Determination of oil well production performance using artificial neural network (ANN) linked to the particle swarm optimization (PSO) tool

Mohammad Ali Ahmadi, Reza Soleimani, Moonyong Lee, Tomoaki Kashiwao, Alireza Bahadori

1. Introduction

Petroleum engineers are always interested in finding appropriate and reliable tools to predict the productivity of horizontal wells. Accurate predictions are very important to conduct technical and economical feasible studies before drilling. The wells which are very costly. It is a very important factor to decide on the economical feasibility of drilling horizontal wells [1–7].

According to the experimental investigation conducted by Yasar et al. [8], applied load and torque for both specific energy and penetration rate in drilling operations have great importance. Based on their results, penetration rate decreases dramatically with raising Unconfined Compressive Strength (UCS) of the cement grout being drilled, proposing that rock characters will intensely effect advance in the drilling operation, and determined specific energy decreases dramatically with both raising applied load and raising penetration rate in the drilling operation [8]. Moreover, Wasantha et al. [9] studied the effect of strain rate on the mechanical behavior of sandstones with various grain sizes. They concluded that size of constituent grains is an essential factor which requires to be encompassed in the reservoir simulation.
thoughts of the mechanical behavior of sandstones under various strain rates [9].

The pseudo-steady state occurs when the fluid production manifests its impact in the boundaries far from the production well where the oil is approaching from the drainage boundary on the way to the depletion area around the well [10,11]. It indicates that the reservoir has reached a condition where the pressure at all reservoir boundaries and also the average reservoir pressure will reduce over time as more fluid is extracted from the reservoir [12,13].

A suitable technique to estimate the pressure transient response of producing horizontal wells was developed by Clonts and Ramey [14]; where the source functions are combined with the Newman product approach. Recently, an analytical approach was also introduced by Hagoort [15] to calculate the productivity of a horizontal well with infinite conductivity placed in a reservoir which is considered as a closed rectangular system.

The resultant of the source functions was utilized by Daviau et al. [16] for oil reservoirs with particular characteristics. A reliable solution was presented by Ozkan et al. [17] and Joshi [18] to estimate pressure response for the anisotropic and infinite reservoirs which experience either uniform influx or include a horizontal well with infinite conductivity.

Several researchers performed a variety of works to use the source function scheme to obtain profile or/and magnitudes of pressure in surrounded reservoirs [16,45,19,20,721]. The horizontal well-bore was considered as a strip source by Goode and Thambynayagam [22]. In their work a uniform distribution of flow rate in the well length is maintained. The pressure response was found for the horizontal producing well through combination of the finite Fourier cosine and Laplace transforms.

Depletion mode, especially in linear flow patterns depends on well productivity in non-circular flow configurations. Helmy and Wattenbarger [23] noticed that this observation matches well with the analytical long-time solutions for the analogous linear heat flow problems.

In general, the horizontal wells are not fully drilled in horizontal direction such that a significant change in vertical elevation happens during the drilling, leading to huge effects on outlet pressure of the horizontal well. In addition, the computation usually is difficult due to negative skin factor of horizontal wells. On the other hand, accurate calculation of the length of horizontal production well is not simple in many cases [2].

Production at a constant well rate plays a major role in current analytical approaches to study the productivity of horizontal wells in closed reservoirs [24,25].

Helmy and Wattenbarger [26] presented simple productivity correlations using several number of reservoir simulations to calculate the productivity of horizontal wells. The reservoir model contained a homogeneous, isotropic, closed reservoir with the shape of a rectangular box and an open hole horizontal well parallel to one of the sides of the box.

A semi-analytical model to determine the productivity of wells in Darcy flow systems was presented by Hagoort [27].

Finally, summary of the productivity models in steady state and pseudo-steady state conditions is demonstrated in Table 1. It should be noted that, each of equations should only use in the specific circumstances. In other words, various assumptions should be met to use each correlations (for instance, steady state flow regime, homogeneous porous media and etc) and this is a big disadvantage for petroleum engineers to employ the aforementioned correlations. Moreover, most of the correlations developed are depending on the size of the system because they are not dimensionless and this is another drawback for using these formulas.

