Hindawi
Complexity
Volume 2021, Article ID 6675638, 12 pages
https://doi.org/10.1155/2021/6675638

Research Article

A Novel Stacking Heterogeneous Ensemble Model with Hybrid Wrapper-Based Feature Selection for Reservoir Productivity Predictions

Changlin Zhou,[1] Lang Zhou,[2] Fei Liu,[1] Weihua Chen,[1] Qian Wang,[1] Keliang Liang,[3] Wenqiu Guo,[4] and Liying Zhou[5]

[1]*The Fracturing and Acidizing Research Institute, The Engineering Technology Research Institute, Petro China Southwest Oil & Gasfield Company, Chengdu 610031, China*
[2]*The Engineering Technology Department, Petro China Southwest Oil & Gasfield Company, Chengdu 610051, China*
[3]*The Downhole Operation Company, China National Petroleum Corporation Chuanqing Drilling Engineering Co., Ltd, Chengdu 610051, China*
[4]*Sichuan Wisdom Think Tank Consulting Co., Ltd, Chengdu 610041, China*
[5]*Business School, Sichuan University, Chengdu 610064, China*

Correspondence should be addressed to Liying Zhou; zhouly@scu.edu.cn

Received 11 November 2020; Revised 18 December 2020; Accepted 28 December 2020; Published 8 January 2021

Academic Editor: M. Irfan Uddin

Acid fracturing is the most important stimulation method in the carbonate reservoir. Due to the high cost and high risk of acid fracturing, it is necessary to predict the reservoir productivity before acid fracturing, which can provide support to optimize the parameters of acid fracturing. However, the productivity of a single well is affected by various construction parameters and geological conditions. Overfitting can occur when performing productivity prediction tasks on the high-dimension, small-sized reservoir, and acid fracturing dataset. Therefore, this study developed a stacking heterogeneous ensemble model with a hybrid wrapper-based feature selection strategy to forecast reservoir productivity, resolve the overfitting problem, and improve productivity prediction. Compared to other baseline models, the proposed model was found to have the best predictive performances on the test set and effectively deal with the overfitting. The results proved that the hybrid wrapper-based feature selection strategy introduced in this study reduced data acquisition costs and improved model comprehensibility without reducing model performance.

1. Introduction

With the fast economic development in the past decades, China's oil and gas energy consumption has been consistently increasing. According to the BP Statistical Review of World Energy 2019, China's oil consumption reached 650 million tons, with a year-on-year increase of 4.9%; and China's consumption of natural gas reached 306.7 billion cubic meters, with a year-on-year increase of 9.4% in 2019 [1]. However, China's oil and gas production can not meet domestic demand and relies heavily on imports. In 2019, China's external dependence on oil was as high as 70.8%, and the external dependence on natural gas reached 43% [2]. China has always attached great importance to the energy security risks caused by the increasing external dependence on oil and gas energy. Therefore, with rich oil and gas reserves of carbonate reservoirs in China, China is developing additional oil and gas reservoirs to ensure a stable supply of long-term mechanism for oil and gas strategic reserves [1].

In the development of oil and gas reservoirs, acid fracturing is found to be the most effective method for improving reservoir production [2]. To provide commercial production decision support for the acid fracturing parameters, effective productivity prediction models are needed to accurately assess well productivity before the acid fracturing or in the early production stages. However,

productivity prediction for carbonate reservoirs is still a challenge because of the heterogeneity of the formations and complexity of the hydraulic fractures [3].

To date, several production analysis techniques have been developed to assess and forecast reservoir productivity and the estimated ultimate recovery (EUR). Most reservoir productivity prediction techniques have been numerical simulations [4, 5] or decline curve analyses (DCAs) [6, 7]. Numerical simulations accurately predict well productivity by establishing highly complex multiphase systems; however, any uncertainty in the input data can easily lead to a misinterpretation of the simulation results. Because of the complexities associated with accurately assessing the geological conditions and property data, establishing accurate mathematical and physical models is very difficult as they generally have high computational and modeling burdens [8].

The DCA method was originally proposed by Arps [9] and has since been widely used to predict productivity because of its rapid predictive capabilities. Consequently, many improved versions of the original DCA have been developed [7, 10]. However, these methods require at least one year of productivity data. Therefore, because of the short economic reservoir recovery period and the need to know reservoir productivity before drilling, DCA-based methods are not ideal for oil and gas reservoir productivity predictions.

Due to technological advances, machine learning (ML) approaches such as neural networks (NN) and random forest (RF) have become more popular for prediction activities [11]. To overcome the limitations of numerical simulation and DCA-based methods, Wang et al. [12] introduced ML techniques to forecast reservoir productivity. However, there is still insufficient research on the application of ML methods to reservoir productivity predictions.

Wang et al.'s [12] study employed a deep neural network (DNN) model for which 2919 well logs were collected and analyzed, with stopping and dropout adopted to prevent any overfitting. However, as the number of oil wells is generally much smaller than 2919, there are often limited samples available to train a machine learning model, which again could result in overfitting; therefore, Wang et al.'s [12] strategy could be compromised with smaller samples. Brantson et al. [13] applied a backpropagation artificial neural network, a radial basis function neural network, and a generalized regression neural network to predict reservoir production decline from 220 data points, and to limit overfitting caused by the small-sample size, Pearson's correlation coefficient is employed to select the most suitable input features and remove the irrelevant, redundant, or noisy features. While this filter-based feature selection method limited the overfitting and improved prediction performances, Pearson's correlation coefficient can only identify linear relationships. Other research has also found that filter-based feature selection methods perform more poorly than wrapper-based feature selection methods [14]. Han et al. [15] proposed a hybrid machine learning model which combines an individual cluster analysis (K-means clustering, partitioning around medoids clustering, and hierarchical clustering) and an individual regressor (random forest, Gradient boosting tree, and support vector machine) for productivity forecasting of shale reservoirs and validated the performance of their proposed model on the data set with 129 well logs. For the purpose of avoiding the overfitting problem and reducing computation time, a wrapper-based feature selection process was utilized, which also improves the model's interpretability.

Although overcoming the limitations of numerical simulation and DCA-based methods, previous ML-based reservoir productivity forecasting studies still have some limitations. Firstly, previous relative studies employed individual regressors, such as NN, and homogeneous ensemble regressors, such as random forest (RF), to conduct reservoir productivity forecasting activities [3]. The heterogeneous ensemble regressors have received little attention in reservoirs productivity forecasting. Heterogeneous ensemble regressors combine multiple different regression algorithms, so that they achieve a higher diversity of base learners that can better complement each other [16]. Heterogeneous ensemble regressor has been demonstrated a more effective regressor than both individual regressors and homogeneous ensemble regressors in some tasks [16, 17]. Therefore, a stacking heterogeneous ensemble model is proposed and developed to improve the reservoir productivity forecasting performance in this study.

