Research Article

Artificial Intelligence Method for Shear Wave Travel Time Prediction considering Reservoir Geological Continuity

Shanshan Liu ¹, Yipeng Zhao ², and Zhiming Wang ¹

¹College of Petroleum Engineering, China University of Petroleum, Beijing 102249, China

²CNPC Engineering Technology R&D Company Limited, Beijing 102206, China

Correspondence should be addressed to Zhiming Wang; wellcompletion@126.com

Received 13 January 2021; Revised 28 January 2021; Accepted 13 February 2021; Published 4 March 2021

Academic Editor: Maolin Liao

Copyright © 2021 Shanshan Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The existing artificial intelligence model uses single-point logging data as the eigenvalue to predict shear wave travel times (DTS), which does not consider the longitudinal continuity of logging data along the reservoir and lacks the multiwell data processing method. Low prediction accuracy of shear wave travel time affects the accuracy of elastic parameters and results in inaccurate sand production prediction. This paper establishes the shear wave prediction model based on the standardization, normalization, and depth correction of conventional logging data with five artificial intelligence methods (linear regression, random forest, support vector regression, XGBoost, and ANN). The adjacent data points in depth are used as machine learning eigenvalues to improve the practicability of interwell and the accuracy of single-well prediction. The results show that the model built with XGBoost using five points outperforms other models in predicting. The R^2 of 0.994 and 0.964 are obtained for the training set and testing set, respectively. Every model considering reservoir vertical geological continuity predicts test set DTS with higher accuracy than single-point prediction. The developed model provides a tool to determine geomechanical parameters and give a preliminary suggestion on the possibility of sand production where shear wave travel times are not available. The implementation of the model provides an economic and reliable alternative for the oil and gas industry.

1. Introduction

Sonic logs (compressional and shear wave travel times) are an important tool in determining production and exploration geophysics characterization [1]. Rock elastic parameters are critical to alleviate the risks associated with the drilling and wellbore stability [2]. Consequently, the availability of sonic logs is very essential for the development and production phases of hydrocarbon recovery. In sand production management, according to the acoustic travel time, density, and other logging data to predict the sand production index of a single well [3, 4] and the multiwell data to predict the regional sand production distribution can guide the macro decision-making of sand production and sand control of the whole block in weakly consolidated sandstone reservoir [5].

Shear wave travel times are usually obtained from full-wave train logging data, but due to its high cost, most wells lack shear wave logging data. Because logging data reflects reservoir information, different logging curves are essentially the reflection of the same reservoir in different physical quantities, so there is a mapping relationship between logging data. Using conventional logging data to realize shear wave travel times inversion by multiple regression method [1], the derived physical model often involves a certain degree of simplification, assumption, and subjective experience, which cannot guarantee the quality of the synthetic logging curve.

In recent years, with the application of machine learning in various fields of science and engineering [6–11], many researchers use data-driven methods to solve geological problems. Some intelligent systems [2, 12, 13] have been

used to improve the prediction and accuracy of sonic wave velocity prediction when sonic logs have been lost due to poor storage, poor logging, failure of logging instruments, and bad hole conditions. In 2012, Quantico began to use artificial intelligence to generate acoustic and density logs from existing data flow [14]. Ramcharitar and Hosein established an artificial neural network with 10 hidden layers to estimate shear wave and P-wave by using depth, porosity, clay content, and bulk density [15]. Compared with the empirical model, the neural network model shows a lower absolute average error, so it is more suitable for the estimation of rock mechanical properties. Tariq et al. established an artificial neural network model based on conventional logging data (density, gamma, and neutron porosity) to predict acoustic travel time [2]. Zou predicted the shear wave travel times of shale gas horizontal wells in the Jiaoshiba area by using a random forest algorithm through the comparison of various machine learning methods [16]. Ni et al. based on conventional logging data, such as natural gamma-ray, density, and resistivity, used support vector regression to predict shear wave travel times and achieved good application results [17]. Onalo et al. proposed a three-layer forward multilayer perceptron artificial neural network model. The purpose of this model is to estimate the P-wave and shear wave transit times using real gamma-ray and formation density logging data [3].

The sedimentary environment and the source, origin, and dynamic evolution of the sediments have some continuity in the vertical direction. Most of the machine learning models used for acoustic logging prediction only obtain the input data from the same depth as the output data and construct the point-to-point mapping, which means that the output logging data is only related to the specific input logging data of the same depth and the prediction results are completely independent of each other in space, ignoring the trend information of logging curve changing with depth and the information contained in the previous and subsequent (spatial) correlation of data, and the vertical continuity characteristics of logging data are not fully utilized. In addition, the existing models lack a unified processing method for regional logging data, so it is very important to establish a unified prediction model for improving the comparability of cross-well calculation results, which puts forward higher requirements for data preprocessing.