An attempt was made by Mutalik et al. [28] to compile different data sets of shape factors and the corresponding skin factors (SCah), where different horizontal wells are considered at various locations of the drainage area in the reservoir [2]. The presence of an enclosed reservoir and an infinite-conductivity well are the main assumptions in their methodology. The horizontal well productivity is determined by an equation as given below [2]:

$$J_h = \frac{q_0}{P_{R} - P_{wf}}$$

$$= \frac{kh}{141.2 \beta_o \mu_o \left( \ln \left( \frac{L}{r_e} \right) - A' + S_r + S_m + S_{CAh} - C' + Dq_o \right)}$$

where

$$r_e = \sqrt{\frac{A}{\pi}}$$

$$S_r = -\ln \left( \frac{L}{4r_w} \right)$$

Bahadori et al. [29] introduced a predictive tool by means of Vandermonde matrix concepts, for estimating pseudo skin factor of horizontal wells through rectangular drainage area. They noticed that the calculation of productivity of a horizontal oil well depends on the pseudo-skin factor for centrally located wells within different drainage areas.

Given the above matters, developing a proper and straightforward correlation seems necessary. Compared to available approaches, the developed equation is expected to be less complicated, and more accurate for predicting the pseudo-skin factor as a function of dimensionless length (Ld) and the ratio of horizontal well length over drainage area side (L/2Xc) for square and rectangular shapes with ratios of sides 1, 2, and 5.

$$L_d = \frac{L}{2R} \sqrt{\frac{K_v}{K_h}}$$

In this article, an intelligent method utilizing a new type of network modeling which called “Least Square Support Vector Machine (LSSVM)” is developed to serve as a rapid and inexpensive predictive model for monitoring well productivity of horizontal wells in box shape drainage area. The proposed LSSVM model is developed implementing extensive actual well productivity data.

To depict the robustness, integrity and accuracy of the suggested LSSVM model, the obtained outcomes from the introduced approach are contrasted with the relevant actual productivity data. Outputs from this research reveal that the evolved approach can monitor the horizontal well productivity with high accuracy. The introduced predictive model can be utilized as a reliable way for quick and cheap but efficient prediction of well productivity of horizontal well parameter in absence of appropriate experimental or/and real data, specifically through the initial stages of evolution of horizontal well drilling.
which demonstrate the Cartesian coordinates such as length of duration which is characterized by the straightforward analytical

| Fig. 1 | illustrates the schematic of the horizontal well configuration which is characterized by the straightforward analytical approach: a cubic reservoir with a horizontal well parallel to the top and bottom of the cube. The reservoir has a width w and thickness h and length 2X. An open hole horizontal well with a length of L is assumed drilled in the box shaped reservoir.

The location of the drilled well is determined by three points which demonstrate the Cartesian coordinates such as x, y, and z. It should be noted that the Cartesian coordinates of the midpoint of the horizontal well throughout the Cartesian system which is tier collateral to the edges of the cubic reservoir. Moreover, direction of the horizontal well is parallel to the x-axis of Cartesian coordinate system. The aforementioned reservoir is isotropic and homogeneous and consists a compressible fluid with a constant compressibility and constant viscosity. No flow boundaries are assumed for the outer boundaries of the reservoir. Production from well may be at constant pressure or constant rate. Well conductivity in aforementioned reservoir is consumed infinite, which conveys pressure distribution throughout the well bore is uniform.

The 2D flow configuration of horizontal well in the abovementioned system is depicted in Fig. 2. As clear be seen in this figure, two types of flow exist in horizontal wells including horizontal flow to a fully penetrating vertical fracture with length L1, and the flow to a fully penetrating horizontal well in a rectangular reservoir of thickness h.