Secondly, previous reservoir productivity forecasting studies identified the main features based on one criterion/algorithm. However, the inherent shortcomings in the individual variable selection algorithm may lead to mistaken important feature identification results [18]. For example, Han et al. [15] solely employed RF to select the main features, which may neglect the main features which may have a significant linear correlation with the forecast target; and the method utilized by Brantson et al. [13] may neglect the nonlinear main features. Therefore, linear-based wrapper feature selection and non-linear-based wrapper feature selection are utilized jointly to select the (linear and nonlinear) main features in this study.

The rest of the article is organized as follows. Section 2 briefly introduces feature selection approaches and machine learning regression models. Section 3 describes the experimental dataset and every component of the proposed stacking heterogeneous ensemble model in detail. The experimental results and further discussion are reported in Section 4. Finally, Section 5 concludes the study and presents the future research direction.

2. Methodology

2.1. Feature Selection. Feature selection, which involves the selection of a feature subset from the original input data by removing irrelevant, redundant, or noisy features while maintaining model performance, generally requires the application of dimensionality reduction techniques [19]. Feature selection is an important processing stage in prediction models [20] because it eliminates the dimensionality problems in datasets with a large number of features [21],

helps ensure that soft-computing models are well trained, and reduces overfitting risks [15].

There are three main feature selection techniques: filter-based, wrapper-based, and embedded [22, 23]. Filter-based strategies are performed before the regression model training stage and consist of two stages. The first stage identifies the relevance of each original feature and predicts the target based on a certain criterion, with the relevance being evaluated using mathematical expressions such as Pearson's correlation coefficient, mutual information, and relief algorithms. In the second stage, all features that have relatively low relevance to the prediction target are removed. Compared with other feature selection approaches, the filter-based strategies are fast and scalable. However, their drawback is notable. They totally ignore the interaction of the selected feature subset with the regression algorithm's performance. So, they show poor performance in benchmark studies [24].

Wrapper-based strategies, however, are implemented after the regression model training process, with the aim being to select the tailored feature subset most beneficial to the learner's performances. In general, as wrapper-based feature selection strategies are focused on the performance of the final learner, it is better than filter-based feature selection [14]; however, on the down side, the wrapper-based feature selection strategies are time-consuming.

Embedded strategies perform the feature selection and learner training simultaneously. And, the embedded strategies are less complexity than the wrapper-based strategies. However, they can only be applied to specific model classes.

2.2. Individual Regressors

2.2.1. Linear Regression. Linear regression (LR) is a widely utilized prediction method that models the relationships between one or more variables and dependent variables, as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \tag{1}$$

where k is the number of the input features, x_i $(i = 1, \ldots, k)$ are the input features, and the coefficients, β_i $(i = 1, \ldots, k)$, can be fitted using ordinary least squares.

LR can also identify the correlations between x_i $(i = 1, \ldots, k)$ and y; however, as LR models assume that the prediction targets are linear with regard to the attributes, and they are unable to capture the nonlinear relationships between the attributes and target and therefore, even though LR models are highly interpretable, they have been found to suffer from inaccuracy in some complex problems.

2.2.2. Support Vector Regression. Support vector machine, which was proposed by Cortes & Vapnik [25], is a powerful machine learning method for classification and regression tasks, with support vector regression (SVR) being specifically used to perform regression activities. The basic idea behind the SVR is the identification of a suitable hyperplane to make the distances between each data point and the hyperplane as small as possible, that is, to minimize the losses between the prediction value $f(x)$ and the real output y. If the data are nonlinear, a kernel function, such as linear, polynomial, or the radial basis function (RBF), is frequently used used to map the data to a higher dimensional space. For a sample (x, y), traditional regression models usually calculate the loss directly based on the difference between the output $f(x)$ and the real output y, and if $f(x)$ and y are completely equal, the loss is considered to be zero. However, SVR assumes that there is a deviation, ε, between $f(x)$ and y, and only when the absolute value of the difference between $f(x)$ ad y is greater than ε, will there be a loss.

SVR has the flexibility of balancing the trade-off between minimizing the empirical error and the complexity of the resulting fitted function, reducing the risk of overfitting. However, SVR is not interpretable and sensitive to missing values; when there are a large number of samples, the efficiency of the SVR is low [26].

2.2.3. Neural Networks. Neural networks are powerful supervised learning machines that can perform regression and classification tasks. A standard neural network (NN) comprises an input layer, several hidden layers, and an output layer. Figure 1 shows the structure of a single-hidden-layer neural network, where n is the number of model inputs and m represents the number of model outputs. In Figure 1, each node represents a neuron, each of which produces an output input to each neuron in the next layer, with the neurons in the hidden layers and the output layer having activation functions, such as rectified linear function (ReLU) and sigmoid. If a_j $(j = 1, 2, \ldots, k)$ is the neuron input, that is, there are k neurons in the previous layer, o_i is the neuron output, w_{ji} is the weight between the neuron i and input a_j, and b_i and $f(\cdot)$ are the bias for neuron i; the neuron's operation can be expressed using the following formula:

$$o_i = f\left(\sum_{j=1}^{k} w_{ji} a_j + b_i\right). \tag{2}$$

Previous studies have demonstrated that NN can automatically approximate any complex no-linear function forms representing the data characteristics well [27]. Though NN showed their ability to solve many regression problems, their performances are significantly influences by many factors, such as the number of the hidden layers and the learning rate. More importantly, NNs are lack of interpretation [28].

2.3. Homogeneous Ensemble Method.

Homogeneous ensemble regressors pool the predictions of multiple individual regression models. Homogeneous ensemble regressors can be roughly divided into two groups: bagging, such as random forest, and boosting, such as gradient boosting decision tree.