In this paper, the regional logging data are standardized, normalized, and depth aligned, and the longitudinal continuous sample points are selected as the training characteristic parameters. The data-driven shear wave logging curve prediction models are established by using four intelligent machine learning methods, that is, random forest, support vector regression, XGBoost, ANN, and linear regression based on the largest R^2 of the validation dataset. The model can extract the information contained in the input more effectively according to the spatial correlation, and the output is generated from a series of data inputs. It considers the internal relationship between logging curves and the variation trend of different logging curves with depth, which is more in line with the geological thought and is helpful to improve the practicability of cross-well prediction. Each

scenario is evaluated based on the coefficient of determination, RMSE and MSE (between actual log data and predicted data). The results show that multipoint as a training feature can effectively extract the information contained in logging data, and the prediction accuracy is higher than that of single-point mapping.

These prediction results can be used as a reliable tool for predicting reservoir geomechanical properties (including sand production potential) and provide a reliable method to determine the sand production potential of the formation in real time when limited data or sonic logging data are not available. The significance of the model to the industry is that limited logging data does not need to be transmitted to the field for geoscience analysis to preliminarily determine the possibility of sand production, thus reducing the cost and time of exploration operation, as shown in Figure 1.

2. Methodology to Develop the Artificial Intelligence Model

Artificial intelligence (AI) is a tool to simulate the intelligent behavior of human learning from examples to solve complex problems [18]. Compared with other computational automation methods, AI is a very efficient decision-making system with less time and cost. Artificial intelligence technology uses data to solve the complex problems of classification, diagnosis, prediction, estimation, optimization, selection, and control. Machine learning is the core of artificial intelligence and an important method to realize artificial intelligence. Its statistical and implicit physical driving methods can learn autonomously based on a large number of training data, effectively find the complex mapping relationship between variables, and build the prediction model. This type of solution is called data-driven because it is “learning” or directly driven from data without assuming a predetermined equation as a model. The rapid development of artificial intelligence technology and machine learning method has greatly promoted its application in geophysical logging. Using the machine learning method to analyze and generate logging curves provides a new way to solve the technical problems in the logging data processing. Machine learning is divided into supervised learning and unsupervised learning according to whether the training data is labeled. Supervised learning is divided into a regression algorithm and a classification algorithm. The regression method is a supervised learning algorithm for forecasting and modeling continuous numerical variables, while the classification method is a supervised learning algorithm for modeling or forecasting discrete numerical variables.

In supervised learning, a data-driven model is built by processing a known label dataset, which includes expected input (feature) and output (label/response). A physical driven model is a kind of mathematical mapping based on theory, which connects input and output, while supervised learning identifies patterns in available data sets, learns from observation, and makes a necessary prediction based on statistical mapping of input and output. In the process of establishing the supervised learning model, the predicted

FIGURE 1: Using logging data to update the workflow of sand production risk assessment.

value is compared with the output value, and the model is improved based on the loss function. This process continues until the data-driven model reaches a high level of accuracy and performance, thus minimizing the loss function.

The machine learning method is mainly used in geophysical logging technology to solve the classification and regression problems encountered in the process of logging data processing and interpretation. It is a supervised learning regression problem in machine learning to get the prediction model by training and learning the existing shear wave data and then predict the unknown shear wave curve. By excavating the internal relationship between the shear wave and conventional logging curves, the prediction value with high accuracy can be obtained, which can make up for the lack of shear wave data in the study area and lay the foundation for subsequent sand production prediction. This paper involves the following methods.

Linear regression is the earliest fitting method estimated by the least-squares method which provides a linear relationship between input variables (x) and one output variable (y).

Random forest (RF) is a machine learning method proposed by Breiman [19], which uses multiple decision trees to establish a multidecision tree model with predictive effect. The decision tree deduces simple decision rules from data characteristics to predict the value of target variables. Each tree depends on the value of a random vector, which samples independently and has the same distribution for all

trees in the forest. Random forest algorithm has excellent performance. It can detect the correlation between feature variables and rank the importance of each feature, and the generated model has good interpretability. It has high prediction accuracy, has good tolerance to outliers and noise, and is not easy to overfit.

Support vector machine (SVM) for regression analysis is called support vector regression (SVR). The algorithm has the advantages of strong adaptability, global optimization, rigorous theory, high training efficiency, and good generalization performance. SVR is usually used as an effective machine learning method to predict rock properties [20]. By using kernel function [21] to map data points to inner product space, efficient regression function in high-dimensional space can be obtained.