3. Methodology

3.1. Artificial neural network

The artificial neural networks (ANNs) have been built on the basis of function and configuration of the brain of human, nerve networks, and complicated procedure of learning and reacting.
aims to produce the magnitudes of target parameters from input information through internal computations and analysis. In general, this connectionist technique involves well interconnected models along with a simple processing constituent (or neuron) that is able to find out the right links between independent and dependent parameters. The multilayer feed forward neural network is the most common and acceptable ANN system, where the neurons are coordinated into different layers; namely, input, output, and hidden.

Multilayer feed forward neural networks usually have one or more hidden layers, which enable the networks to model non-linear and complex functions. The hidden layer is responsible to convey meaningful information between the input and network stages through an effective approach. ANNs interface predictors through the neurons of an input layer and transmit output of dependent variable(s) through the neurons of an output layer. Each neuron in hidden and output layers carries a transfer function that illustrates internal activation phenomenon. A neuron output is normally achieved through transformation of its input employing an opposite transfer function.

Transfer functions may be linear or non-linear. It should be mentioned that, each interconnection in an ANN has a strength that is expressed by a number referred to as weight. Furthermore, bias is an extra input appended to neurons, and which all the time holds a value of 1 and treated similar to other weights [32]. Fig. 3 illustrates the structure of a feed forward ANN with one hidden layer used in this study.
A brief procedures to produce the output variable, using the input data, are described below. Given the info for inputs neurons, the net inputs (S) for the hidden neurons are computed as follows:

\[ S^H_j = \sum_{l=1}^{2} w^H_{j} \cdot a_l + b^H_j \]  

(5)

where \( a_l \) is the vector of the input parameters, \( j \) is the index of hidden neuron, \( w^H_{j} \) denotes the interconnection weight between the input neurons with the hidden layer, and the term \( b^H_j \) stands for the bias of the \( j \)th hidden neuron. Then, the outputs \( (L_j) \) of the hidden neurons are computed, utilizing a transfer function \( f_H \) which is associated with the hidden neurons.

\[ L_j = f_H (S^H_j) = f_H \left( \sum_{l=1}^{2} w^H_{j} \cdot a_l + b^H_j \right) \]  

(6)

The produced outputs from the hidden layer are presented to the output layer as inputs. The output of the output layer (the final output, \( Y \)) can be also calculated in similar way. Indeed, it is responsibility of all neurons in a network to relate the input information to the desired output function through an adequate and suitable correlation.

3.2. Least squares support vector machines

Support vector machine (SVM) is a new kind of learning machine which was developed within the context of statistical learning theory and structural risk minimization by Vapnik and his colleagues [33–36,37–41]. SVM is a very nice methodology which has recently attracted special attention from a great deal of researchers for solving problems in nonlinear classification, function estimation as well as for solving density estimation problems [42–36,37–41], due to it enjoys several interesting properties such as the use of the kernel-induced feature spaces, the sparseness of the solution, high generalization ability, together with excellent performance.

An evolved version of SVM, called LSSVM was recently introduced by Suykens and co-workers in which the inequality constraints of an SVM were changed to a set of equality constraints, in order to facilitate the original SVM method [33–36,37–41,43,44]. Therefore, the solution is obtained by solving a system of linear equations, resulting in an easier-to-implement and faster alternative to the original SVM method [45–48,43,44]. It should be mentioned that, this section of the paper serves as an overview to LSSVM. Details of this method can be found in the literature [33–36,37–41,43,44].

Assume a set of \( N \) training data \( \{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\} \), where \( x_k \in \mathbb{R}^p \) is the \( k \)th input data and \( y_k \in \mathbb{R} \) is corresponding output value. The aim is to estimate a model of the formula [33–36,37–41]:

\[ y(x) = \varphi^T(x) + b \]  

(7)

where \( \varphi(\cdot): \mathbb{R}^p \rightarrow \mathbb{R}^m \) is a nonlinear mapping function, which maps the input data to higher dimensional feature space; \( b \) is the bias term and \( \varphi \) denotes the weight vector needed to be determined, which can be calculated by minimizing the following function [33–36,37–41]:

\[ J(\varphi,e) = \frac{1}{2} \varphi^T\varphi + \frac{1}{2} \gamma \sum_{k=1}^{N} e_k^2 \]  