2.3.1. Random Forest. Random forest (RF) is a homogeneous regression/classification ensemble technique that constructs multiple decision trees using different bootstrap data samples [29]. A standard decision tree has several

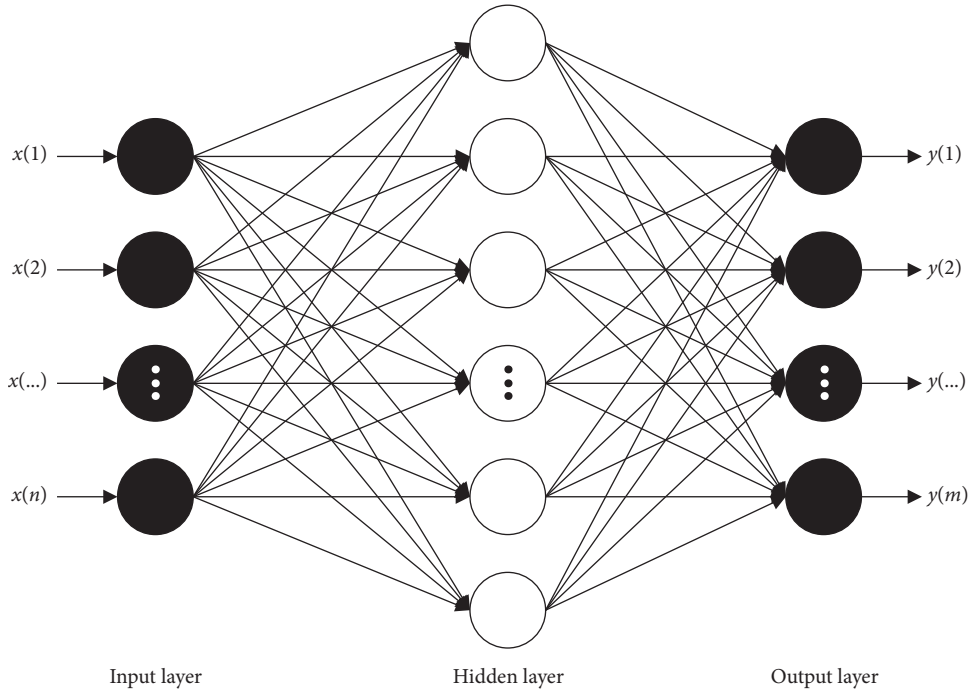

Figure 1: Single-hidden-layer neural network structure.

nodes and directed edges. The training process divides the feature space into several units, each of which has a specific output. When using test data, it is only necessary to classify the data into a unit based on its attributes, with the corresponding output being given by the decision tree. The training process for the regressor decision tree is detailed in the following:

(1) Suppose x and y are, respectively, the input and output variables, $x = \{x_1, x_2, \ldots, x_n\}$, and $y = \{y_1, y_2, \ldots, y_n\}$, where n is the sample size.

(2) A heuristic method is used to partition the feature space, and all the values for all the features in each partition in the data set are inspected one by one. Finally, the best segmentation feature x^j and segmentation point s are determined based on the least square error minimization, $\min_{x^j,s}[\min_{c_1} \sum_{x_i \in} R_1(x^j, s)(y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(x^j,} s)(y_i - c_2)^2]$, where $R_1(j, s) = \{x|x^j t \le ns\}$ and $R_2(j, s) = \{x|x^j t > ns\}$ are determined using feature x^j and its value s; c_1 (c_2) is the output when the specific sample is divided into the region R_1 (R_2).

(3) After determining the optimal segmentation feature x^j and the segmentation point s, the subregions are subsequently segmented until the stop condition is satisfied. When the multiple decision tree predictions have been made, the random forest regressor averages these predictions to determine the final prediction.

RF has good interpretability and can give the feature importance value. However, when the training data set is noisy, there is a risk of overfitting.

2.3.2. Gradient Boosting Decision Tree and Xgboost. The gradient boosting decision tree (GBDT) is also an algorithm that integrates the predictions from multiple decision trees using a boosting approach [30]. Unlike the bagging algorithm, which fits the base learners in parallel, boosting approaches to sequentially build the models; the GBDT model builds additive new base learners to reduce the residual errors in the previous model and keeps adding decision trees until the stop condition is satisfied or the losses reach an acceptable level.

Xgboost is a recently developed advanced GBDT model [31], which because of its powerful regression/classification ability has been employed by many projects in Kaggle machine learning competitions. Similar to the GBDT [30], Xgboost approximates the loss function by employing a Taylor expansion and adds a regularized term to the loss function to smooth the base-learner contributions and limit any overfitting.

GBDT and Xgboost can both calculate the feature importance value and robust to outliers. However, GBDT and Xgboost are not suitable for processing high-dimensional sparse data. In addition, GBDT may suffer from overfitting problem because it subsequently adds new base learners for minimizing the loss function, while Xgboost can avoid overfitting by adding the complexity of the tree model into the regular term [32].

2.4. Heterogeneous Ensemble Method. Heterogeneous regression ensemble methods increase the diversity and complementarity of base predictors by combining a set of different regressors to improve regression performances. Usually, a weighted method and especially the simple average method are utilized to combine the base regressors,

with stacking-based base regressor combination methods having been found to significantly improve heterogeneous regression ensemble model performances [32]. The stacking-based base regressor combination method has two stages: the construction of a set of different base regressors and the feeding of the first stage predictions into a meta-regressor to obtain the final predictions.

3. Experimental Setup

3.1. Data description. The empirical evaluations were based on a real-world dataset coming from an acid fractured vertical well. The sample dataset size was 59 and had 24 features and one target variable, the test production. Table 1 summarizes the parameters and abbreviations for the input and target variables used in this study.

3.2. Workflow for the Proposed Stacking-Based Ensemble Method. Figure 2 shows the flowchart for the proposed prediction model. As can be seen, there are three main prediction model phases: data collection and preprocessing; feature engineering; and prediction model construction.

3.2.1. Phase One: Data collection and Preprocessing. After data collection, the raw dataset is randomly partitioned into two subsets: a training set (80%) and a testing set (20%). Although tree-based classifiers, such as random forest and XGboost, are seldom affected by the feature range, LR, SVMs, and NNs are sensitive to feature scaling; therefore, the data are normalized, as shown in the following equation

$$x_i = \frac{X_i - \min\left(X^{\text{train}}\right)}{\max\left(X^{\text{train}}\right) - \min\left(X^{\text{train}}\right)}, \tag{3}$$

where x_i is the normalized value for X_i, X^{train} is the original training set, and $\max(\cdot)$ and $\min(\cdot)$ calculate the maximum and minimum values for a given dataset.

3.2.2. Phase Two: Feature engineering. In this phase, the initial feature set is retained, and a new feature set, the main feature set, is generated using a wrapper-based feature selection strategy that calculates the relative importance of the features in the model and filters out those that are redundant or irrelevant. As linear models are unable to capture the nonlinear relationships between the features and the targets [33], both linear (LR) and nonlinear models (RF and Xgboost) are used to generate the new feature set.