XGBoost is a supervised learning model which can be used for regression and classification problems. The concept of XGBoost combines gradient descent and tree ensemble learning. Through additive training, one new tree is added at a time to the previous XGBoost model to improve its prediction performance and optimize the value of a pre-specified objective function [22].

The artificial neural network model consists of an input layer, one or more hidden layers, and an output layer with different numbers of neurons. The adjacent layers are fully connected, and each connection is assigned weight. The artificial neural network with a hidden layer can simulate any nonlinear mapping from n -dimension to m -dimension with any precision on the closed set [23].

2.1. Data Preprocessing

2.1.1. Single-Well Logging Data Preprocessing. Data preprocessing is of great significance for machine learning to obtain an accurate prediction model. Well log data serves as the input for the proposed models. Different data formats are generated for logging data due to different data coding format acquisition software adopted by various technical service companies. All the LAS data files are converted into a dataset of one CSV file.

The case study presented in this work is actual well log data of four wells (W1, W2, W3, and W4) from an offshore sandstone oil field in North China. The data contains conventional logging data (including CAL, CNL, AC, GR, PE, RD, RMLL, RS, SP, DEN, DTS, and SP) and orthogonal dipole array acoustic logging data (in this paper, shear wave logging curve DTS is used). Figure 2 shows the original wireline log curve for an interval of 0.1 m with partial characteristics. Firstly, the data is analyzed to identify the suitable depth range that provides an accurate representation of the well. Table 1 shows the measuring depth range of logging curves of each well.

In order to restore the objectivity of the data and get more accurate analysis results, we need to process the original data, deal with the missing values, and remove the abnormal and null data. Caliper logs are used to eliminate borehole irregularities, key seats, and wash-out sections where the tools may have generated false readings. In order to predict sand production, the logging data of the reservoir section is selected according to logging interpretation results. Table 2 lists the statistical summary of the well log data, such as count, mean, standard deviation, min, median, and max after preprocessing.

2.1.2. Standardization of Regional Logging Data. It is necessary to establish a unified data set with multiwell data to establish a unified shear wave travel time prediction model for each block. The establishment of a unified mathematical model for multiwell logging data is of great significance to improve the prediction accuracy of regional shear wave logging. Based on editing, standardization, normalization, and depth drift correction of logging data, the original logging curve is corrected by the standardization method, and the standardized logging data of the whole work area is obtained. Only in this way can the logging interpretation of the whole work area have a unified accuracy.

The key of standardization is to select a reasonable standard layer. The sandstone, mudstone, or other lithologic bodies of the same layer system generally have the same sedimentary environment and approximate parameter distribution characteristics, and their logging responses are consistent. By using this characteristic, the strata with stable regional distribution, similar physical properties, or regular changes and certain thickness are selected as the standard layer. According to the standard layer, the logging curves of each well are translated to realize the standardization of regional logging curves. The common standardization methods include the histogram method and trend surface analysis method. The histogram

method first determines the standard interval of each well, makes the frequency histogram of each well standard interval, analyzes the histograms and logging curves of all wells, determines the most representative frequency distribution, and takes it as the standard value to check and correct the logging curves of other wells, so that the logging characteristics of other wells are consistent with the standard value.

The difference Δx between the average logging value of the standard histogram of the base well and the reference well is taken as the correction value. The mean translation method is used to standardize logging curves.

The log mean value before standardization is x and that after standardization is x_s . Then,

$$x_s = x + \Delta x, \quad (1)$$

where Δx is the difference between the average logging value x_0 of the base well and the average logging value x of the reference well:

$$\Delta x = x_0 - |x|. \quad (2)$$

Different types of logging data have different units. Scaling the data within the same range of values for each input feature can minimize the deviation from one feature to another and speed up the model training time. The training data set is trained after the normalization of mean and variance, and the mean and variance of the training set should be used in the test data set normalization. Because the test data simulate the real environment, the normalization of the data is also a part of the algorithm. The real environment may not get all the tests. Therefore, it is necessary to save the mean and variance of the training set. The data of multiple wells are merged to generate a historical data set. For certain feature data, the following formula can be used for standardization and normalization. For random forest and XGBoost algorithm, there are no feature normalization limits, and SVR, ANN, linear regression algorithms need to normalize the data.

