Neural Comput & Applic (2018) 30:1171–1185
https://doi.org/10.1007/s00521-016-2729-2

ORIGINAL ARTICLE

Prediction of water saturation in a tight gas sandstone reservoir by using four intelligent methods: a comparative study

Sadegh Baziar[1] · Habibollah Bavarsad Shahripour[2] · Mehdi Tadayoni[3] · Majid Nabi-Bidhendi[4]

Received: 20 March 2015 / Accepted: 19 November 2016 / Published online: 9 December 2016

Abstract Reservoir water saturation is an important property of tight gas reservoirs. Improper calculation of water saturation leads to remarkable errors in following studies for development and production from reservoir. There are conventional methods to determine water saturation, but these methods suffer from poor generalization and cannot be applicable for various conditions of reservoirs. These methods also depend on core measurements. On the other hand, well log data are usually accessible for all the wells and provide continuous information across the well. Customary techniques are not fully capable to prepare meaningful results for predicting petrophysical properties, especially in presence of small data sets. In this regard, soft computing approaches have been used here. In this research, Support Vector Machine, Multilayer Perceptron Neural Network, Decision Tree Forest and Tree Boost methods have been employed to predict water saturation of Mesaverde tight gas sandstones located in Uinta Basin. Tree Boost and Decision Tree Forest are powerful predictors which have been applied in many research fields. Multilayer Perceptron is the most common neural network, and Support Vector Machine has been used in many petrophysical and reservoir studies. In this research, by using a small data set, the ability of these methods in predicting water saturation has been studied. Based on the data from four wells, two data set patterns were designed to evaluate training and generalization capabilities of methods. In each pattern, different combinations of well data were used. Three error indexes including correlation coefficient, average absolute error and root-mean-square error were used to compare the methods results. Results show that Support Vector Machine models perform better than other models across data sets, but there are some exceptions exhibiting better performance of Multilayer Perceptron Neural Network and Decision Tree Forest models. Correlation coefficient values vary from 0.6 to 0.8 for support vector machine, which exhibits better performance in comparison with other methods.

Keywords Tight gas reservoir · Water saturation · Multilayer perceptron · Decision tree forest · Tree boost · Support vector machine

Abbreviations

AAE	Average absolute error
ANN	Artificial neural network
DT	Sonic travel-time log
ERM	Empirical risk minimization
GR	Gamma ray log
ILD	Deep induction resistivity log
MLP	Multilayer perceptron
MD	Millidarcy
NPHI	Neutron porosity log
OCR	Optical character recognition
r	Correlation coefficient
RBF	Radial basis function
RCAL	Routine core analyses
RHOB	Bulk density log
RMSE	Root-mean-square
SRM	Structural risk minimization

✉ Sadegh Baziar
baziar@aut.ac.ir

[1] Department of Petroleum Engineering, Amirkabir University of Technology, Tehran, Iran

[2] International Petro Offshore Niam Kish, Tehran, Iran

[3] National Iranian Oil Company, Tehran, Iran

[4] Institute of Geophysics, University of Tehran, Tehran, Iran

SV	Support vector
SVM	Support vector machine
SVR	Support vector regression
S_w	Water saturation
VC	Vapnik–Chervonenkis

1 Introduction

Tight gas is the term commonly used to refer to low-permeability reservoirs that produce mainly dry natural gas. Tight gas reservoirs are often defined as formations with permeability <0.1 millidarcy. Many “ultratight” gas reservoirs may have in situ permeability down to 0.001 MD [1]. Tight gas reserves constitute a significant percentage of the natural gas resources worldwide and offer tremendous potential for future reserve growth and production. Economical production from these resources depends on comprehensive understanding from reservoir and determining petrophysical parameters such as water saturation.

Water saturation is one of the most challenging petrophysical properties of a hydrocarbon reservoir that is mainly used to predict the volume of hydrocarbon in place and determine pay zones. Many researchers have probed various methods to measure water saturation [2–4]. This property can be measured directly from routine core analyses (RCAL) or estimated by petrophysical methods. There are some relationships for predicting water saturation in specific formations such as Archie equation, which is formulated for clean sand formations [5, 6]. These models are non-universal and nonlinear empirical relations that need to be fitted to real data. They can be applicable only for specified reservoirs which satisfies the model assumptions. These are the primary reasons for using artificial intelligence techniques like Decision Tree models, artificial neural network (ANN) and support vector machine (SVM) to predict water saturation. Employing these methods reduces the problems associated with costs and generalization of the empirical models of water saturation.

Decision tree forests are an ensemble learning technique used for classification, regression and other tasks, which operate by building a large number of decision trees at training process and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This method has been rarely used in reservoir characterization and petroleum engineering studies. Anifowose et al. [7] employed random forest method for prediction of permeability and porosity in an oil reservoir. They also employed this method for prediction of petroleum reservoir characterization [8].

In the past few years, the Tree Boost method has transpired as a robust method for predictive data mining. It has been extensively used for regression classification tasks, with continuous and categorical predictors. Detailed technical explanation of Tree Boost has been collected in the studies done by Friedman et al. [9]. The authors could not find the application of Tree Boost technique in previous petrophysical and reservoir studies.

An artificial neural network (ANN) is a parallel processor including neurons with ability of performing mathematical calculations through a learning algorithm. The knowledge is encoded in the interconnection weights between input, hidden and output layers [10]. A neural network is made of nonlinear activation functions. Its capability to control nonlinearity is important specially if the underlying distribution responsible for creating the input–output data is intrinsically nonlinear. The network learns by building an input–output mapping for the learning problem. This supervised learning underlies correction of interconnection weights by means of training samples including input signal and a corresponding desired response. The goal is to minimize the difference between the desired response and estimated response by the network in accordance to a proper criterion. The training of the network is repeated until the network reaches a predefined accuracy [10, 11].

Artificial Neural Networks have been extensively used in petroleum engineering studies. ANNs have proved their application in predicting petrophysical and reservoir properties [12–54]. They also have been employed to predict water saturation by some researchers [55–60].

Recently, Support Vector Machines (SVMs) have gained attention in regression and classification tasks due to their excellent generalization performance. The SVM formulation is based on the structural risk minimization (SRM) inductive principle where the empirical risk minimization (ERM) inductive principle and the Vapnik–Chervonenkis (VC) confidence interval are simultaneously minimized [61–63]. There are at least three reasons for the success of SVM: Its ability to learn well with only a very small number of parameters, its robustness against the error of data and its computational efficiency. By minimizing the structural risk, SVM works well not only in classification but also in regression [64, 65].

SVM has gained popularity in petroleum engineering studies and have been used for prediction of reservoir and petrophysical properties [30, 37, 66–76]. There are some studies focusing on prediction of water saturation by using SVM [38, 77].

Using Artificial Intelligence and Learning methods leads to an efficient and universal solution for obtaining reservoir properties for any location on the world. While experimental correlations are applicable to determined reservoir and may have many limitations, these methods are universal and by updating and setting their key parameters can be used in any reservoir.

In this research, four techniques including Decision Tree Forest, Tree Boost, multilayer perceptron neural network (MLP) and support vector machine (SVM) have been employed to predict water saturation of Mesaverde tight gas reservoir located in Uinta Basin, USA. These four methods have previously proved their brilliant performance in many fields of science and technology. Decision Tree Forest and Tree Boost were less employed in petrophysical studies, while Artificial Neural Network and Support Vector Machine have been used in many researches in recent years. The main reason of selecting these methods was to evaluate Decision Tree Forest and Tree Boost methods in predicting water saturation and comparing their performance with those of MLP Neural Network and SVM. Results obtained for different techniques have been compared to each other, and the best predictor for water saturation has been determined.

1.1 Decision Tree Forest

Decision Tree Forest consists of an ensemble of decision trees whose predictions are combined to make the overall prediction for the forest. A decision tree forest is similar to a Tree Boost model in the sense that a large number of trees are grown. However, Tree Boost generates a series of trees with the output of one tree going into the next tree in the series. In contrast, a decision tree forest grows a number of independent trees in parallel, and they do not interact until after all of them have been built. Schematic of a Decision Tree Forest model has been presented in Fig. 1.

Both Tree Boost and decision tree forests produce high-accuracy models. Experiments have shown that Tree Boost works better with some applications and decision tree forests with others, so it is best to try both methods and compare the results. The Decision Tree Forest technique used here is an implementation of the “random forest” algorithm developed by Breiman [78].