First, a backward feature elimination process based LR and t test are applied to select the linear features that have significant linear relationships with the prediction target, TP. The backward feature elimination process starts with all the initial features and then discards all features with p values higher than the chosen significance level, which in this study was set to 0.05. Then, RF and Xgboost are employed to select the nonlinear features that have nonlinear relationships with the prediction target. This paper integrates feature ranking into the tree-based model for the selection of the main features using the tree-based regressors, RF and Xgboost, which provide feature-importance scores that measure the average objective reduction after taking the specific features for splitting. A feature with a higher score has a higher importance when building the tree-based model. Following Xia et al. [32], the feature-importance scores generated by the tree-based models were employed to select the main features. However, this study's experimental result shows that wrapper-based feature selection methods with different regressors give totally different feature importance score ranking, and the study of Simsek et al. [18] shows that the nonlinear main features commonly selected by two different wrapper-based feature selection methods outperformed the other nonlinear main feature set while incorporating the smallest number of variables. Consequently, the common features in the Top-K important features identified by two different feature selection methods, the RF and Xgboost in this study, are selected as the nonlinear main features (excluding those linear main features). Finally, the main feature set that contains the linear and nonlinear main features is generated.

Table 1: Parameters and abbreviations for the input and target variables.

Parameter type	Parameter	Abbreviation
Input variables	Effective thickness	ET
	Showing thickness	ST
	Leakage	Leakage
	Total hydrocarbon	TH
	Gamma ray	GR
	Compensated sonic	CS
	Compensated neutron	CN
	Compensated density	CD
	Deep lateral resistivity	DLR
	Short lateral resistivity	SLR
	The difference of depth lateral resistivity	DDLR
	The ratio of depth lateral resistivity	RDLR
	Porosity	Porosity
	Water saturation	WS
	Formation reserve coefficient	FRC
	Gas reserve coefficient	GRC
	Slickwater	SW
	Gelled acid	GA
	The volume of injected liquid	VIL
	Injection rates	IR
	Tubing pressure	TuP
	Stopped pump pressure	SPP
	Pressure decline speed	PDS
	Flow-back rate	FR
Target variables	The test production	TP

3.2.3. Phase Three: Prediction Model Construction. The construction of the proposed forecasting model can be broken down into two subtasks. First, a diverse set of base-regressors is generated, for which LR, SVR, RF, Xgboost, and NN were applied in this study. As all these base-regressors

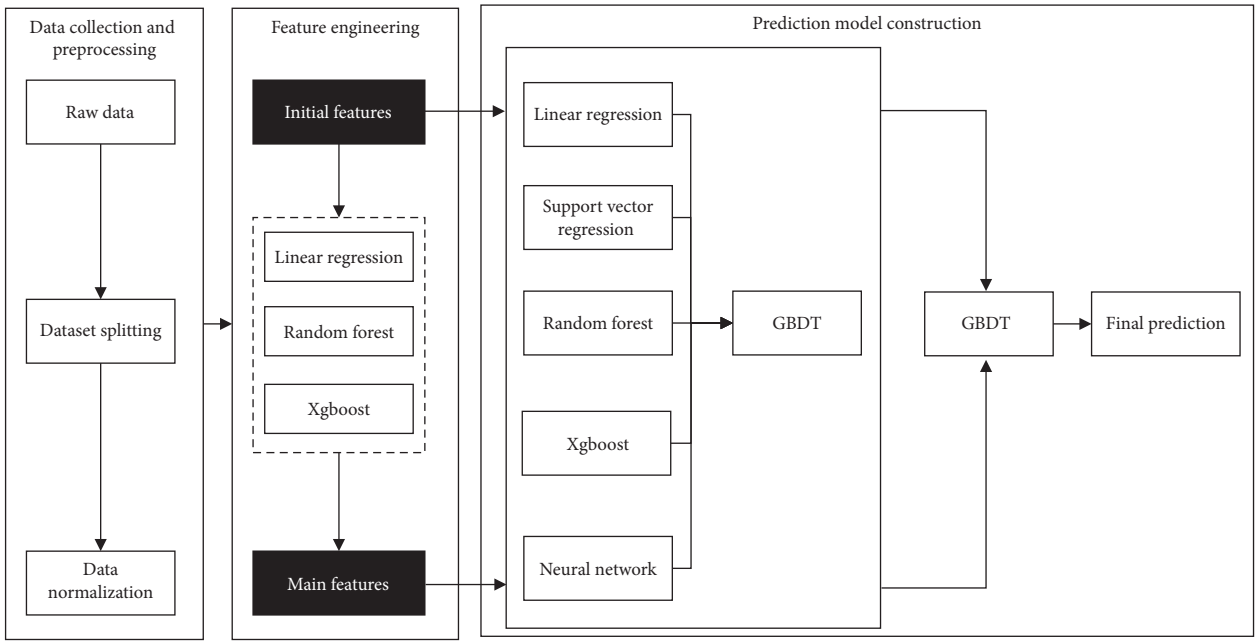

Figure 2: Flowchart for the proposed predicting model.

except LR have a few hyper-parameters that can substantially influence their performances, a 5-fold cross validation grid search was performed on the training set to validate the proposed forecasting model and optimize the base-regressor hyper-parameters. The full parameter grid is presented in Table 2.

When the base-regressors are generated, it is necessary to combine their respective predictions, for which weighted-based and especially average-based methods have been widely used in the previous studies. However, while more complicated combination methods such as stacking with meta-regressors can significantly improve the heterogeneous ensemble model performances [34], the meta-regressor needs to be robust enough to handle high-dimension base predictions and the multicollinearity of the base-regressors [35]. Therefore, GBDT was selected as the meta-regressor to combine the base regressor predictions in this study, that is, the base-regressor predictions were fed to the GBDT to train the meta-regressor. The stacking heterogeneous ensemble algorithmic details are presented in Algorithm 1.

3.3. Benchmarks. To test the superiority of the proposed stacking heterogeneous ensemble prediction model, it was compared with fifteen benchmarks, the details for which are shown in Table 3.

3.4. Performance Evaluation Criteria. In this study, two evaluation criteria, mean-squared error (MSE) and R-squared (R^2) were utilized to subjectively assess the prediction performances of each reservoir productivity prediction model using the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2}, \tag{4}$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{5}$$

where y_n and $\hat{y}_i$ are the observed and predicted values, $\overline{y}$ is the mean for the observed values, and N is the size of the predictions. When the prediction value of a prediction model is more similar to the observation value, the R^2 is larger and the MSE is smaller.

4. Results and Discussion

4.1. Feature Importance Analysis. As the feature selection process can (i) reduce the data acquisition costs, (ii) improve model comprehensibility, and most importantly, (iii) avoid overfitting problems [36–38], the feature importance and main features are analyzed first.

Leakage, DDLR, GRC, VIL, SPP, and FR all have significant linear relationships with the prediction target, with the t statistics and p values for these linear main features given in Table 4. Then, the feature importance scores are obtained by training the RF and Xgboost on the initial feature set, the respective feature importance score results for which are shown in Figures 3 and 4, with the each features shown on the y axis and the importance scores on the x-axis, which are ordered from most to least important.