(8)

Subject to the following equality constraint [33–36,37–41]:

\[ y_k = \varphi^T(x_k) + b + e_k, \quad k = 1, 2, \ldots, N \]  

(9)

where \( \gamma \) is the regularization constant used for preventing over fitting, and \( e_k \) is the training data error. After then, the Lagrangian function is adopted as follow in order to solve the constrained optimization problem [33–36,37–41]:

\[ J(\varphi,b,e,a) = J(\varphi,e) - \sum_{k=1}^{N} a_k \left( \varphi^T(x_k) - b - e_k - y_k \right) \]  

(10)

where \( a_k \) are Lagrange multipliers. By executing the first order partial derivatives of Lagrange function (8) with respect to \( \varphi, b, e_k \) and \( a_k \) the conditions for optimality can be written as follow [33–36,37–41]:

\[ \frac{\partial J}{\partial \varphi} = \varphi - \sum_{k=1}^{N} a_k \varphi(x_k) = 0 \]  

(11)

\[ \frac{\partial J}{\partial b} = \sum_{k=1}^{N} a_k = 0 \]  

(12)

\[ \frac{\partial J}{\partial e_k} = a_k - \gamma e_k = 0, \quad k = 1, \ldots, N \]  

(13)

\[ \frac{\partial J}{\partial e_k} = a_k - \gamma e_k = 0, \quad k = 1, \ldots, N \]  

(14)

When the variables \( \varphi \) and \( e \) are removed, the optimization problems can be described as a linear system [33–36,37–41].

\[ \begin{bmatrix} 0 & 1^T \alpha \Omega + \gamma^{-1} \mathbb{I} \end{bmatrix} \begin{bmatrix} b \alpha \end{bmatrix} = \begin{bmatrix} 0 \ y \end{bmatrix} \]  

(15)

where \( y = [y_1 \ldots y_N]^T, \alpha = [\alpha_1 \ldots \alpha_N]^T, l \) is an identity matrix and \( \Omega \) is an \( N \)-dimensional symmetric matrix:

\[ \Omega_{kl} = \varphi(x_k)^T \varphi(x_l) = K(x_k,x_l) \quad \forall k, l = 1, \ldots, N \]  

(16)

where \( (b, \alpha) \) is the solution to the linear system (15).

4. Developed models

In the current research, we applied two types of artificial intelligence based methodologies, viz. ANN and LSSVM to build the predictive models to calculate the shape-related skin factor or pseudo-skin factor in horizontal wells as a function of dimensionless length and the ratio of horizontal well length over drainage area side for various drainage areas based on previously published database [2,28].

4.1. ANN model

4.1.1. Data distribution (training and testing subsets)

When training multilayer feed-forward neural networks, the general practice is to first separate the data into two subsets [45]. The first subset is the training set, which is used for updating the
network weights and biases. The accuracy of trained neural network can be independently measured by the second data subset called testing set.

This will give us a sense of how well the network will do when applied to data from the real world. Herein, total of 96 data points (80% of whole data set) were used as training data and main data set were considered as a test data.

4.1.2. Training method and transfer functions

A key consideration in ANN design is the type of transfer functions. Different transfer functions can be used for neurons in the hidden and output layers. Among different transfer functions available, the log-sigmoid transfer function (logsig), which is an appropriate choice for many nonlinear functions, was employed as the transfer function for neurons in hidden layer which is mathematically expressed as:

$$f(S) = \frac{1}{1 + \exp(-S)} \quad (17)$$

The logsig function is bounded between 0 and 1. One of the most remarkable features of this transfer function lies in its $S$ shape which provides appropriate gain for a wide range of input levels. It should be mentioned that, ANNs be indebted their non-linear ability to exploit such non-linear transfer functions [50].