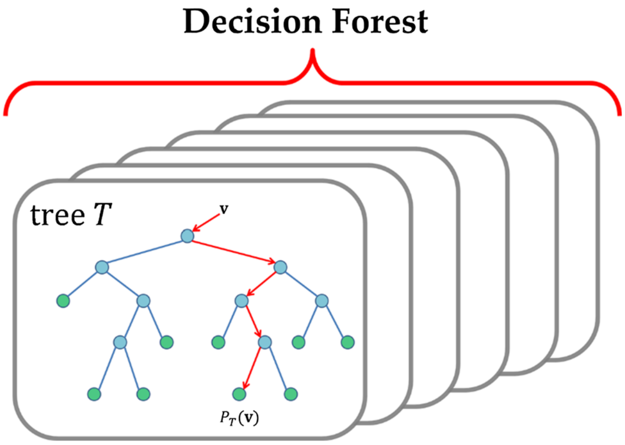

Fig. 1 Schematic of Decision Tree Forest

Decision Tree Forest models are among the most accurate models yet invented. These models can be applied to regression and classification models. Decision tree forest models often have a degree of accuracy that cannot be obtained using a large, single-tree model. This method can handle hundreds or thousands of potential predictor variables.

Decision tree forests use the “out-of-bag” data rows for validation of the model. This provides an independent test without requiring a separate data set or holding back rows from the tree construction. About one-third of data rows are excluded from each tree in the forest, and each tree will have a different set of out-of-bag rows.

In many cases, decision tree forests do not have a problem with over-fitting. Generally, the more the trees in the forest, the better the fit. The randomization element in the decision tree forest algorithm makes it highly resistant to over-fitting.

The primary disadvantage of decision tree forests is that the model is complex and cannot be visualized like a single tree. It is more of a “black box” like a neural network.

1.2 Tree Boost method

Boosting is one of the most popular learning methods which combines many weak learners to create a single-strong learner [79, 80]. Boosting is a technique for improving the accuracy of a predictive function by applying the function repeatedly in a series and combining the output of each function with weighting so that the total error of the prediction is minimized. In many cases, the predictive accuracy of such a series greatly exceeds the accuracy of the base function used alone. The Tree Boost algorithm is functionally similar to decision tree forests because it creates a tree ensemble, but a Tree Boost model consists of a series of trees whereas a decision tree forest consists of a collection of trees grown in parallel. Tree Boost is also known as “Stochastic Gradient Boosting” and “Multiple Additive Regression Trees” (MART).

The Tree Boost algorithm used here was developed by Friedman [81] and is optimized for improving the accuracy of models built on decision trees. Graphically, a Tree Boost model can be represented as demonstrated in Fig. 2.

The first tree is fitted to the data. The residuals (error values) from the first tree are then fed into the second tree which attempts to reduce the error. This process is repeated through a series of successive trees. The final predicted

Tree 1
+
Tree 2
+
Tree 3

Fig. 2 Schematic of a Tree Boost model

value is formed by adding the weighted contribution of each tree.

Usually, the individual trees are fairly small (typically 3 levels deep with 8 terminal nodes), but the full Tree Boost additive series may consist of hundreds of these small trees. Tree Boost models often have a degree of accuracy that cannot be obtained using a large, single-tree model. Tree Boost models can handle hundreds or thousands of potential predictor variables. Irrelevant predictor variables are identified automatically and do not affect the predictive model. Tree Boost uses the Huber M-regression loss function [82] which makes it highly resistant to outliers and misclassified cases. The randomization element in the Tree Boost algorithm makes it highly resistant to over-fitting. Tree Boost can be applied to regression models and k-class classification problems.

The primary disadvantage of Tree Boost is that the model is complex and cannot be visualized like a single tree. It is more of a "black box" like a neural network.

1.3 Multilayer perceptron neural network

Multilayer perceptron (MLP) networks are currently the most widely used neural networks. MLP is a popular estimator to construct nonlinear models of data. It consists of an input layer one or more internal layers of hidden neurons and an output layer. They are also called multilayer feedforward networks (MLFF). The hidden layers are also called internal layers as they receive internal inputs. The network is provided with a training set of patterns having inputs and outputs. The learning algorithm for this type of network is called the back-propagation (BP) algorithm [83, 84]. Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. This is an example of supervised learning and is carried out through back propagation, a generalization of the least mean squares algorithm in the linear perceptron. Figure 3 demonstrates the architecture of an MLP model.

In Fig. 3, $x_p(N)$ is input variable, $W_{hi}(N_h, N)$ is input connection weight, $net_p(N_h)$ is Net input function, $O_p(N_h)$ is Activation function, $W_{oh}(M, N_h)$ is output connection weight and $y_p(N)$ is output variable. MLP network generates nonlinear relationship between inputs and outputs by interconnection of nonlinear neurons. The nonlinearity is distributed throughout the network. It does not require any assumption about the underlying data distribution for designing the networks; hence, the data statistics do not need to be estimated. For an MLP network, the topology is important for the solution of a given problem, i.e., the number of hidden neurons and the size of the training data set. The network has a strong capability for function approximation, learning and generalization.

1.4 Support Vector Regression

Support Vector Machines (SVMs) are learning machines implementing the structural risk minimization inductive principle to obtain good generalization on a limited number of learning patterns [37, 63, 67, 85]. Structural risk minimization (SRM) involves simultaneous attempt to minimize the empirical risk and the VC (Vapnik–Chervonenkis) dimension [85]. The VC dimension of a set of functions is the size of the largest data set due to that the set of functions can scatter. VC theory has been developed over the last three decades by Vapnik and Chervonenkis [61] and Vapnik [62, 63]. This theory characterizes properties of learning machines which enable them to effectively generalize the unseen data [85].

In its present form, the Support Vector Machine was largely developed at AT&T Bell Laboratories by Vapnik et al. [86–91]. The SV (Support Vector) algorithm is a nonlinear generalization of the generalized Portrait algorithm developed in Russia in the sixties [92, 93].

SVM is a learning system that uses a high-dimensional feature space. It yields prediction functions that are extended on a subset of support vectors. SVM can generalize intricate gray level structures with only a very little support vectors. A version of a SVM for regression has been proposed [91], which is called support vector regression (SVR). The model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close (within a threshold ε) to the model prediction [85].

2 Geological background

The data set for this study is obtained from Mesaverde tight gas sandstones located in Uinta Basin in USA. Mesaverde group sandstones represent the principal gas productive sandstone unit in the largest Western US tight gas sandstone basins including Washakie, Uinta, Piceance, northern Greater Green River, Wind River and Powder River.

The Mesaverde group is divided into the regressive deposits of Iles Formation and the overlying massively stacked, lenticular non-marine Williams Fork Formation. The Iles Formation comprises the lower part of the Mesaverde. It contains three marine sandstone intervals, the Corcoran, Cozzette and Rollins. The Williams Fork Formation extends from the top of the Rollins to the top of