There are many different types of normalization usually used to scale data including Z-score normalization, min-max normalization, sigmoid normalization, statistical column normalization, etc. For this study, to transform and normalize the data, the standardized expression of the mean value is as follows:

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)}. \quad (3)$$

The feature scaling expression is as follows:

$$x_j = \frac{x_j - \mu_j}{S_j}, \quad (4)$$

where μ_j is the normalized parameter, x_j is the actual parameter, and S_j is the standard deviation of the actual parameters.

2.2. Feature Engineering. The purpose of feature engineering is to reduce the calculation cost and improve the upper limit of the model. The best practice is to find which input

FIGURE 2: Continuous profiles of wireline log data of four wells, as input data for shear wave travel times prediction. (a) Well 1, (b) well 2, (c) well 3, and (d) well 4.

TABLE 1: Measuring depth range of logging curve.

Well	Conventional survey section (m)	Mac survey section (m)	Reservoir section (m)
W1	245.7–2587.5	287.9–2280.2	1260–1580
W2	390–3384	376.8–2404.5	1468–3172
W3	453.20–3466.80	472.30–3454.20	2388–3416
W4	459.30–1625.00	456.8952–1626.5652	1400–1592

TABLE 2: Statistical analysis of the field data (wells 1–4).

	MD	CAL	CNL	AC ($\mu\text{s}/\text{ft}$)	GR (GAPI)	PE	RD	RMLL	RS	DEN (g/cm^3)	DTC	DTS	SP
Count	23859	23859	23859	23859	23859	23859	23859	23859	23859	23859	23859	23859	23859
Mean	2237.178	11.262	12.783	97.122	87.311	3.777	5.994	3.625	5.709	2.373	97.09	209.258	32.546
Std	661.702	2.405	16.465	13.903	18.188	1.261	13.023	7.992	12.576	0.147	13.753	51.918	16.232
min	1260	8.193	0.069	57.633	39.585	1.784	0.2	0.218	0.2	1.698	57.633	111.716	-3.776
25%	1574.8	8.816	0.254	87.062	73.424	2.875	1.955	0.725	1.807	2.269	86.976	171.47	19.423
50%	2149	12.07	0.335	96.297	86.858	3.543	2.705	1.198	2.469	2.39	96.196	202.961	30.323
75%	2839.35	12.763	29.06	106.585	101.139	4.561	4.286	2.712	3.933	2.502	106.966	240.347	42.559
Max	3435.8	38.326	63.127	150.556	141.992	17.314	339.161	171.096	295.832	2.722	150.206	435.68	114

parameter has more influence on the output parameter by means of determining the correlation coefficient (CC). The value of CC between two parameters always lies in the range of -1 to 1 . The value of CC close to -1 shows a strong inverse relation between two parameters, the value of CC close to 1 shows a strong direct relation between two parameters, while the value of CC if 0 shows no relation between two parameters. The correlation matrix is shown in Figure 3. According to the correlation matrix, finally, learning variables CAL, CNL, SP, and AC are used.

In addition to correlation analysis, this paper also proposes the viewpoint of multipoint mapping to construct feature engineering. The principle is based on the vertical continuity of strata. The point-to-point feature mapping method does not reflect the actual geological analysis experience and geological point of view very well and is not the optimal method to generate logging curves. The sampling interval of logging is usually small (0.1 m), and the logging data of different depth formations obtained by logging tools affect each other in the longitudinal depth. Therefore, in the logging curve, each data point within the range of mutual influence includes several data points, which means that the prediction of shear wave travel times can be regarded as a sequence data analysis problem with spatial correlation. To make better use of the longitudinal continuity of the logging curve, we take the input vector of the training sample as the logging curve value of three points corresponding to shear wave travel times and its upper and lower points and compare the training effect of the two models with single-point mapping. The principle of feature selection is shown in Figure 4. The input variable is expressed as

$$X_K^N = [x_1^T, x_2^T, \dots, x_{K-1}^T, x_K^T], \quad (X^n \in R_x^K, n \in [1, N]), \quad (5)$$

where R_x^K is the K -dimensional input space, X_K^N are input variables, K is the total number of features, and N is the number of samples.

The output variable is expressed as

$$y^{[N]} = y^T, \quad (y^n \in R_y, n \in [1, N]), \quad (6)$$

where R_y is the output space, and $y^{[N]}$ is the output.

The training sample consists of feature parameters and labels, according to the correlation coefficient matrix, and CAL, CNL, SP, and AC are recorded as n and DTS as the training label. N samples are stored as $n^* (m+1)$ matrix by row, and m is the number of characteristic points. When modeling with a single-point feature, the training sample of the point at a certain depth is the single-point feature and label, as shown in Figure 5. When modeling with a 3-point feature, the construction feature of a sample contains 12 features of 3 adjacent points, namely, 3^*4 features, and the label is the DTS of the depth, as shown in Figure 6. In x_q^p , p stands for points, and q stands for features of a single point. The principle of five points is the same as that of three points.