Also, for output layer, linear transfer function (purelin) is applied as a transfer function which is represented as:

$$f(S) = S \quad (18)$$

Outputs in the range of $-\infty$ to $+\infty$, can be produced from the linear transfer function [50].

ANNs' ability to deal with problems successfully greatly depends on applying an appropriate and effective training algorithm. Training process is basically an optimization process by which the weights of interconnections between neurons together with the biases are adjusted so that the network can predict the correct outputs for a given set of inputs.

There are many different types of training algorithms which have been applied for various applications such as Back Propagation (BP) [45], Genetic Algorithm (GA) [51], Particle Swarm Optimization (PSO) [47,52,50,53], Hybrid Genetic Algorithm and Particle Swarm Optimization (HGAPSO) [54,55], Unified Particle Swarm Optimization (UPSO) [56] and Imperialist Competitive Algorithm (ICA) [57,58].

In this study, PSO algorithm was implemented as a training algorithm for updating the weights and biases of multilayer feed forward network based on success stories of this new evolutionary computation algorithm for training ANNs in addition to its outstanding features including robustness, simplicity, flexibility, superior convergence characteristics and many others.

PSO is a population-based optimization algorithm based on the simulation of the social behavior of bird flocks, fish schooling and swarm of insects developed by [59], where the system is randomly initialized with a swarm of particles (potential solutions) and then searches the optima within the search space through updating generations. Particles inside the swarm are drawn progressively towards the global optimum as the system iterates. In each iteration, each particle knows and stores the two best values. The earliest value is the personal best position $p_{i,pbest}$ that is the location of the particle $i$ in the search space, where it has attained the preeminent solution up to now, based on the fitness function. The global best solution $p_{gbest}$ is taken into account as the second one that introduces a position, leading to the best solution among all the particles on the same basis. Using this information, the velocity and position of each particle are updated according to the following formulae and consequently the new swarm of particles is generated with improved characteristics:

$$v_{i}^{n+1} = \omega v_{i}^{n} + c_{1} r_{1}^{n} [p_{i,pbest}^{n} - p_{i}^{n}] + c_{2} r_{2}^{n} [p_{gbest}^{n} - p_{i}^{n}] \quad (19)$$

$$p_{i}^{n+1} = p_{i}^{n} + v_{i}^{n+1} \quad (20)$$

where $v_{i}^{n}$ and $v_{i}^{n+1}$ are velocities of particle $i$ at iterations $n$ and $n + 1$; $p_{i}^{n}$ and $p_{i}^{n+1}$ are positions of particle $i$ at iterations $n$ and $n + 1$; $\omega$ is the inertia weight that plays an important role to handle the exploration and exploitation of the search space, as it continuously corrects the value of velocity; $c_{1}$ and $c_{2}$ are termed as cognition and social components respectively are the acceleration constants which changes the velocity of a particle towards $p_{i,pbest}^{n}$ and $p_{gbest}^{n}$ (generally somewhere between $p_{i,pbest}^{n}$ and $p_{gbest}^{n}$) [60]. $r_{1}^{n}$ and $r_{2}^{n}$ are two random numbers uniformly distributed in [0,1]; $p_{gbest}^{n}$ is the best position in the current swarm over generation $n$ determined by the fitness function; and $p_{i,pbest}^{n}$ is the best position of particle $i$ over generation $n$ determined by the fitness function. A complete description and more details of PSO can be found at a number of publications [60–65]. Herein, only the basic principles were briefly presented.

An opposite representation and fitness function ($f$) appear to be necessary before the PSO is utilized to train the connectionist network model. Herein, neural network fitness was decided to be the mean square of errors (MSE) of the whole training data set [66]:

$$f = \frac{1}{N} \sum_{i=1}^{N} \left(y_{i}^{exp} - y_{i}^{pre}\right)^{2} \quad (21)$$

where $N$ represents the total number of data points (input and output pairs), $y^{exp}$ is the actual (experimental) at the sampling point $i$, $y^{pre}$ is the ith output of the network. Every particle characterizes a possible solution to the optimization case as biased of a trained neural network and the corresponding weights are counted a solution, one comprehensive network is represented by a single particle.