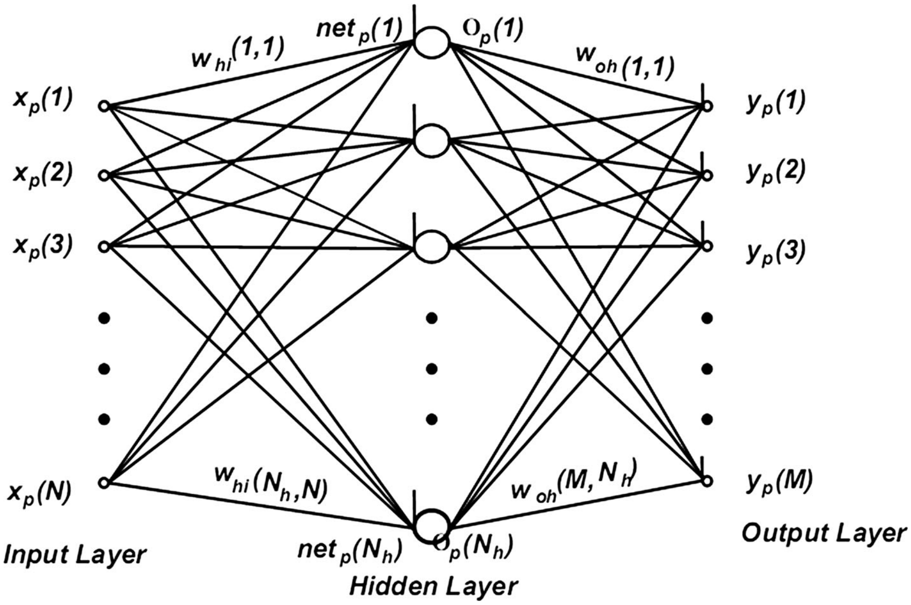

Fig. 3 Multilayer perceptron neural network architecture

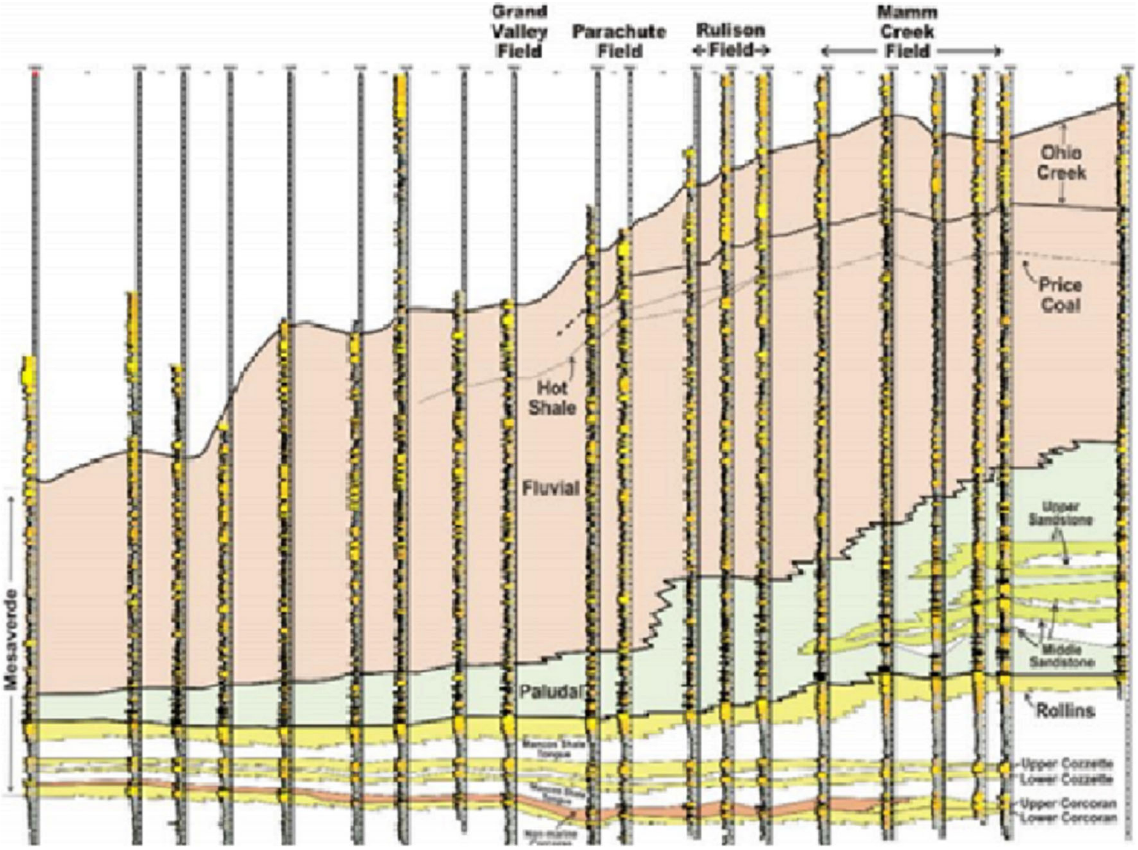

Fig. 4 Cross section showing the stratigraphy of the Mesaverde group [94]

the Mesaverde. The lower part of the Williams Fork contains coals, and is commonly referred to as the Cameo coal interval. Most of the sandstones in the Williams Fork are discontinuous fluvial sands. The stratigraphy of the Mesaverde group is shown in Fig. 4 [94].

3 Data set

To measure the accuracy of the models, log and core information from four wells were used. These wells are cited in Uinta Basin. Database here is small in number of

Table 1 Well order for evaluating training capability

Data set number	Training wells	Testing wells
1	1	1
2	2	2
3	1.2	1.2
4	1.2	1
5	1.2	2

Table 2 Well order for evaluating generalization capability

Data set number	Training wells	Testing well
1	1	2
2	2	1
3	1	3
4	2	3
5	1.2	3
6	1	4
7	2	4
8	1.2	4

data points. Well 1 has a total of 190 data points, well 2 has 107 data points, well 3 has 67 data points and well 4 has 41 data points. It is worth mentioning that prediction methods generally perform better in presence of large number of training data points, but data points here are small in number and it may be a challenge for methods to present their ability to predict water saturation in presence of few number of data points. Data source is clear and has been calibrated before using in this research.

In this research, training and generalization capabilities of methods have been evaluated. To assess training capability of methods, the models were tested by employed data in training procedure. Table 1 represents the pattern of using data of various wells for evaluating training ability.

Furthermore, models trained with training wells data were tested by other wells data to evaluate their capability in generalizing the relationships between various parameters of training data set into new data set in testing procedure. Table 2 describes the set of well data in evaluation of generalization capability of methods.

There was not any geological preference for selecting training wells. Wells number 1 and number 2 were selected as training well, because they had enough number of data points for training procedure, while wells number 3 and number 4 because of small number of data points cannot be selected as training wells.

In training process, regression methods employed in this research got several parameters as input data and a scalar variable as output and then they tried to establish relations between input parameters in order to estimate values as close as possible to output values. In the testing process, the relations obtained in training were used to predict output variable.

In this research, input data consists of log data including gamma ray log (GR), a neutron porosity log (NPHI), a deep induction resistivity log (ILD), a bulk density log (RHOB) and a sonic travel-time log (DT) as input vectors. The scalar output is core-based water saturation. Table 3 represents each parameter range of values. Moreover, in Fig. 5 scatter plots of water saturation versus each well log values are demonstrated. Measurement scales of log data are cited in Table 3. It is not necessary that input measurement scale be within the range of output values, because output is water saturation which is a scalar parameter and have no scale.

4 Methods

Regression models (Decision Tree Forest, Tree Boost, MLP and SVM) were constructed by using DTREG intelligent software. The number of trees for constructing Decision Tree Forest models was 400. Generally, the larger a decision tree forest is, the more accurate the prediction. Maximum tree levels of the Decision Tree Forest were chosen 100, which specify the maximum number of levels (depth) that each tree in the forest may be grown to. When a tree is constructed in a decision tree forest, a random subset of the predictor variables are selected as candidate splitters for each node. Two predictors were chosen as candidates for each node split. The regression methods were verified and validated in a manner that models trained with a well data were tested by using data from other well(s);then, predicted and actual values of saturation were compared by using error indexes.

For constructing Tree Boost models, 400 trees were generated in Tree Boost series. Each tree in the Tree Boost series had 10 levels of splits. The Tree Boost algorithm uses Huber M-regression loss function to evaluate error measurements for regression models [82]. This loss function is a hybrid of ordinary least squares (OLS) and least absolute deviation (LAD). For residuals less than a cutoff point, the squared error values are used. For residuals greater than the cutoff point, absolute values are used. Huber cutoff point was chosen 0.1. A tenfold cross-validation resampling technique was used to strike the right trade-off between over-fitting and under-fitting.

MLP neural network models constructed here had 4 layers (one input, two hidden and one output). An algorithm was used to automatically determine the number of neurons in hidden layers. This algorithm tries building multiple networks with different numbers of neurons in hidden layers and evaluates how well they fit by using cross-validation. Twelve neurons were selected for hidden layer 1, and 4 neurons were selected for hidden layer 2. A

Table 3 Range of parameters values

	Well no. 1	Well no. 2	Well no. 3	Well no. 4
Neutron porosity	0.026–0.410	0.033–0.617	0.043–0.783	0.073–0.189
Sonic travel-time (US/F)	50.373–146.004	64.757–99.236	61.538–83.745	58.750–75.125
Bulk density (g/cm^3)	1.907–2.650	1.236–2.762	1.849–2.622	2.284–2.725
Gamma ray (GAPI)	26.770–128.273	25.075–119.483	32.491–93.741	38.342–84.753
Deep induction resistivity log (OHMM)	4.106–81.224	12.592–139.640	6.486–16.787	6.281–27.922
Depth (m)	2279–3658	3327–3758	2953–3205	3103–3184
Water saturation	0.270–0.942	0.107–0.933	0.217–0.876	0.255–0.899

tenfold cross-validation method was used for validation. Sigmoid function was selected as activation function of hidden layers an output layer. The conjugate gradient method was used to find optimal network weights.

For SVM models, correct selection of the kernel function is so important. RBF, sigmoid and linear kernels are common kernel functions which have been employed in many researches previously [37, 66, 73]. In this research, at the first step, a collection of 320 data points from all 4 wells were used to train the SVM models built by sigmoid and RBF, and later, the models were tested against 85 data points. The best results were obtained by RBF kernel function. So, for constructing SVM models, the RBF kernel function was used. Tenfold cross-validation method was used for validation. The accuracy of an SVM model definitely depends on correct choice of the parameters C, ε and the kernel parameters. The problem of optimal parameter selection is more complex by the principle that a SVM model complexity depends on all mentioned parameters. In procedure of constructing a SVM model, the user has to choose the proper kernel function, and for the selected kernel, how to adjust the parameters. In order to find optimal parameter values, two methods including grid search and pattern search were used. A grid search tries values of each parameter across the specified search range using geometric steps. A pattern search starts at the center of the search range and makes trial steps in each direction for each parameter. If the fit of the model improves, the search center moves to the new point and the process is repeated. If no improvement is found, the step size is reduced and the search is tried again. The pattern search stops when the search step size is reduced to a specified tolerance. When using both grid search and pattern search, the grid search is performed first. Once the grid search finishes, a pattern search is performed over a narrow search range surrounding the best point found by the grid search. Hopefully, the grid search will find a region near the global optimum point and the pattern search will then find the global optimum by starting in the right region.