From these two figures and the obtained linear main features, we can observe that RF-/Xgboost-based feature selection method neglects several variables that have significant (linear) relations with the target. And, it also can be observed that the feature importance score rankings of RF and Xgboost are totally different. These distinctly different

Table 2: Hyperparameter grid.

Model	Parameter	Candidate values
LR	—	—
SVR	C	10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}
	Gamma	0.1, 0.2, 0.3, 0.4
	Kernel	RBF, linear
RF	n_estimator	700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200
	max_features	0.9, 0.8, 0.7, 0.6, 0.5
	max_depth	9, 8, 7, 6, 5, 4, 3
Xgboost	n_estimator	700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200
	max_depth	9, 8, 7, 6, 5, 4, 3
NN	max_iter	700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200
	learning_rate	0.001, 0.005, 0.01, 0.02

(i) **Require:**
(ii) Training dataset1 $D^1 = \{x_i^1, TP_i\}$, $i = 1, \ldots, N$, where x_i^1 are the initial features and N is the size of the training set;
(iii) Training dataset2 $D^2 = \{x_i^2, TP_i\}$, $i = 1, \ldots, N$, where x_i^2 are the main features and N is the size of the training set;
(iv) Base-regressors $R_1, R_2, \ldots, R_n$, where n is the number of base-regressors;
(v) **Ensure:**
(vi) Stacking heterogeneous ensemble regressor R;
(1) **for** $j \in [1, n]$ **do**
(2) Train base-regressor R_j^1 on D^1;
(3) **end for**
(4) **for** each $i \in [1, N]$ **do**
(5) Construct new dataset $D_I = \{x_i^I, TP_i\}$ of predictions, where $x_i^I = \{R_1^1(x_i^1), \ldots, R_n^1(x_i^1)\}$; Train a meta-regressor R^I on D_I.
(6) **end for**
(7) **for each** $j \in [1, n]$ **do**
(8) Train base-regressor R_j^2 on D^2;
(9) **end for**
(10) **for each** $i \in [1, N]$ **do**
(11) Construct new dataset $D_M = \{x_i^M, TP_i\}$ of predictions, where $x_i^M = \{R_1^2(x_i^2), \ldots, R_n^2(x_i^2)\}$; Train a meta-regressor R^M on D_M.
(12) **end for**
(13) **for** each $i \in [1, N]$ **do**
(14) Construct new dataset $D = \{x_i, TP_i\}$ of predictions, where $x_i = \{R^I(x_i^I), R^M(x_i^M)\}$; Train a meta-regressor R on D.
(15) **end for**
(16) **return** R;

Algorithm 1: Proposed reservoir productivity forecast model.

Table 3: Benchmark forecasting models.

Model	Description
IF_LR IF_SVR IF_RF IF_Xgboost IF_NN	Feed the initial features to the base-regressor
MF_LR MF_SVR MF_RF MF_Xgboost MF_NN	Feed the main features to the base-regressor
IF_Ensemble_Average	Employ the average-based method to combine the predictions of the IF_base-regressor models
IF_Ensemble_Stack	Employ a meta-regressor to combine the predictions of the IF_base-regressor models
MF_Ensemble_Average	Employ the average-based method to combine the predictions of the MF_base-regressor models
MF_Ensemble_Stack	Employ a meta-regressor to combine the predictions of the MF_base-regressor models
IF_MF_Ensemble_Average	Employ the average-based method to combine the predictions of the IF_Ensemble_Stack and MF_Ensemble_Stack model

Table 4: Correlation coefficient significance test.

Feature	t statistic	p value
Leakage	3.260	0.002
DDLR	2.254	0.030
GRC	5.542	≤0.001
VIL	2.656	0.011
SPP	−4.333	≤0.001
FR	−2.110	0.041

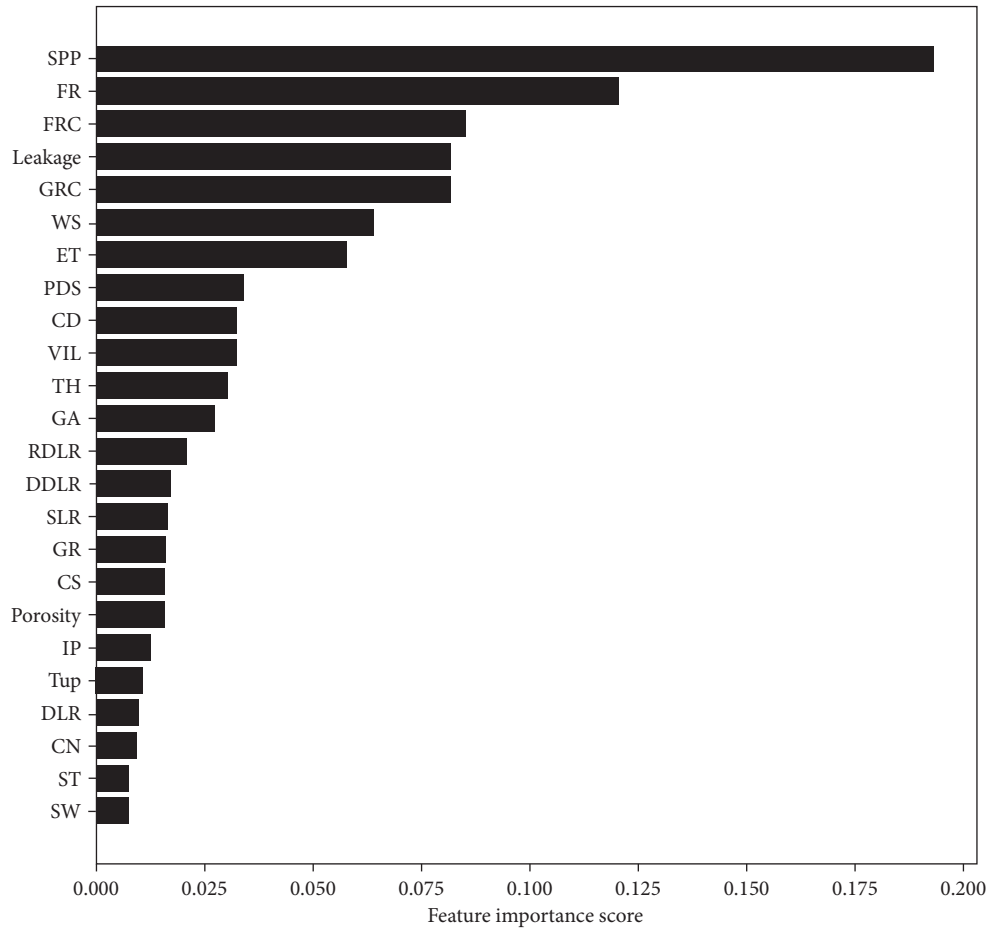

Figure 3: Random forest feature importance scores.

feature importance ranking results showed that these two approaches work differently and have different criteria to score features. To select the features which have significantly nonlinear relations with the predicted target, we adopt a voting strategy in this study. The common top-K important features of RF and Xgboost are considered to have significantly nonlinear relations with the predicted target. It can be seen that there is a large break between the 7^{th}-placed feature and the 8^{th}-placed feature. Therefore, the K in this study is set to 7. The nonlinear main features are selected from Top-7 important features given by RF (SPP, FR, FRC, leakage, WS, ET, and GRC) and Xgboost (SPP, FR, FRC, Leakage, WS, ET, and ST) in this study. Voting by the results given by RF and Xgboost, FRC, WS, and ET are indicated as the nonlinear main features.