Each component of a particle’s position vector represents one neural network weight or bias [54]. When the pre-specified criterion is reached, the iterations cease and the set of these optimized weights and biases found is finalized and used in the proposed architecture (see Fig. 3). The detailed flowcharts of the implementations of PSO algorithm for training the proposed ANN architecture are shown in Fig. 4.

4.1.3. ANN structure

Network structure is another important issue that needs consideration due to its considerable impacts on the estimated values. The number of input and output neurons corresponds to the number of input and output data, respectively (8 and 1 in this study). However, the most favorable numbers of hidden layers and neurons in each separate layer are entirely dependent on the complexity of the problem being solved by the neural network and there is no a straightforward scheme to obtain these parameters.

Hornik et al. showed that multilayer feed forward neural networks consist of only one hidden layer with enough number of hidden neurons can map any input to each output to an arbitrary degree of accuracy [67]. Based on the above, in the current study only networks with one hidden layer was examined. In order to find optimum number of neurons, different $2 - x - 1$ architectures ($x$ varies from 1 to 10) were investigated.
MSE and $R^2$ values evaluated for different neural network configurations of varying number of neurons in the hidden layer are shown in Fig. 5.

The best configuration for the smart system is normally determined based on the minimum value of MSE and maximum magnitude of $R^2$. As Fig. 5 depicts, the most appropriate ANN model consists of eight hidden neurons in one layer. Thus, it is found that a three-layer network with a $2-8-1$ different processing stages is the best structure. The final trained ANN model with 8 hidden neurons is shown in Fig. 3.

Table 2 presents the adjustable parameters for the proposed hybrid PSO–ANN approach.

4.2. LSSVM model

4.2.1. Data distribution (training and testing subsets)

Like ANNs, for the LSSVM based-model analyses, the database is broken down into training and testing subsets. The training

![Flowchart of PSO-based optimization algorithm for evolving the weights and biases of the constructed ANN.]

[Fig. 4. Flowchart of PSO-based optimization algorithm for evolving the weights and biases of the constructed ANN.]

[Fig. 5. Effect of number of hidden neurons on performance of the PSO–ANN model in terms of $R^2$ and MSE values.]
data are used for developing the model involves determining of parameters embedded in the LSSVM model. The testing data are used to measure the performance of the model and consequently to evaluate the precision and the effectiveness of the trained LSSVM model on data that play no role in building it. For the analysis, out of 120 data points, 96 data sets are taken for the training process and the rest of the data sets are used for testing the predictive capability of the model.

4.2.2. Kernel function

It is important to choose an appropriate kernel function. There are many types of kernels; however, three typical choices for the kernel function are:

- \( K(x, x_k) = x^T x_k \) (Linear kernel)
- \( K(x, x_k) = (\tau + x^T x_k)^d \) (Polynomial kernel of degree \( d \))
- \( K(x, x_k) = \exp(-x - x_k^2 / \sigma^2) \) (Radial basis function RBF kernel)

Entire magnitudes of \( \sigma \) and positive extents of \( \tau \) meet the Mercer condition [43]. There are many studies in the literature which have made comparisons between the most common kernels [45–47,49,68,48]. Provided nonlinear function prediction and nonlinear modeling cases, one can safely apply the RBF kernel as it is an effective way that can produce accurate results [45–47,49,68,48]. Based on the above, this study employed the radial basis function (RBF) as a kernel function.

4.2.3. Optimization method to tune the embedded parameters (\( \gamma \) and \( \sigma^2 \))

It is required to obtain the optimum values of kernel and regularization parameters through the training stage of the LSSVM technique. One effective way to conduct this task is linking the LSSVM model to genetic algorithm (GA) technique which is employed in this study. Indeed, GA which has been initially developed by Holland [69] is recognized as a strong evolutionary algorithm to optimize Kernel RBF parameter (\( \sigma^2 \)) and regularization parameter (\( \gamma \)). This tool can be properly designed to handle a variety of optimization problems in the context of biological processes. GA operates with a collection of candidate solutions, called a population.

