural Network Approach to Predict the Life Condition of Crude Oil Pipeline. Processes 2020;8:661.

[7] Wang K, Luo J, Wei Y, Wu K, Li J, Chen Z. Practical application of machine learning on fast phase equilibrium calculations in compositional reservoir simulations. Journal of Computational Physics 2020;401:109013.

[8] Le Van S, Chon BH. Evaluating the critical performances of a CO2–Enhanced oil recovery process using artificial neural network models. Journal of Petroleum Science and Engineering 2017;157:207–22.

[9] Ma Z, Leung JY. A knowledge-based heterogeneity characterization framework for 3D steam-assisted gravity drainage reservoirs. Knowledge-Based Systems 2020;192:105327.

[10] Pinto H, Gates I, Wang X. Bayesian Biclustering by dynamics: A clustering algorithm for SAGD time series data. Computers & Geosciences 2019;133:104304.

[11] Mayo-Molina I, Leung JY. Optimization of the Steam Alternating Solvent Process Using Pareto-Based Multi-Objective Evolutionary Algorithms. Journal of Energy Resources Technology 2023;145:033202.

[12] Akbilgic O, Zhu D, Gates ID, Bergerson JA. Prediction of steam-assisted gravity drainage steam to oil ratio from reservoir characteristics. Energy 2015;93:1663–70.

[13] Zheng J, Leung JY, Sawatzky RP, Alvarez JM. A Proxy Model for Predicting SAGD Production From Reservoirs Containing Shale Barriers. Journal of Energy Resources Technology 2018;140:122903.

[14] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

[15] Wang S, Chen Z, Chen S. Applicability of deep neural networks on production forecasting in Bakken shale reservoirs. Journal of Petroleum Science and Engineering 2019;179:112–25.

[16] Breiman L. Random Forests. Machine Learning 2001;45:5–32.

[17] Gu Z, Cao M, Wang C, Yu N, Qing H. Research on Mining Maximum Subsidence Prediction Based on Genetic Algorithm Combined with XGBoost Model. Sustainability 2022;14:10421.

[18] Sipper M, Moore JH. Symbolic-regression boosting. Genet Program Evolvable Mach 2021;22:357–81.