Finally, the feature set consists of leakage, DDLR, GRC, VIL, SPP, FR, FRC, WS, and ET called the main feature set. The main feature set contains geologic parameters (the leakage, different depth lateral resistivities, the gas reserve coefficient, the formation reserve coefficient, the water saturation, and the effective thickness) and engineering parameters (the stopped pump pressure, the volume of injected liquid, and the flow-back rate) and are considerably close and found to be in reasonable agreement with the fundamental principle of reservoir characterization practice. For example, the different of depth lateral resistivity is considered to negatively correlate with the reservoir productivity [39]; improving the flow-back rate is an effective measure to improve the productivity of reservoirs [40]. As a consequence, the features selected by the proposed method

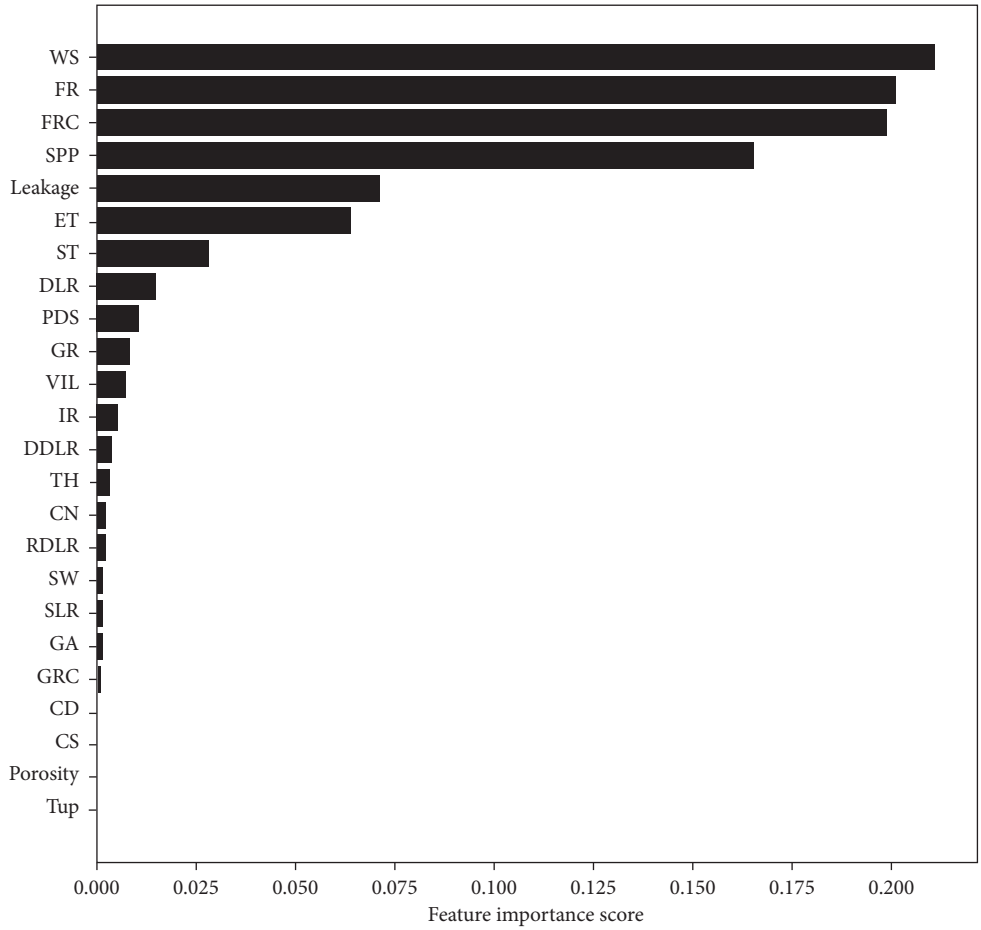

Figure 4: Xgboost feature importance scores.

could be recommended to petroleum engineers as guidelines in their reservoir reconstruction tasks.

4.2. Performance Comparisons of Different Regressors. The performances of the proposed model were compared with benchmarks in terms of R^2 and MSE. The experimental results are shown in Table 5, in which the bold value is the best performance from all prediction models. From Table 5, it can be observed that LR is inferior to other machine learning methods in the training stage. Although individual regressors (SVM and NN) and homogeneous ensemble regressors (RF and Xgboost) achieve good training set performance, their testing set performances are poor. That is, these individual and homogeneous ensemble machine learning regression models suffer overfitting problems to some extent in the reservoir productivity prediction tasks. Specifically, the Xgboost-based regressors (IF_Xgboost and MF_Xgboost) are found to have the best training set performances, with the R^2values of Xgboost-based regressors on the training set both being 0.9999, which indicates an excellent goodness of fit; however, these Xgboost-based regressors have a serious overfitting problem. The R^2values of IF_Xgboost and MF_Xgboost on the out of sample predictions are 0.22 and 0.14, indicating that the Xgboost-based regressor performances on the testing set are the worst.

This serious overfitting problem indicates that individual regressors and homogeneous ensemble regressors are not adequate prediction models for practical applications. The result of the case of the small-size dataset further confirms the need for more data, on one hand, and the incapability of individual regressors and homogeneous ensemble regressors to handle small-size datasets, on the other hand [41]. However, the acquisition of well logs is difficult and time-consuming. The proposed prediction model achieved relatively good performances for the training set and the best testing set performances on the given small-size dataset, with the respective training set R^2 and MSE being 0.9987 and 0.0001, and the respective testing set R^2 and MSE being 0.91 and 0.01, which indicated that the proposed model was a reliable productivity forecasting tool. It shows that our proposed model has the capability to handle the small dataset that individual regressors (SVM and NN) and homogeneous ensemble regressors (RF and Xgboost) could not handle.