2.3. Model Development

2.3.1. Data Set Partition. The generalization of a model is an important feature of any model. The developed model must be able to adequately describe the relationship in the training dataset such that the relationship is applicable for a dataset outside the training dataset. If the model successfully describes the relationship in the training dataset but fails to validate and test the model on an external dataset, the model is said to be poor. To avoid this, 70% of the dataset are used for cross-validation purpose and 30% used for testing purpose.

2.3.2. Optimized Hyperparameters for the Models. Model configuration is generally referred to as hyperparameters, such as n -value in random forest algorithm and different kernel functions in SVM. In most cases, the choice of hyperparameters is infinite. The minimum generalization error on the test set is considered as the optimum parameters of the models, as shown in Figure 7. When the model is too complex or too simple, it will be overfitted or underfitted and the generalization error will be very high. The balance point with the optimum parameters of the models in the middle is to be sought by drawing the learning curve of the training set and test. In order to evaluate the performance of the model, the evaluation index R^2 (coefficient of determination), mean squared error (MSE), mean absolute error (MAE), mean squared error (MSE), and Root Mean Square Error (RMSE) are also recorded for both training and testing phases of the models. A highly efficient parallel approach is used for computing the XGBoost by an excellent utilization of the GPU resources offering a significant computational speedup without sacrificing any predictive accuracy. In this paper, five machine learning models are implemented based on third-party Python modules on a server with 64 cores, 256G RAM. Setting hyperparameters of proposed models and optimal hyperparameters solution are listed in Tables 3 and 4.

2.4. Results and Discussion

2.4.1. Comparison of Model Prediction Results. The corresponding model fit is generated with optimal hyperparameters after each of the machine learning models passes through the training data. Table 5 lists the performance comparison of each model during training and testing. The average R^2 for all the models is 0.946 for training and 0.93 for testing, respectively. The difference is small, which indicates that the training process is reliable (i.e., no overfitting). This comparison clearly shows that the predicting model built with XGBoost using five points outperforms other models in prediction. The R^2 of 0.993660634 and 0.964349 are obtained for the training set and testing set respectively.

The analyses between actual and predicted values for testing phases are, respectively, shown in Figure 8.

Figure 9 shows that the proposed XGBoost model using five points gives the highest R^2 and lowest RMSE between actual and predicted data in the training set. During testing of unseen data, XGBoost outperforms other models. It

FIGURE 3: Pearson's correlation heatmap.

FIGURE 4: Construct features considering reservoir geological continuity.

means that for the given set of data, the XGBoost model trained is generalized and can give good results on unseen data.

Computational efficiency is another important aspect to evaluate an algorithm, apart from the prediction accuracy. 16701 points are used in training. Table 6 shows that the running time of linear regression is the shortest, but the calculation accuracy is the lowest. With the development of

graphics processing unit- (GPU-) accelerated computing in recent years, other learning algorithms still cannot compete with XGBoost. For example, the running time of XGBoost keeps stable to generate the prediction result. With the increase of training characteristics, training time increases rapidly in the ANN model. The running time of ANN is almost 36 times slower than that of XGBoost. When the dataset size increases dramatically, the difference in the

FIGURE 5: Single point as a feature.

FIGURE 6: Three points as a feature.

FIGURE 7: Generalization error and model complexity.

TABLE 3: Setting hyperparameters of proposed models.

Algorithm	Hyperparameters				
RF	$n_estimators = 100$	$max_depth = 1$	$min_samples_leaf = 1$	$min_samples_split = 2$	
SVR	kernel = "rbf"	gamma = scale	degreeint = 3	C = 1.0	coef0 = 0
XGBoost	Obj: "reg:linear"	gamma:0	eta:0.3	max_depth:5	subsample: 0.5
ANN-1	activation = "relu"	solver = "Adam"	hidden_layer_sizes = (100)	max_iter = 200	
ANN-3	activation = "relu"	solver = "Adam"	hidden_layer_sizes = (100)	max_iter = 200	
ANN-5	activation = "relu"	solver = "Adam"	hidden_layer_sizes = (100)	max_iter = 200	

TABLE 4: Optimal solutions of hyperparameters.