Different error statistics including correlation coefficient (r), root-mean-square error (RMSE) and average absolute error (AAE) were used to evaluate the accuracy of different models. The mathematical expressions of these error measures are demonstrated in Table 4.

5 Results and discussion

The error analysis for selection of SVM kernel is represented in Table 5. As it can be inferred from the Table 5, RBF kernel function relatively has the lowest values of error among various kernel functions, and so, it has been selected to be the kernel function of SVM models in this study.

5.1 Training capability of models

In this section, capability of methods in training has been evaluated and it has been argued that which method has been well trained. In this regard, models based on Table 1 were tested by data which has been used in training process of them. Correlation coefficients of 4 employed methods (Decision Tree Forest, Tree Boost, MLP and SVM) among 5 regarded data sets in Table 1 have been compared in Fig. 6.

As it can be seen in Fig. 6, it is obvious that all employed methods have been well trained. SVM has gained best results in predicting water saturation, although MLP performs better in data set number 4. MLP has better training ability rather than Tree Boost and Decision Tree Forest. Furthermore, it can be mentioned that Decision Tree Forest has better training sufficiency than Tree Boost technique. Moreover, Training capability of these methods has been assessed by using AAE and RMSE error measurements. Values of mentioned errors are presented in Table 6. Error measures in Table 6 verify the comparison depicted in Fig. 6 and show that SVM and MLP have better training capability.

5.2 Generalization capability of models

In this section, capability of methods in generalization has been assessed and it has been investigated that which

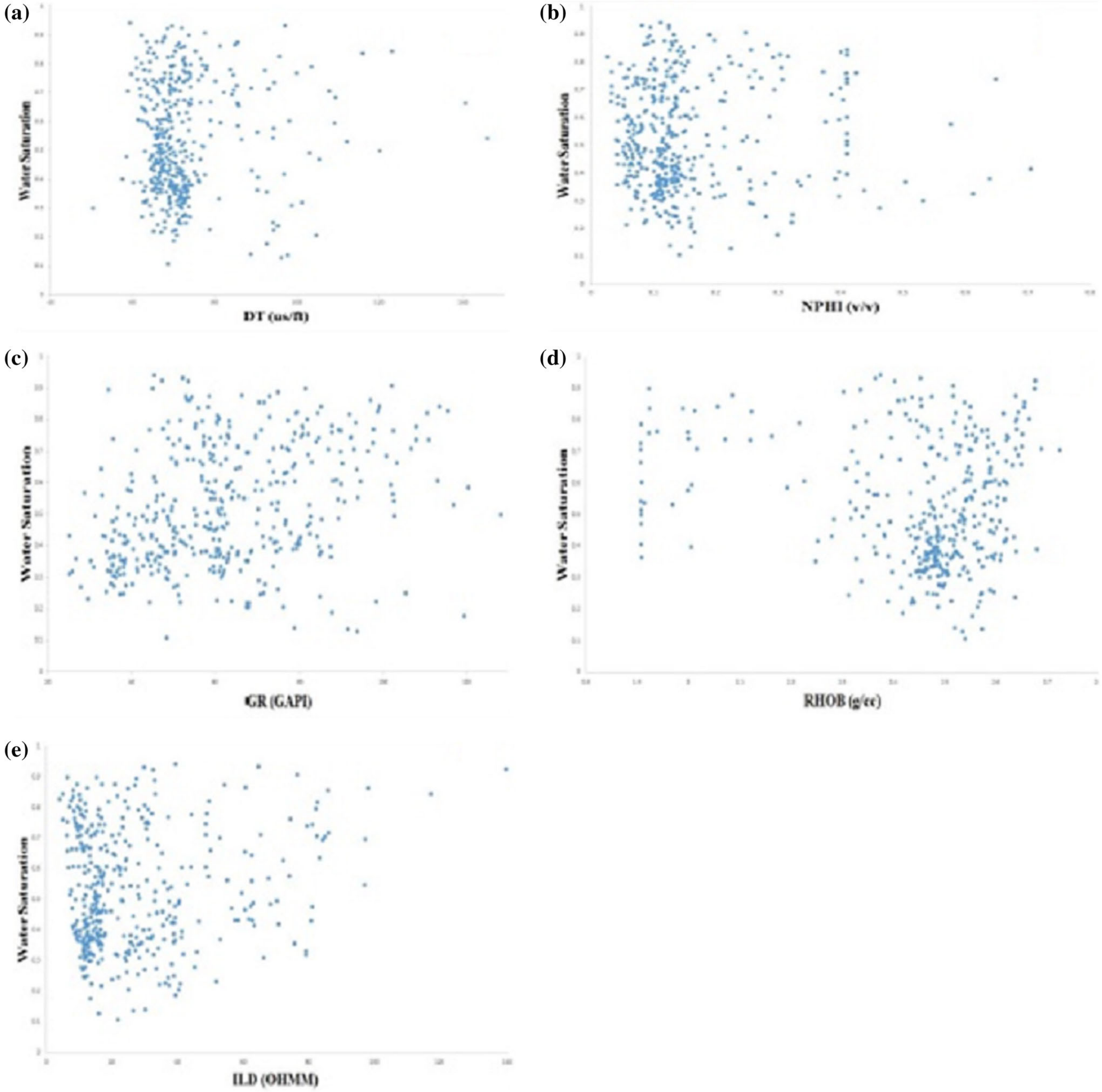

Fig. 5 Scatter plots of selected petrophysical logs versus water saturation. **a** DT versus water saturation. **b** NPHI versus water saturation. **c** GR versus water saturation. **d** RHOB versus water saturation. **e** ILD versus water saturation

technique has the best ability in predicting water saturation by using input data which has not been introduced to model earlier. In this regard, models based on Table 2 were tested by data which have not been used in their training process. Correlation coefficients of 4 employed methods (Decision Tree Forest, Tree Boost, MLP and SVM) among 8 regarded data sets in Table 2 have been compared in Fig. 7.

It can be well understood that SVM performs efficiently and better than other methods based on results showed in Fig. 7. It is also notable that in some data sets other methods has better correlation coefficient than SVM, as in data set 1 MLP and in data set 4 MLP and Decision Tree Forest represent better performance. Decision Tree Forest and MLP have similar results, and both of them can be applicable techniques for prediction of water saturation. It should be mentioned that MLP has relatively better performance rather than Decision Tree Forest. The weakest results are gained by Tree Boost models which reveal that although it has good training capability, it cannot be a reliable method in predicting water saturation based on previously unseen input data.

As it can be seen in Fig. 7, predictions made by all methods in data sets 5 and 8 have more accuracy rather than predictions made in other data sets. It should be regarded that in these two data sets, training data includes data gathered from two wells and provide more training data points, and as result, models in data sets 5 and 8 are well trained and it enables them to predict water saturation better than models trained with fewer data points.

Table 4 Error statistics formulas

Error type	Formula
Correlation coefficient, r	$\frac{\sum_{i=1}^{l}(y_i-\overline{y_i})(\hat{y}_i-\overline{\hat{y}_i})}{\sqrt{\sum_{i=1}^{l}(y_i-\overline{y_i})^2\sum_{i=1}^{N_p}(\hat{y}_i-\overline{\hat{y}_i})^2}}$
Average absolute error, AAE	$\frac{1}{l}\sum_{i=1}^{l}\lvert y_i-\hat{y}_i\rvert$
Root-mean-square error, RMSE	$\sqrt{\frac{1}{l}\sum_{i=1}^{l}(y_i-\hat{y}_i)^2}$

Table 5 Error measures in different kernel functions of SVM

Kernel function	r (%)	AAE	RMSE
Sigmoid	68	0.136	0.114
RBF	65	0.092	0.078
Linear	59	0.147	0.121

Error analyses of generalization capability of methods also have been studied by employing AAE and RMSE error statistics. Values of these errors have been presented in Table 7.

Error measures in Table 7 verify the error analyses done by using correlation coefficient. Interpretation of Table 7 is detailed as below:

Data set 1 Decision Tree Forest in comparison with SVM, Tree Boost and MLP models has the lowest average error. SVM also has acceptable error measures, but MLP and Tree Boost have large values of errors.

Data set 2 SVM in comparison with Decision Tree Forest, Tree Boost and MLP models has the lowest average error. MLP and Decision Tree Forest also have acceptable error measures, but Tree Boost has large values of errors.

Data set 3 SVM in comparison with Decision Tree Forest, Tree Boost and MLP models has the lowest average error. MLP and Decision Tree Forest also have acceptable error measures, but Tree Boost has large values of errors.