4.3. Performance Comparisons of the Prediction Combination Methods. A heterogeneous ensemble method can be divided into two steps: base-estimator generation and predict result fusion. The effectiveness of the heterogeneous ensemble method is highly dependent on result fusion methods [34].

Table 5: Prediction model performances.

Model	Training set		Testing set	
	R^2	MSE	R^2	MSE
IF_LR	0.8305	0.0143	0.57	0.04
IF_SVR	0.8593	0.0119	0.54	0.05
IF_NN	0.9413	0.0050	0.72	0.03
IF_RF	0.9290	0.0060	0.48	0.05
IF_Xgboost	**0.9999**	$\mathbf{5.3402 \times 10^{-7}}$	0.22	0.08
MF_LR	0.7901	0.0178	0.66	0.04
MF_SVR	0.9158	0.0071	0.72	0.03
MF_NN	0.9318	0.0058	0.72	0.03
MF_RF	0.9348	0.0055	0.44	0.06
MF_Xgboost	0.9999	1.1110×10^{-6}	0.14	0.09
IF_Ensemble_Average	0.9454	0.0046	0.62	0.04
IF_Ensemble_Stack	0.8254	0.0148	0.85	0.02
MF_Ensemble_Average	0.9318	0.0058	0.61	0.04
MF_Ensemble_Stack	0.9327	0.0057	0.89	0.01
IF_MF_Ensemble_Average	0.9269	0.0062	0.89	0.01
Our proposed model	0.9987	0.0001	**0.91**	**0.01**

The proposed stacking heterogeneous ensemble model employs the LR, SVR, RF, Xgboost, and NN as the base-regressors, and GBDT is stacked to combine the base-regressor predictions in the proposed model. To validate the stacked-GBDT superiority, it is compared to a widely utilized results fused method, the average method.

From Table 5, it can be seen that the stacked-GBDT models' testing set performances are always better than the simple average models' testing set performances, which is in perfect agreement with the literature. The average testing set R^2 of simple average models is 0.71, while the average testing set R^2 of stacked-GBDT models is 0.88, which is relatively high. Although the training set performances of stacked-GBDT heterogeneous ensemble prediction models are inferior to that of most individual and homogeneous ensemble regressors, the testing set performances of stacked-GBDT heterogeneous ensemble models are significantly superior to that of any other prediction models. It shows that the stacked-GBDT prediction models avoid the overfitting problem generally. Therefore, we can conclude that the stacked-GBDT is proven to be an ideal method for combining the base regressor predictions.

4.4. Performance Comparisons of Different Feature Sets. Previous studies have demonstrated that feature selection approaches could help to overcome the overfitting problem and improve the prediction accuracy [41]. However, even though the input feature dimensions are reduced in the MF_base-regressor models, MF_ base-regressor models, especially the RF- and Xgboost-based regressors, still suffer the overfitting problem. Therefore, we can conclude that dimension reduction methods are not fully effective to resolve the overfitting problems caused by small-sample-sized, high-dimensional datasets [42].

From Table 5, we can observe that the MF_based model (MF_base-regressors models and MF_Ensemble models) performances are better than the corresponding IF_based model performances in general. This result is consistent with that of the previous studies. And, this result indicates that, without losing any statistical performance, the data acquisition costs, model comprehensibility, and computational complexity can be optimized using the main feature selection method proposed in this study.

In addition, when the MF_Ensemble_Stack is utilized to forecast reservoir productivity, the R^2 and MSE based on the output of sample predictions are 0.89 and 0.01, which are very close the best performances obtained by the proposed forecast model. Therefore, if cost minimization is one of the decision objectives, the MF_Ensemble_Stack model can be employed because it requires a fewer number of features, which reduces the data acquisition costs and realizes high prediction accuracy simultaneously.

5. Conclusion and Future Work

The primary objective of this study was to construct an effective reservoir productivity prediction model that could provide a basic acid fracturing design, for which a novel stacking-based heterogeneous ensemble model with hybrid wrapper-based feature selection was proposed.

This study focused on two main aspects. First, to select the significantly linear and nonlinear main features, linear regression, random forest, and Xgboost are jointly utilized in this study. Finally, the effective thickness, difference of depth lateral resistivity, gas reserve coefficient, volume of injected liquid, stopped pump pressure, flow-back rate, formation reserve coefficient, and water saturation are selected as main features. The experimental results showed that the prediction models fed with the main features had similar and even better performances compared to the corresponding prediction models fed with the initial features, which indicated that the data acquisition costs and the model comprehensibility and computational complexities could be optimized using the hybrid wrapper-based main feature selection strategy without losing any statistical performance. More importantly, the MF_Ensemble_Stack model performance, the inputs for which were the main features, was close to the

best prediction performances obtained in this study. Therefore, if minimizing cost is as important as maximizing prediction accuracy, it would be reasonable to perform reservoir productivity predictions only using the main features as the stacking-based heterogeneous ensemble model inputs.

Second, a novel stacking-based heterogeneous ensemble model was adopted to improve the prediction abilities, the superiority of which was demonstrated in the comparison results for the small-sample-sized reservoir dataset. The proposed reservoir productivity prediction model significantly outperformed the benchmark individual and homogeneous ensemble models and effectively resolved the overfitting problems.

While the proposed model was validated on only one set of real-world reservoir data, this research determined its applicability for this specific reservoir. Therefore, further research is needed on data from other reservoirs to train and validate the proposed predictive model. The selection of the ensemble model base classifiers is important [43] as the base-regressors and meta-regressors in the proposed stacking heterogeneous ensemble model are fixed. As increasingly more powerful individual regressors are being proposed, further research is necessary to design a regressor selection strategy to select optimal base regressor/meta-regressor combinations.

Data Availability

The data used to support the findings of this study can be made available from the corresponding author upon reasonable request (zhouly@scu.edu.cn).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Dong, K. Li, B. Li et al., "Study in deep shale gas well to prevent shoulder protruding packer with high pressure sealing," *Engineering Failure Analysis*, vol. 118, Article ID 104871, 2020.

[2] K. Jiawei, J. Yan, Z. Kunpeng, and R. Pengju, "Experimental investigation on the characteristics of acid-etched fractures in acid fracturing by an improved true tri-axial equipment," *Journal of Petroleum Science and Engineering*, vol. 184, Article ID 106471, 2020.

[3] J.-H. Li and L. Ji, "Productivity forecast for multi-stage fracturing in shale gas wells based on a random forest algorithm," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, pp. 1–10, 2020.

[4] K. C. Wilson and L. J. Durlofsky, "Optimization of shale gas field development using direct search techniques and reduced-physics models," *Journal of Petroleum Science and Engineering*, vol. 108, pp. 304–315, 2013.