The system guides the candidate solutions towards an optimum using the principles of evolution and natural genetics. As the search process advances, the population incorporates better and better individuals, and ultimately the convergence is attained, revealing that a single solution defeats it [69]. There are three key operators including mutation, selection, and crossover utilized at every generation of the GA model. It should be mentioned that a chromosome corresponds to a solution vector in GA. An extensive information on GA is found in the literature [65,61,69–72].

The flowchart of GA-based LSSVM incorporating to tune \( \gamma \) and \( \sigma^2 \) values is shown in Fig. 6. It should be noted that the hybridization of LSSVM method with GA had already given encouraging results for various applications [45–48]. The hybrid GA–LSSVM model searching mechanism for optimizing \( \gamma \) and \( \sigma^2 \) parameters is briefly described as follows.

a. Encoding and generating Initial population: Defining a chromosome or an array of variable values to be optimized. Herein, the chromosome has 2 variables (a two-dimensional optimization problem) given by (\( \gamma, \sigma^2 \)). Note that, a
chromosome is attributed to a unique solution in the solution space. Therefore, a mapping practice between the chromosomes and the solution space is needed. Encoding is another name for this mapping [69]. After representation of candidate solutions, an initial population of chromosomes was randomly generated.

b. **Fitness assignment**: Evaluate the fitness of each chromosome in the population using a fitness function. Herein, the mean squared error of all training data set was established as fitness function.

c. **Selection**: This stage involves repeating the most triumphant individuals available in a population with a rate which is proportional to their relative quality. Indeed, chromosomes with better fitness values have a greater chance of being selected than those having worse fitness values.

d. **Crossover**: Two different solutions are putrefied in this step and then the parts are randomly combined to create new solutions.

e. **Mutation**: The process disturbs a possible solution through a random way.

f. **Replace**: Use new generated population for the next generation.

g. **Stop criterion**: The procedure is continued unless a number of stopping conditions (e.g., when an acceptable solution has been found or the number of generations that the GA is allowed to execute) are satisfied.

According to the hybrid GA–LSSVM developed in the current study, the final magnitudes of \( g \) and \( \sigma^2 \) are determined to be 2,158,321.5215 and 22.210787768, respectively; implying high predictive potential of the model to determine the outputs with reliable and proper precision. Table 3 represents various settings in adjustment of the LSSVM model and GA optimization parameters.

### 5. Results and discussions

#### 5.1. LSSVM output results

Before launching the outcomes of the developed vector machine model for pseudo skin factor in rectangular drainage area, it is worth to mention that, distributions of the addressed target (pseudo skin factor) versus independent variables such as dimensionless length, \( L/2X_e \) and \( X_e/Y_e \) are demonstrated in Fig. 7. Fig. 7 depicts the variation of pseudo skin factor versus corresponding dimensionless length at different drainage area for \( L/2X_e = 0.2 \).

A regression plot between the estimated pseudo skin factor values and the real values are illustrated in Fig. 8. Fig. 8 demonstrates the scatter diagram that compares experimental pseudo skin factor against LSSVM approach solutions. A tight cloud of points about diagonal line \( (Y = X) \) for training and testing data sets present the high performance of the developed LSSVM approach.

The addressed figure depicts that superior agreement exist between the gained results of LSSVM model and the actual pseudo skin factor in rectangular drainage area. The statistical criteria of the evolved LSSVM approach which includes mean squared errors (MSE), average absolute relative deviations (AARD), and determination coefficients \( (R^2) \) are summarized to Table 4.

Fig. 9 depicts the contrast of the measured values of pseudo skin factor and gained outputs of the suggested low parameter model versus corresponding data index. As shown in Fig. 9, the model outputs are reliable because follow exactly the measured pseudo skin factor.