Data set 4 MLP in comparison with Decision Tree Forest, Tree Boost and MLP models has the lowest

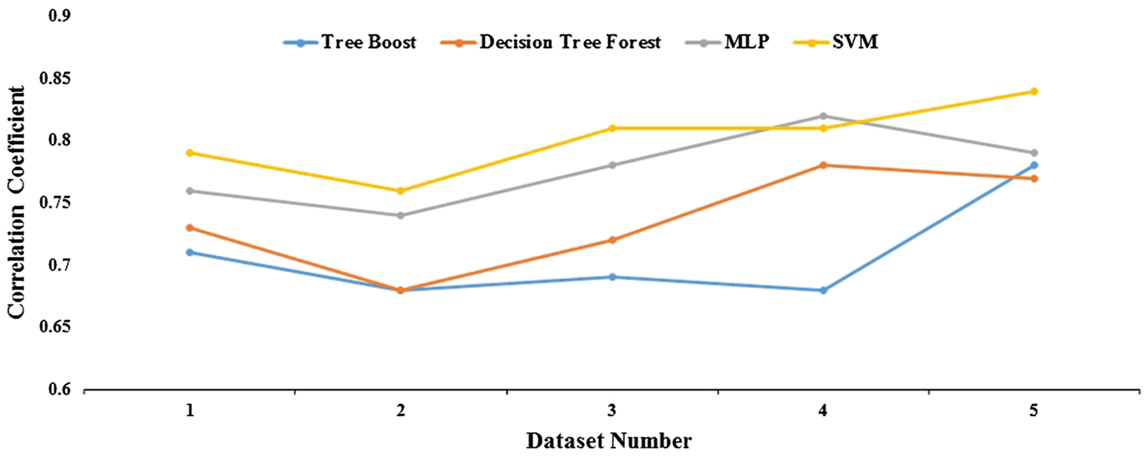

Fig. 6 Correlation coefficient of Tree Boost, Decision Tree Forest, MLP and SVM in predicting water saturation from previously seen data

Table 6 Comparison of RMSE, and AAE error measures between SVM, MLP neural network, Decision Tree Forest and Tree Boost models in order to evaluate training capability

Data set number	Tree Boost		Decision Tree Forest		MLP		SVM	
	AAE	RMSE	AAE	RMSE	AAE	RMSE	AAE	RMSE
1	0.114	0.102	0.102	0.094	0.096	0.081	0.071	0.059
2	0.083	0.087	0.089	0.082	0.083	0.074	0.076	0.062
3	0.103	0.093	0.072	0.066	0.071	0.057	0.068	0.044
4	0.079	0.065	0.074	0.070	0.048	0.039	0.050	0.035
5	0.078	0.071	0.062	0.065	0.059	0.043	0.052	0.033

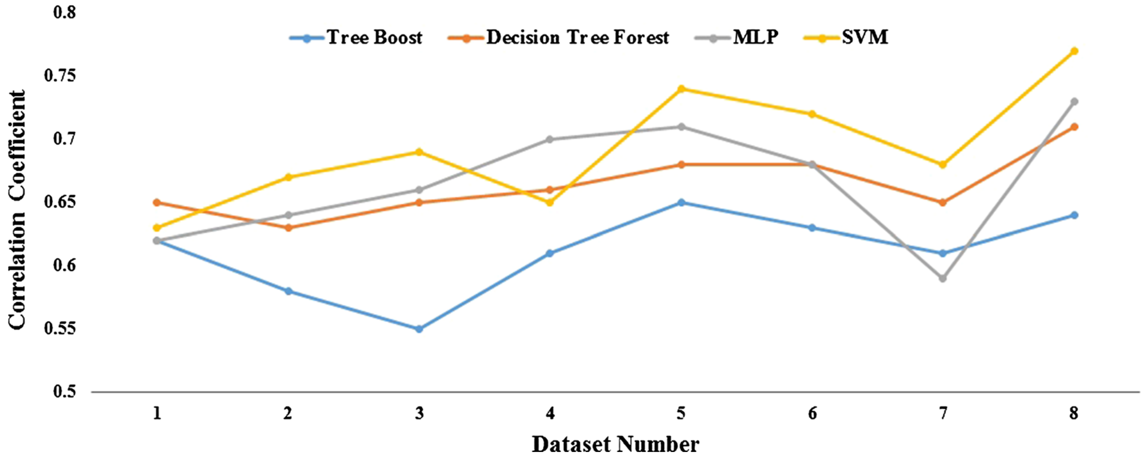

Fig. 7 Correlation coefficient of Tree Boost, Decision Tree Forest, MLP and SVM in predicting water saturation from previously unseen data

Table 7 Comparison of RMSE, and AAE error measures between SVM, MLP neural network, Decision Tree Forest and Tree Boost models in order to evaluate generalization capability

Data set number	Tree Boost		Decision Tree Forest		MLP		SVM	
	AAE	RMSE	AAE	RMSE	AAE	RMSE	AAE	RMSE
1	0.133	0.127	0.083	0.084	0.103	0.099	0.088	0.091
2	0.157	0.135	0.105	0.100	0.095	0.091	0.079	0.073
3	0.166	0.152	0.093	0.107	0.091	0.085	0.081	0.071
4	0.126	0.115	0.101	0.094	0.089	0.074	0.099	0.097
5	0.113	0.108	0.097	0.091	0.087	0.077	0.068	0.064
6	0.125	0.111	0.108	0.099	0.098	0.102	0.075	0.069
7	0.127	0.120	0.131	0.124	0.107	0.100	0.089	0.081
8	0.105	0.093	0.071	0.058	0.074	0.064	0.062	0.051

average error. SVM and Decision Tree Forest also have acceptable error measures, but Tree Boost has large values of errors.

Data set 5 SVM in comparison with Decision Tree Forest, Tree Boost and MLP models has the lowest average error. MLP and Decision Tree Forest also have acceptable error measures, and Tree Boost has better performance in comparison with previous data sets.

Data set 6 SVM in comparison with Decision Tree Forest, Tree Boost and MLP models has the lowest average error. MLP and Decision Tree Forest also have acceptable error measures, but Tree Boost has large values of errors.

Data set 7 SVM in comparison with Decision Tree Forest, Tree Boost and MLP models has the lowest average error. MLP also has acceptable error measures, but Tree Boost and Decision Tree Forest have large values of errors.

Data set 8 SVM in comparison with Decision Tree Forest, Tree Boost and MLP models has the lowest average error. MLP performs better than Decision Tree Forest and Tree boost, but these two methods also have acceptable error measures.

It can be understood that SVM has the best performance in predicting water saturation, because as shown in Figs. 6 and 7, Tables 6 and 7, SVM has the lowest values of error indexes including correlation coefficient, average absolute error and root-mean-square error. MLP and Decision Tree forest are moderate predictors, but Tree Boost cannot be regarded as a powerful method in predicting water saturation.

Scatter plots of core-based values and predicted values of water saturation by each method in data set 6 are presented in Fig. 8. Figures 9, 10, 11 and 12 represent the results of estimating water saturation by all four methods (Tree Boost, Decision Tree Forest, MLP and SVM) in data set 8.

6 Conclusions

In this study, Support Vector Machine, Multilayer Perceptron Neural Network, Decision Tree Forest and Tree Boost methods were used to predict water saturation measures in Mesaverde tight gas sandstones located in Uinta Basin, USA. Also, performances of these methods