Algorithm	Hyperparameters				
RF	$n_estimators = 197$	$max_depth = 6$	$min_samples_leaf = 1$	$min_samples_split = 2$	
SVR	kernel = "rbf"	gamma = 1	degreeint = 3	C = 1.0	coef0 = 0
XGBoost	Obj: "reg:linear"	gamma:0	eta:0.3	max_depth:6	subsample:1
ANN-1	activation = "relu"	solver = "lbfgs"	hidden_layer_sizes = (8, 8)	max_iter = 4000	
ANN-3	activation = "relu"	solver = "lbfgs"	hidden_layer_sizes = (14, 14)	max_iter = 4000	
ANN-5	activation = "relu"	solver = "lbfgs"	hidden_layer_sizes = (20, 20)	max_iter = 4000	

TABLE 5: Performance comparison with different models.

Training algorithm	Training set				Testing set			
	R^2	MSE	MAE	RMSE	R^2	MSE	MAE	RMSE
Linear-1	0.890	300.245	11.913	17.328	0.886	299.977	11.952	17.320
Linear-3	0.917	225.643	10.154	15.021	0.902	257.637	10.420	16.051
Linear-5	0.924	206.982	9.754	14.387	0.903	255.466	10.218	15.983
RF-1	0.920	217.056	10.228	14.733	0.913	228.440	10.607	15.114
RF-3	0.929	192.336	9.636	13.869	0.918	214.787	10.165	14.656
RF-5	0.936	173.793	9.156	13.183	0.924	199.431	9.741	14.122
SVR-1	0.931	188.328	9.051	13.723	0.926	193.703	9.324	13.918
SVR-3	0.960	109.210	6.756	10.450	0.950	131.479	7.404	11.466
SVR-5	0.975	67.824	5.392	8.236	0.962	100.597	6.346	10.030
XGB-1	0.981	50.370	4.980	7.097	0.946	142.733	8.004	11.947
XGB-3	0.990	27.193	3.831	5.215	0.953	122.514	7.358	11.069
XGB-5	0.994	17.238	3.062	4.152	0.964	93.628	6.460	9.676
ANN-1	0.925	204.841	9.811	14.312	0.919	213.549	10.164	14.613
ANN-3	0.952	129.916	7.791	11.398	0.943	150.871	8.327	12.283
ANN-5	0.962	103.634	6.938	10.180	0.948	135.592	7.604	11.644

running time may not be neglected. Considering the calculation time and accuracy, five points are selected which meet the prediction requirements.

2.4.2. Influence of Feature Point Density on Model Accuracy.

For the machine learning model, when an example of logging data is input into the model, the final output data is often determined by the characteristics of the training set close to the example. The density of feature points in the training set has a certain impact on the final prediction accuracy. In theory, if a large number of feature points are gathered in a certain area of the input space, the area will be better covered, so the model can better describe the mapping relationship between the input and output space. The logging data features selected in this paper form a four-dimensional input space.

The Gaussian kernel density estimation algorithm [24] is used to calculate the Gaussian kernel density distribution of the feature points of the samples in the four-dimensional input space of the training set and to estimate the feature points density of the training set at the sample points of the test set. The calculation results and absolute value of relative error are projected into the "AC-SP" space and plotted into a two-dimensional scatter diagram, as shown in Figure 10. The "blue-red" scatters are the distribution of 7158 test set feature points and the absolute value of the relative error of DTS, "purple-yellow" circles are the Gaussian kernel density of the current point, and the grey scatter is the distribution of training set samples. The absolute value of the relative error of test set sample points in the high-density area is generally low. Although the absolute value of relative error is not strictly inversely proportional to Gaussian kernel density, there will be a large number of error points in low-density

(a)
FIGURE 8: Continued.

FIGURE 8: Continued.

(c)

FIGURE 8: Continued.

(d)

FIGURE 8: Continued.

FIGURE 8: Comparison of measured versus predicted shear wave travel times (left) and cross-plot (right) of the testing set with different methods. (a) Linear regression with different points, (b) RF with different points, (c) SVR with different points, (d) XGBoost with different points, and (e) ANN with different points.

FIGURE 9: Comparison between the developed models.

regions and only low error points in high-density regions. If the feature points of the test set are distributed in the high-density area of the input space, the prediction accuracy in this area will be higher.

In practice, logging curves of different wells in the same block have high similarity, so the characteristic range of input space is small. At this time, as long as a certain number of samples are collected to make the input spatial feature points reach a certain density, the DTS within this range can be predicted with good accuracy.

3. Model Validation and Sanding Potential Prediction

The developed XGBoost model with a five-point mapping has been validated using field data for another well within the area under study. The dataset used for the validation process has not been used during building the model. The validation data involved a sandstone reservoir section including 1500 data points. Sanding potential can be determined from the shear modulus and bulk modulus estimated from sonic travel time.