[5] C. L. Cipolla, E. P. Lolon, J. C. Erdle, and B. Rubin, "Reservoir modeling in shale-gas reservoirs," *SPE Reservoir Evaluation & Engineering*, vol. 13, no. 4, pp. 638–653, 2010.

[6] A. N. Duong, "Rate-decline analysis for fracture-dominated shale reservoirs," *SPE Reservoir Evaluation & Engineering*, vol. 14, no. 3, pp. 377–387, 2011.

[7] D. Han and S. Kwon, "Selection of decline curve analysis method using the cumulative production incline rate for transient production data obtained from a multi-stage hydraulic fractured horizontal well in unconventional gas fields," *International Journal of Oil, Gas and Coal Technology*, vol. 18, no. 3/4, pp. 384–401, 2018.

[8] S. Zendehboudi, N. Rezaei, and A. Lohi, "Applications of hybrid models in chemical, petroleum, and energy systems: a systematic review," *Applied Energy*, vol. 228, pp. 2539–2566, 2018.

[9] J. J. Arps, "Analysis of decline curves," *Transactions of the AIME*, vol. 160, no. 1, pp. 228–247, 1945.

[10] H. N. Mead, "Modifications to decline curve analysis," *Transactions of the AIME*, vol. 207, no. 1, pp. 11–16, 1956.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] S. Wang, Z. Chen, and S. Chen, "Applicability of deep neural networks on production forecasting in bakken shale reservoirs," *Journal of Petroleum Science and Engineering*, vol. 179, pp. 112–125, 2019.

[13] E. T. Brantson, B. Ju, Y. Y. Ziggah et al., "Forecasting of horizontal gas well production decline in unconventional reservoirs using productivity, soft computing and swarm intelligence models," *Natural Resources Research*, vol. 28, no. 3, pp. 717–756, 2019.

[14] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*, Springer-Verlag, Berlin, Germany, 2006.

[15] D. Han, J. Jung, and S. Kwon, "Comparative study on supervised learning models for productivity forecasting of shale reservoirs based on a data-driven approach," *Applied Sciences*, vol. 10, no. 4, p. 1267, 2020.

[16] M. Ala'raj and M. F. Abbod, "A new hybrid ensemble credit scoring model based on classifiers consensus system approach," *Expert Systems with Applications*, vol. 64, pp. 36–55, 2016.

[17] J. Mendes-Moreira, A. M. Jorge, J. Freire de Sousa, and C. Soares, "Improving the accuracy of long-term travel time prediction using heterogeneous ensembles," *Neurocomputing*, vol. 150, pp. 428–439, 2015.

[18] S. Simsek, A. Dag, T. Tiahrt, and A. Oztekin, "A bayesian belief network-based probabilistic mechanism to determine patient no-show risk categories," *Omega*, Article ID 102296, 2020, In press.

[19] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[20] M. I. Miah, S. Zendehboudi, and S. Ahmed, "Log data-driven model and feature ranking for water saturation prediction using machine learning approach," *Journal of Petroleum Science and Engineering*, vol. 194, Article ID 107291, 2020.

[21] G. V. Trunk, "A problem of dimensionality: a simple example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 3, pp. 306-307, 1979.

[22] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning ICML '00*, pp. 359–366, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2000.

[23] V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo, and M. P. Mendes, "Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods," *Science of The Total Environment*, vol. 624, pp. 661–672, 2018.

[24] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes, and R. García-Herrera, "Feature selection in machine learning prediction systems for renewable energy applications," *Renewable and Sustainable Energy Reviews*, vol. 90, pp. 728–741, 2018.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] C. E. D. S. Santos, R. C. Sampaio, L. D. S. Coelho, G. A. Bestard, and C. H. Llanos, "Multi-objective adaptive differential evolution for svm/svr hyperparameters selection," *Pattern Recognition*, vol. 110, Article ID 107649, 2021.

[27] Z. Qu, W. Mao, K. Zhang, W. Zhang, and Z. Li, "Multi-step wind speed forecasting based on a hybrid decomposition technique and an improved back-propagation neural network," *Renewable Energy*, vol. 133, pp. 919–929, 2019.

[28] Z. Alameer, M. A. Elaziz, A. A. Ewees, H. Ye, and Z. Jianhua, "Forecasting gold price fluctuations using improved multilayer perceptron neural network and whale optimization algorithm," *Resources Policy*, vol. 61, pp. 250–260, 2019.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] J. H. Friedman, "Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[31] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16*, pp. 785–794, Association for Computing Machinery, New York, NY, USA, August 2016.

[32] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring," *Expert Systems with Applications*, vol. 78, pp. 225–241, 2017.

[33] Y. Huang, W. Jin, Z. Yu, and B. Li, "Supervised feature selection through deep neural networks with pairwise connected structure," *Knowledge-Based Systems*, vol. 204, Article ID 106202, 2020.

[34] L. Todorovski and S. Džeroski, "Combining classifiers with meta decision trees," *Machine Learning*, vol. 50, no. 3, pp. 223–249, 2003.

[35] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[36] M. Tubishat, N. Idris, L. Shuib, M. A. M. Abushariah, and S. Mirjalili, "Improved salp swarm algorithm based on opposition based learning and novel local search algorithm for feature selection," *Expert Systems with Applications*, vol. 145, Article ID 113122, 2020.

[37] M. Wang and H. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Applied Soft Computing*, vol. 88, Article ID 105946, 2020.

[38] N. Kozodoi, S. Lessmann, K. Papakonstantinou, Y. Gatsoulis, and B. Baesens, "A multi-objective approach for profit-driven feature selection in credit scoring," *Decision Support Systems*, vol. 120, pp. 106–117, 2019.

[39] W. Su, T. Zhang, J. Huo, T. Dong, S. Huang, and Z. Jiang, "Study on deep shallow negative resistivity difference reservoir in no.9 area of karamay oilfield," *West-China Exploration Engineering*, vol. 123, pp. 91-92, 2006.

[40] T. Zhang, "Effects of shut-in timing on flowback rate and productivity of shale gas wells," *Filtration Industry Analyst*, vol. 37, pp. 48–60, 2017.

[41] L. Chen, K. Shao, X. Long, and L. Wang, "Multi-task regression learning for survival analysis via prior information guided transductive matrix completion," *Frontiers of Computer Science*, vol. 14, Article ID 145312, 2020.

[42] M. S. Mahmud, J. Z. Huang, and X. Fu, "Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification," *International Journal of Computational Intelligence and Applications*, vol. 19, no. 1, Article ID 2050002, 2020.

[43] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: adaption of different imbalance ratios," *Expert Systems with Applications*, vol. 98, pp. 105–117, 2018.