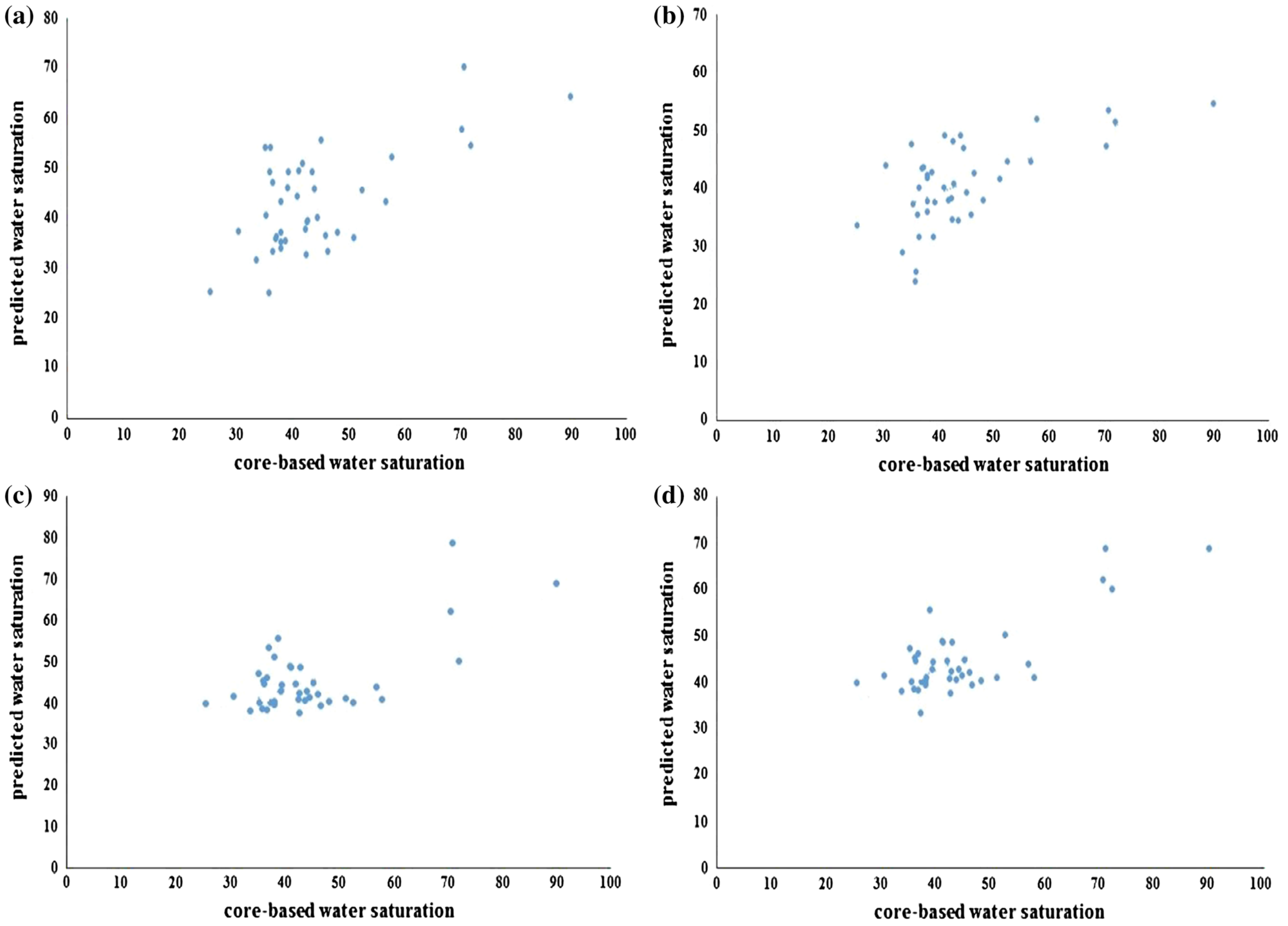

Fig. 8 Scatter plots of predictions made by different methods in data set 6. **a** Prediction made by Tree Boost. **b** Prediction made by Decision Tree Forest. **c** Prediction made by MLP. **d** Prediction made by SVM

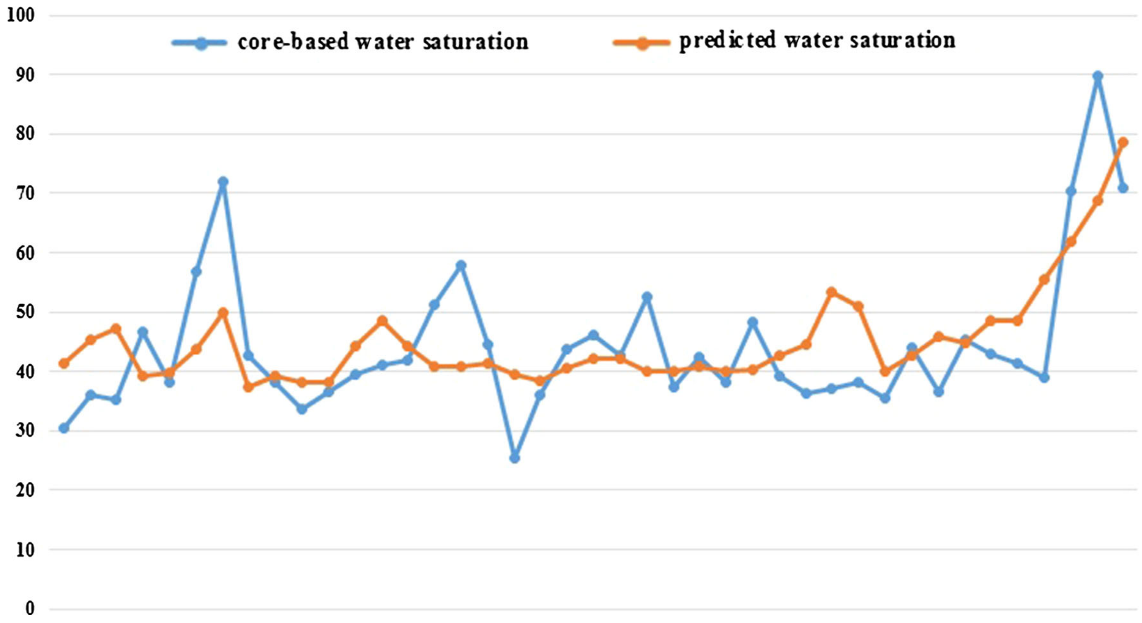

Fig. 9 Comparison of water saturation measured of core and predicted by Tree Boost in data set 8

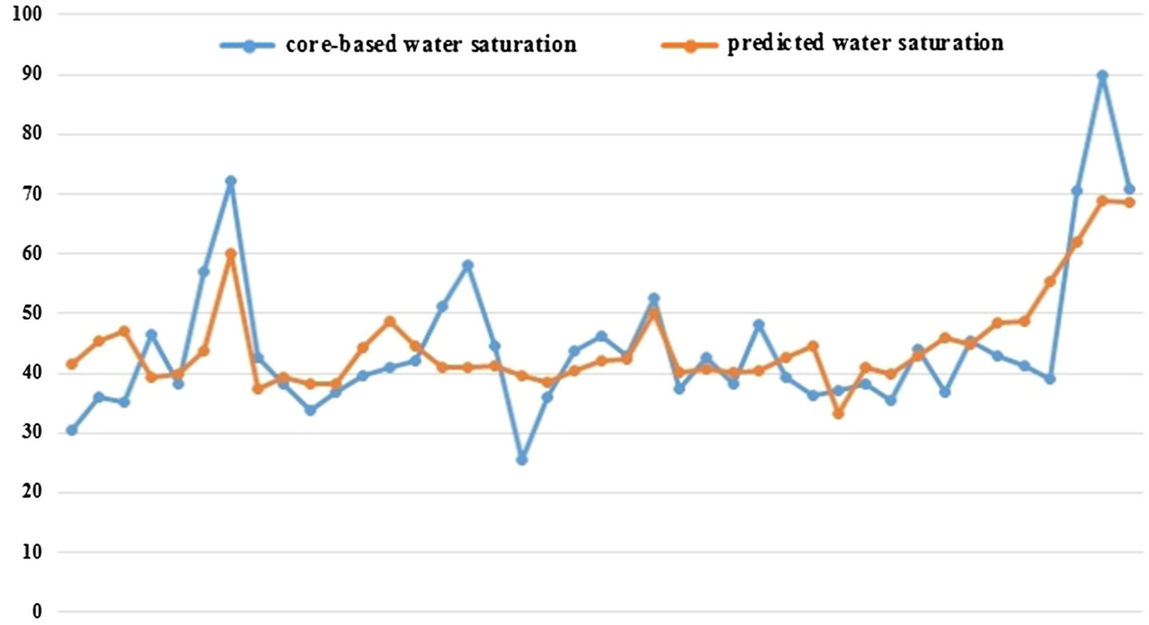

Fig. 10 Comparison of water saturation measured of core and predicted by Decision Tree Forest in data set 8

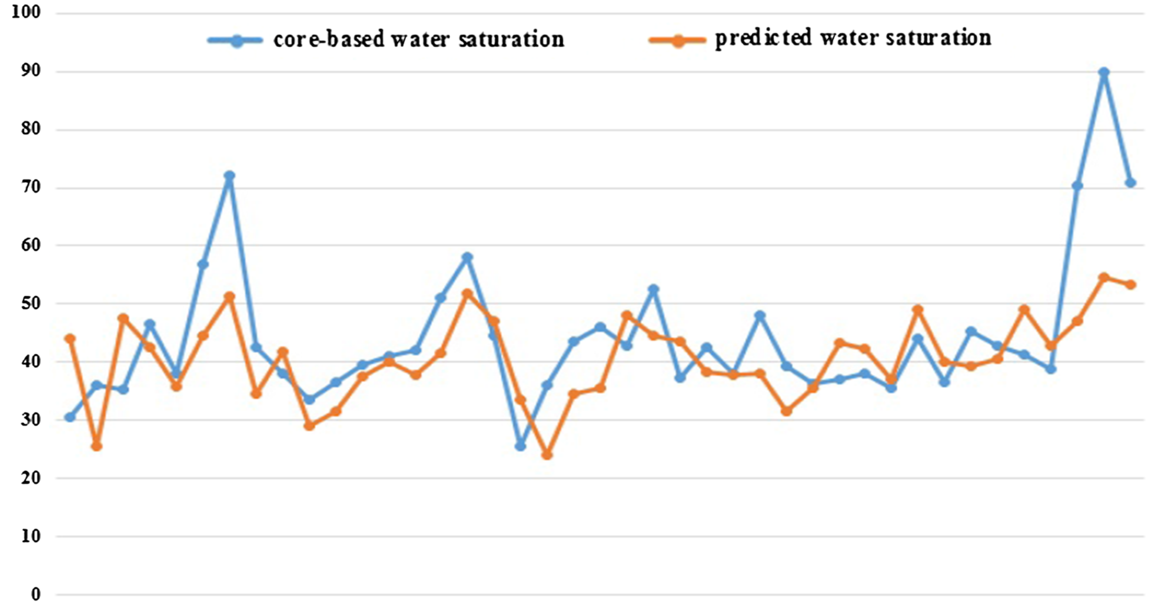

Fig. 11 Comparison of water saturation measured of core and predicted by MLP in data set 8

were compared. Capabilities of methods in predicting water saturation were evaluated in two divided categories including training and generalization. The main conclusions of this study are as follow:

- Support Vector Machine, Multilayer Perceptron Neural Network and Decision Tree Forest are reliable methods in predicting water saturation in the tight gas reservoirs.
- Support Vector Machine has better efficiency in training and generalization rather than other methods.
- Decision Tree Forest performs superior than Tree Boost in the prediction of water saturation, and it represents acceptable results in training and generalization tasks.
- RBF is the best kernel function for SVM in prediction of water saturation.