Bulk modulus (K) can be estimated from the sonic transit times using

$$K = \rho_r \times \left(\frac{3\Delta t_s^2 - 4\Delta t_c^2}{3\Delta t_s^2 \times \Delta t_c^2} \right) \times 1.34 \times 10^{10}. \quad (7)$$

Shear modulus (G) is estimated from the sonic transit time using

$$G = \frac{\rho_r}{\Delta t_s^2} \times 1.34 \times 10^{10}, \quad (8)$$

where ρ_r is the formation rock bulk density, g/cm^3 ; Δt_s is the shear wave travel times, $\mu\text{s/ft}$; Δt_c is the compressional wave travel time, $\mu\text{s/ft}$.

Sand production index B (MPa), combined modulus E_c (MPa), and Schlumberger ratio R (MPa^2) are calculated using

$$E_c = \frac{9.94 \times 10^5 \times \rho_r}{\Delta t_c^2},$$

$$B = K + \frac{4}{3}G, \quad (9)$$

$$R = K^*G.$$

Discriminant criteria: when $\text{DTC} > 345 \mu\text{s/m}$, $E_c < 1.5 \times 10^4 \text{MPa}$, $B < 1.4 \times 10^4 \text{MPa}$, and $R < 5.9 \times 10^7 \text{MPa}^2$, there is a tendency of sand production in oil and gas reservoirs, otherwise, no sand production. Figure 11 shows the vertical distribution law of sand production index based on the distribution of rock mechanics parameters along depth according to logging data. Δt_s is obtained by machine learning prediction. The results of the four sand production indexes show that the tendency of sand production is serious.

TABLE 6: Training time comparison with different models.

Name	Training time
Linear-1	00:00:028002
Linear-3	00:00:097006
Linear-5	00:00:192011
RF-1	00:00:895051
RF-3	00:01:596092
RF-5	00:02:415138
SVR-1	00:03:421196
SVR-3	00:03:854220
SVR-5	00:05:969342
XGB-1	00:05:160295
XGB-3	00:05:611321
XGB-5	00:06:469370
ANN-1	00:35:640039
ANN-3	03:10:164877
ANN-5	03:03:819514

FIGURE 10: “Density-MSE” relationship in two-dimensional heat scatter map for XGB 3 points.

FIGURE 11: Vertical distribution of sand production prediction indexes.

4. Conclusions

- (1) This paper depicts the potential of machine learning in sand management. The problem of lacking shear wave data in the process of building a geomechanical model is solved. Considering the geological continuity of the reservoir, the data-driven shear wave logging prediction model has been established with five artificial intelligence methods (linear regression, random forest, support vector regression, XGBoost, and ANN) from the perspective of geological analysis. The longitudinal continuous points are extracted as the machine learning characteristic parameters according to the logging curve trend and background information.
- (2) It is important to establish a unified mathematical model of artificial intelligence prediction in the whole area to improve the comparability of calculation results between wells, but at the same time, it puts forward higher requirements for data preprocessing. The selection and supplement, standardization, normalization, and depth alignment of regional logging data solve the matching problem between logging data and measured data from different aspects to ensure the rationality of the model and the accuracy of prediction. These methods are worth popularizing in using machine learning methods to predict well-logging reservoir parameters.
- (3) XGBoost model using five points gives the highest R^2 and lowest RMSE between actual and predicted data in the training set. During testing of unseen data, XGBoost outperforms other models. No matter which machine learning method is used, the accuracy of multipoint prediction is higher than that of single-point prediction. Considering the calculation time and accuracy, five points are selected which meet the prediction requirements. The influence of feature point density on model accuracy is also discussed.

Data Availability

The labeled dataset used to support the findings of this study is available from the Corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the Scientific Research and Technological Development Project of CNPC, Research and Development of Integrated Software of Drilling and Completion Engineering Design and Optimization Decision (smartdrilling) (2020B-4019).

References

- [1] M. P. Tixier, G. W. Loveless, and R. A. Anderson, "Estimation of formation strength from the mechanical-properties log (includes associated paper 6400)," *Journal of Petroleum Technology*, vol. 27, no. 03, pp. 283–293, 1975.
- [2] Z. Tariq, S. M. Elkatatny, M. A. Mahmoud, A. Abdurraheem, A. Z. Abdelwahab, and M. Woldeamanuel, "Estimation of rock mechanical parameters using artificial intelligence tools," in *Proceedings of the 51st U.S. Rock Mechanics/Geomechanics Symposium*, San Francisco, CA, USA, June 2017.
- [3] D. Onalo, S. Adedigba, F. Khan, L. A. James, and S. Butt, "Data driven model for sonic well log prediction," *Journal of Petroleum Science and Engineering*, vol. 170, pp. 1022–1037, 2018.
- [4] C. A. McPhee, C. Webster, G. Daniels et al., "Developing an integrated sand management strategy for kinabalu field, offshore Malaysia," in *Proceedings of the SPE Annual Technical Conference and Exhibition*, Amsterdam, Netherland, October 2014.
- [5] T. Balarabe and S. Isehunwa, "Evaluation of sand production potential using well logs," in *Proceedings of the Annual International Conference and Exhibition*, Jeju Island, Korea, July 2017.
- [6] Y. Xu, H. Zhang, and Z. Guan, "Dynamic characteristics of downhole bit load and analysis of conversion efficiency of drill string vibration energy," *Energies*, vol. 14, no. 1, p. 229, 2021.
- [7] M. Najafzadeh and G. Oliveto, "Riprap incipient motion for overtopping flows with machine learning models," *Journal of Hydroinformatics*, vol. 22, no. 4, p. 749, 2020.
- [8] F. Homaei and M. Najafzadeh, "A reliability-based probabilistic evaluation of the wave-induced scour depth around marine structure piles," *Ocean Engineering*, vol. 196, Article ID 106818, 2020.
- [9] F. Saberi-Movahed, M. Najafzadeh, and A. Mehrpooya, "Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: training group method of data handling using extreme learning machine conceptions," *Water Resources Management*, vol. 34, no. 2, pp. 529–561, 2020.
- [10] M. Najafzadeh and F. Saberi-Movahed, "GMDH-GEP to predict free span expansion rates below pipelines under waves," *Marine Georesources & Geotechnology*, vol. 37, no. 3, pp. 375–392, 2018.
- [11] M. Najafzadeh, F. Saberi-Movahed, S. Sarkamaryan et al., "NF-GMDH-based self-organized systems to predict bridge pier scour depth under debris flow effects," *Marine Georesources & Geotechnology*, vol. 36, no. 5, pp. 589–602, 2017.
- [12] R. M. Alloush, S. M. Elkatatny, M. A. Mahmoud, T. M. Moussa, A. Z. Ali, and A. Abdurraheem, "Estimation of geomechanical failure parameters from well logs using artificial intelligence techniques," in *Proceedings of the SPE Kuwait Oil Gas Show Conference*, Kuwait City, Kuwait, October 2017.
- [13] M. Aleardi, "Seismic velocity estimation from well log data with genetic algorithms in comparison to neural networks and multilinear approaches," *Journal of Applied Geophysics*, vol. 117, pp. 13–22, 2015.
- [14] T. Gan, M. A. Kumar, C. B. Ehiwario et al., "Artificial intelligent logs for formation evaluation using case studies in gulf of Mexico and trinidad & tobago," in *Proceedings of the SPE Annual Technical Conference and Exhibition*, Calgary, Alberta, Canada, September 2019.

- [15] K. Ramcharitar and R. Hosein, "Rock mechanical properties of shallow unconsolidated sandstone formations," in *Proceedings of the SPE Trinidad and Tobago Section Energy Resources Conference*, Port of Spain, Trinidad and Tobago, June 2016.
- [16] X. Zou, "Application of machine learning in shear wave prediction of jiaoshiba shale gas horizontal well," *Jiangnan Petroleum Science and Technology*, vol. 29, no. 04, pp. 16–22, 2019.
- [17] W. Ni, Qi Li, W. Guo et al., "Prediction of shear wave velocity in shale reservoir based on support vector machine," *Journal of Xi'an Shiyou University (Natural Science Edition)*, vol. 32, no. 4, pp. 46–49, 2017.
- [18] A. I. Lawal and S. Kwon, "Application of artificial intelligence to rock mechanics: an overview," *Journal of Rock Mechanics and Geotechnical Engineering*, 2020.
- [19] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [20] Y. Kang and J. Wang, "A support-vector-machine-based method for predicting large-deformation in rock mass," *Institute of Rock and Soil Mechanics Chinese Academy of Sciences*, vol. 3, pp. 1176–1180, 2010.
- [21] I. Steinwart, *Support Vector Machines*. Los Alamos National Laboratory, Information Sciences Group (CCS-3), Springer, Berlin, Germany, 2008.
- [22] 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.
- [23] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," in *Proceedings of the international conference on Neural Networks*, pp. 11–14, IEEE Press, San Diego, CA, USA, January 1987.
- [24] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.