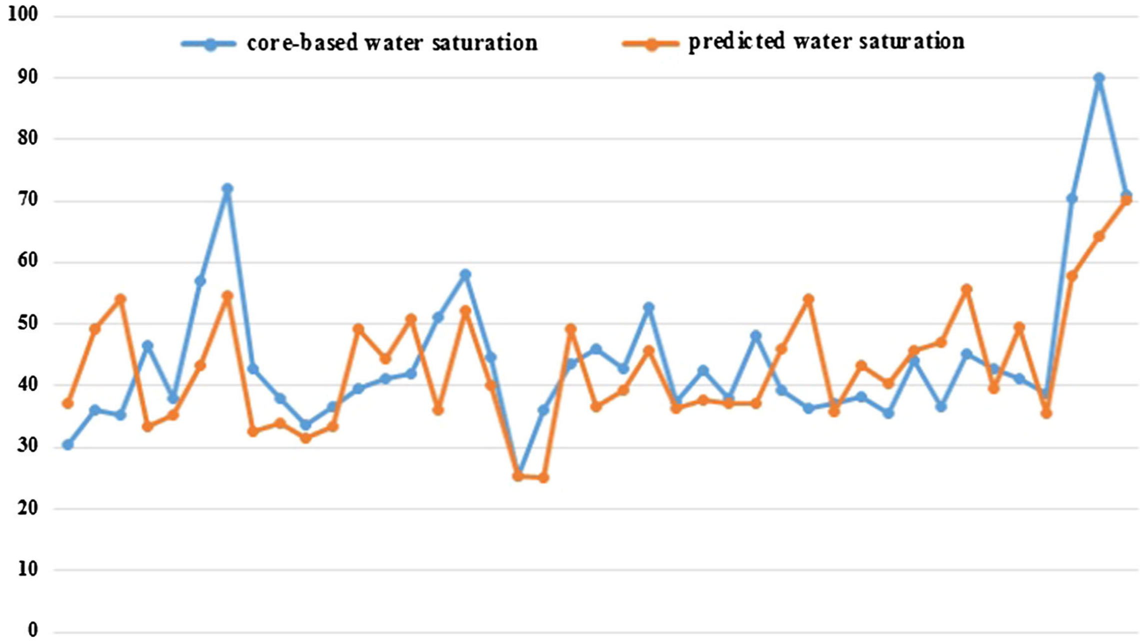

Fig. 12 Comparison of water saturation measured of core and predicted by SVM in data set 8

- Tree Boost cannot be considered as an accurate predictor because of its poor generalization capability.

Acknowledgements The authors appreciate Habibollah Bavarsad Shahripour for his constructive comments and contributions in improving the paper.

References

1. Naik G (2003) Tight gas reservoirs—an unconventional natural energy source for the future. www.sublette-se.org/files/tight_gas.pdf. Accessado em. **1**(07):2008
2. Zhou X, Morrow N, Ma S (2000) Interrelationship of wettability, initial water saturation, aging time, and oil recovery by spontaneous imbibition and waterflooding. SPE J 5(02):199–207
3. Khishvand M, Khamehchi E (2012) Nonlinear risk optimization approach to gas lift allocation optimization. Ind Eng Chem Res 51(6):2637–2643
4. Li K, Horne RN (2001) Characterization of spontaneous water imbibition into gas-saturated rocks. SPE J 6(04):375–384
5. Archie GE (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. Trans AIME 146(1). doi:10.2118/942054-G
6. Poupon A, Leveaux J (1971) Evaluation of water saturation in shaly formations. In: SPWLA 12th annual logging symposium. Society of Petrophysicists and Well-Log Analysts
7. Anifowose F, Labadin J, Abdulraheem A (2013) Predicting petroleum reservoir properties from downhole sensor data using an ensemble model of neural networks. In: Proceedings of workshop on machine learning for sensory data analysis. ACM
8. Anifowose F, Labadin J, Abdulraheem A (2015) Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. Appl Soft Comput 26:483–496
9. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat 28(2):337–407
10. Haykin S (1994) Neural networks: a comprehensive foundation. Prentice Hall PTR, Englewood
11. Kecman V (2005) Support vector machines—an introduction. In: Support vector machines: theory and applications. Springer, Berlin Heidelberg, pp 1–47
12. Karimpouli S, Fathianpour N, Roohi J (2010) A new approach to improve neural networks' algorithm in permeability prediction of petroleum reservoirs using supervised committee machine neural network (SCMNN). J Petrol Sci Eng 73(3):227–232
13. Lachnar J, Zangl G (2006) Treating uncertainties in reservoir-performance prediction with neural networks. J Petrol Technol 58(6):69–71
14. Lim J-S, Kim J (2004) Reservoir porosity and permeability estimation from well logs using fuzzy logic and neural networks. In: SPE Asia Pacific oil and gas conference and exhibition. Society of Petroleum Engineers
15. Mohaghegh S et al. (1996) Petroleum reservoir characterization with the aid of artificial neural networks. J Petrol Sci Eng 16(4):263–274
16. Mohaghegh S et al (1995) Design and development of an artificial neural network for estimation of formation permeability. SPE Comput Appl 7(6):151–154
17. Nikravesh M (2004) Soft computing-based computational intelligent for reservoir characterization. Expert Syst Appl 26(1):19–38
18. Olson TM (1998) Porosity and permeability prediction in low-permeability gas reservoirs from well logs using neural networks. In: Rocky Mountain regional meeting/low permeability reservoirs symposium
19. Ouadfeul S-A, Aliouane L (2012) Lithofacies classification using the multilayer perceptron and the self-organizing neural networks. In: Neural information processing. International conference on neural information processing. Springer, Berlin Heidelberg
20. Ouenes A (2000) Practical application of fuzzy logic and neural networks to fractured reservoir characterization. Comput Geosci 26(8):953–962
21. Rezaee M, Jafari A, Kazemzadeh E (2006) Relationships between permeability, porosity and pore throat size in carbonate rocks

using regression analysis and neural networks. J Geophys Eng 3(4):370

22. Shokir EE-M (2004) Prediction of the hydrocarbon saturation in low resistivity formation via artificial neural network. In: SPE Asia Pacific conference on integrated modelling for asset management. Society of Petroleum Engineers
23. Singh S (2005) Permeability prediction using artificial neural network (ANN): a case study of Uinta Basin. In: SPE annual technical conference and exhibition. Society of Petroleum Engineers
24. Sun Q et al. (2001) Porosity from artificial neural network inversion for Bermejo field, Ecuador. In: SEG expanded abstracts. vol. 20
25. Tadayoni M, Valadkhani M (2012) New approach for the prediction of Klinkenberg permeability in situ for low permeability sandstone in tight gas reservoir. In: SPE middle east unconventional gas conference and exhibition. Society of Petroleum Engineers
26. Tahmasebi P, Hezarkhani A (2012) A fast and independent architecture of artificial neural network for permeability prediction. J Petrol Sci Eng 86:118–126
27. Wiener J, Rogers J, Moll B (1995) Predict permeability from wireline logs using neural networks. Petrol Eng Int 68(5)
28. Wong P, Jian F, Taggart I (1995) A critical comparison of neural networks and discriminant analysis in lithofacies, porosity and permeability predictions. J Petrol Geol 18(2):191–206
29. Zhang Y, Salisch HA, McPherson JG (1999) Application of neural networks to identify lithofacies from well logs*. Explor Geophys 30(1/2):45–49
30. Al-Anazi A, Gates I (2010) Support vector regression for porosity prediction in a heterogeneous reservoir: a comparative study. Comput Geosci 36(12):1494–1503
31. Alcocer Y, Rodrigues P (2001) Neural networks models for estimation of fluid properties. In: SPE Latin American and Caribbean Petroleum Engineering conference. Society of Petroleum Engineers
32. Aliouane L et al. (2012) Petrophysical parameters estimation from well-logs data using multilayer perceptron and radial basis function neural networks. In: Neural information processing. international conference on neural information processing. Springer, Berlin Heidelberg
33. Aminian K, Ameri S (2005) Application of artificial neural networks for reservoir characterization with limited data. J Petrol Sci Eng 49(3):212–222
34. Aminian K et al. (2003) Prediction of flow units and permeability using artificial neural networks. In: SPE western regional/AAPG pacific section joint meeting. Society of petroleum engineers
35. Asadisaghandi J, Tahmasebi P (2011) Comparative evaluation of back-propagation neural network learning algorithms and empirical correlations for prediction of oil PVT properties in Iran oilfields. J Petrol Sci Eng 78(2):464–475
36. Baneshi M et al (2013) Predicting log data by using artificial neural networks to approximate Petrophysical parameters of formation. Petrol Sci Technol 31(12):1238–1248
37. Baziar S et al (2014) Prediction of permeability in a tight gas reservoir by using three soft computing approaches: a comparative study. J Nat Gas Sci Eng 21:718–724
38. Zhao B et al (2006) Water saturation estimation using support vector machine. In: SEG/New Orleans 2006 annual meeting
39. Bhatt A (2002) Reservoir properties from well logs using neural networks
40. Boadu FK (2001) Predicting oil saturation from velocities using petrophysical models and artificial neural networks. J Petrol Sci Eng 30(3):143–154
41. Carrasquilla A, Silvab J, Flexac R (2008) Associating fuzzy logic, neural networks and multivariable statistic methodologies in the automatic identification of oil reservoir lithologies through well logs. Rev Geol 21(1):27–34
42. Hamada G, Elshafei M (2009) Neural network prediction of porosity and permeability of heterogeneous gas sand reservoirs. In: SPE Saudia Arabia section technical symposium. Society of Petroleum Engineers
43. Helle HB, Bhatt A, Ursin B (2001) Porosity and permeability prediction from wireline logs using artificial neural networks: a North Sea case study. Geophys Prospect 49(4):431–444
44. Huang Z et al (1996) Permeability prediction with artificial neural network modeling in the Venture gas field, offshore eastern Canada. Geophysics 61(2)
45. Irani R, Nasimi R (2011) Evolving neural network using real coded genetic algorithm for permeability estimation of the reservoir. Expert Syst Appl 38(8):9862–9866
46. Jamialahmadi M, Javadpour F (2000) Relationship of permeability, porosity and depth using an artificial neural network. J Petrol Sci Eng 26(1):235–239
47. Kapur L et al (1998) Facies prediction from core and log data using artificial neural network technology. In: SPWLA 39th annual logging symposium. Society of Petrophysicists and Well-Log Analysts
48. Naseri A, Nikazar M, Dehghani SM (2005) A correlation approach for prediction of crude oil viscosities. J Petrol Sci Eng 47(3):163–174
49. Rogers SJ et al (1992) Determination of lithology from well logs using a neural network (1). AAPG Bull 76(5):731–739
50. Khishvand M, Naseri A (2012) An artificial neural network approach to predict asphaltene deposition test result. Fluid Phase Equilib 329:32–41
51. Baziar S, Shahripour HB (2015) A novel correlation approach to predict total formation volume factor, using artificial intelligence
52. Hemmati-Sarapardeh A et al (2013) Toward reservoir oil viscosity correlation. Chem Eng Sci 90:53–68
53. Van Der Baan M, Jutten C (2000) Neural networks in geophysical applications. Geophysics 65(4):1032–1047
54. Wong PM, Gedeon TD, Taggart IJ (1995) An improved technique in porosity prediction: a neural network approach. IEEE Trans Geosci Remote Sens 33(4):971–980
55. Al-Bulushi N, Araujo M, Kraaijveld M (2007) Predicting water saturation using artificial neural networks (ANNS). Neural Netw 549(198):57
56. Basbug B, Karpyn ZT (2007) Estimation of permeability from porosity specific surface area and irreducible water saturation using an artificial neural network. In: Latin American and Caribbean Petroleum Engineering conference. Society of Petroleum Engineers
57. Goda HM, Maier H, Behrenbruch P (2005) The development of an optimal artificial neural network model for estimating initial water saturation-Australian reservoir. In: SPE Asia Pacific oil and gas conference and exhibition. Society of Petroleum Engineers
58. Goda HM, Maier H, Behrenbruch P (2007) Use of artificial intelligence techniques for predicting irreducible water saturation-Australian hydrocarbon basins. In: Asia Pacific oil and gas conference and exhibition. Society of Petroleum Engineers
59. Ibrahim MA, Potter DK (2004) Prediction of residual water saturation using genetically focused neural nets. In: SPE Asia Pacific oil and gas conference and exhibition. Society of Petroleum Engineers
60. Mollajan A, Memarian H (2013) Estimation of water saturation from petrophysical logs using radial basis function neural network. J Tethys 1(2):156–163
61. Vapnik VN, Chervonenkis AJ (1974) Theory of pattern recognition [in Russian]. Nauka, Moscow
62. Vapnik V (1982) Estimation of dependences based on empirical data. Springer, New York

63. Vapnik V (2000) The nature of statistical learning theory. Springer, New York
64. Mukherjee S, Osuna E, Girosi F (1997) Nonlinear prediction of chaotic time series using support vector machines. In: Proceedings of the 1997 IEEE workshop neural networks for signal processing [1997] VII. IEEE
65. Jeng J-T (2005) Hybrid approach of selecting hyperparameters of support vector machine for regression. IEEE Trans Syst Man Cybern Part B Cybern 36(3):699–709
66. Al-Anazi A, Gates I (2010) Support vector regression to predict porosity and permeability: effect of sample size. Comput Geosci 39:64–76
67. Al-Anazi A, Gates I (2010) On the capability of support vector machines to classify lithology from well logs. Nat Resour Res 19(2):125–139
68. Al-Anazi A, Gates ID (2010) A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. Eng Geol 114(3–4):267–277
69. Al-anazi AF, Gates ID, Azaiez J (2009) Innovative data-driven permeability prediction in a heterogeneous reservoir. In: EUROPEC/EAGE conference and exhibition. Society of Petroleum Engineers
70. Anifowose FA, Ewenla AO, Eludiora SI (2011) Prediction of oil and gas reservoir properties using support vector machines. In: International petroleum technology conference. International Petroleum Technology Conference
71. Gholami R, Shahraki AR, Jamali Paghaleh M (2012) Prediction of hydrocarbon reservoirs permeability using support vector machine. Math Probl Eng 2012(2012). doi:10.1155/2012/670723
72. Nazari S, Kuzma HA, Rector III JW (2011) Predicting Permeability from well log data and core measurements using support vector machines. In: 2011 SEG annual meeting. Society of Exploration Geophysicists
73. Saffarzadeh S, Shadizadeh SR (2012) Reservoir rock permeability prediction using support vector regression in an Iranian oil field. J Geophys Eng 9(3):336
74. Yue Y, Wang J (2007) SVM method for predicting the thickness of sandstone. Appl Geophys 4(4):276–281
75. Kamari A et al (2013) Prediction of sour gas compressibility factor using an intelligent approach. Fuel Process Technol 116:209–216
76. Hemmati-Sarapardeh A et al (2014) Reservoir oil viscosity determination using a rigorous approach. Fuel 116:39–48
77. Mollajan A, Memarian H, Jalali M (2013) Prediction of reservoir water saturation using support vector regression in an iranian carbonate reservoir. In: 47th US rock mechanics/geomechanics symposium. American Rock Mechanics Association
78. Breiman L (2001) Decision-tree forests. Mach Learn 45(1):5–32
79. Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227
80. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
81. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38(4):367–378
82. Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35(1):73–101
83. Rumelhart DE, Hinton GE, Williams RJ (1985) Learning internal representations by error propagation. No ICS-8506. California University of San Diego La Jolla Institute for Cognitive Science
84. Rosenblatt F (1961) Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Spartan Books, Washington
85. Basak D, Pal S, Patranabis DC (2007) Support vector regression. Neural Inf Process Lett Rev 11(10):203–224
86. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. ACM
87. Guyon I, Boser B, Vapnik V (1996) Automatic capacity tuning of very large VC-dimension classifiers. Adv Neural Inf Process Syst (5):147–147
88. Cortes C, Vapnik V (1995) Support-vector networks. Mach learn 20(3):273–297
89. Schölkopf B, Burgest C, Vapnik V (1995) Extracting support data for a given task. In: Proceedings of the 1st international conference on knowledge discovery & data mining
90. Schölkopf B, Burges C, Vapnik V (1996) Incorporating invariances in support vector learning machines. In: Artificial neural networks ICANN 96. Springer pp 47–52
91. Vapnik V, Golowich SE, Smola A (1997) Support vector method for function approximation, regression estimation, and signal processing. Adv Neural Inf Process Syst (6):281–287
92. Vapnik V, Chervonenkis AJ (1964) A class of perceptrons. Autom Remote Control 25(1):1964
93. Vapnik V, Lerner A (1963) Generalized portrait method for pattern recognition. Autom Remote Control 24(6):774–780
94. Cumella SP, Scheevel J (2008) The influence of stratigraphy and rock mechanics on Mesaverde gas distribution. Piceance Basin, Colorado