Fig. 10 illustrates the distribution for the deviation of gained outputs applying the LSSVM model versus experimental values of pseudo skin factor for all of the 120 data sets that implemented for developing the LSSVM approach. As demonstrated in Fig. 10, maximum relative deviations of obtained results from corresponding actual skin factor is about ±5%.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Basic parameter values of GA–LSSVM model for the prediction of pseudo-skin factor.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td><strong>Value/comment</strong></td>
</tr>
<tr>
<td>Input layer</td>
<td>2</td>
</tr>
<tr>
<td>Output layer</td>
<td>1</td>
</tr>
<tr>
<td>Kernel function</td>
<td>RBF kernel function</td>
</tr>
<tr>
<td>Number of data used for training</td>
<td>96</td>
</tr>
<tr>
<td>Number of data used for testing</td>
<td>24</td>
</tr>
<tr>
<td>GA Population size</td>
<td>1000</td>
</tr>
<tr>
<td>Max. number of generations</td>
<td>1000</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>0.8</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Fig. 7. Variation of skin factor versus dimensionless length \( (L_L) \) at different drainage area \( (X_e/Y_e) \) for \( L/2X_e = 0.2 \).

Fig. 8. Regression plot of the proposed vector machine model versus actual pseudo skin factor.
Moreover, Fig. 11 depicts the comparison between estimated skin factor by utilizing our LSSVM model and actual skin factor versus relevant dimensionless length when $L/2x_e = 1$. As illustrated in Fig. 11, the obtained results from our LSSVM model are matched greatly with corresponding real skin factor for this specific system. In addition, Fig. 12 demonstrates the comparison between estimated skin factor values by LSSVM model and actual skin factor in square drainage area ($X_e/Y_e = 1$) versus relevant dimensionless length ($L_D$). As shown in Fig. 12, LSSVM result’s followed real skin factor values with high level of precision and accuracy.

To determine relative importance of the used input variables on the desired target of this study that is pseudo skin factor, a rigorous statistical method has been implemented. The addressed technique named “analysis of variance (ANOVA)”. The obtained materials from the mentioned statistical method are exhibited in Fig. 13. As exhibited in Fig. 13, dimensionless length ($L_D$) has the highest positive effect on the pseudo skin factor of horizontal wells in the rectangular drainage area.

5.2. PSO–ANN output results

A regression plot between the estimated pseudo skin factor values and the real values are demonstrated in Fig. 14. This figure illustrates the scatter diagram that contrasts real pseudo skin factor against LSSVM approach solutions. A tight cloud of points about diagonal line ($Y = X$) for training and testing data sets present the high performance of the developed LSSVM approach. The addressed figure depicts that superior agreement exist between the gained results of LSSVM model and the actual pseudo skin factor in rectangular drainage area. Fig. 15 depicts the contrast of the measured values of pseudo skin factor and gained outputs of the suggested hybridized model versus corresponding data index. As shown in Fig. 15, the model outputs aren’t reliable in all boundaries because correlation coefficient between measured pseudo skin factor and PSO–ANN results is much lower than suggested LSSVM model.

Fig. 16 illustrates the distribution for the deviation of gained outputs applying the PSO–ANN model versus experimental values of pseudo skin factor for all of the 120 data sets that implemented for developing the PSO–ANN approach. As demonstrated in Fig. 16, maximum relative deviations of obtained results from corresponding actual skin factor is about ±30% that is not acceptable for engineering applications.

Moreover, Fig. 17 depicts the comparison between estimated skin factor by utilizing our PSO–ANN model and real skin factor versus relevant dimensionless length when $L/2x_e = 1$. As illustrated in Fig. 16, the obtained results from our PSO–ANN model aren’t matched precisely with relevant real skin factor for this

Table 4
Statistical parameters of the evolved LSSVM approach.

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th></th>
<th>Test set</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.9988</td>
<td></td>
<td>0.9991</td>
<td></td>
<td>0.9983</td>
</tr>
<tr>
<td>Average absolute relative deviation</td>
<td>0.77363</td>
<td></td>
<td>0.98169</td>
<td></td>
<td>0.81524</td>
</tr>
<tr>
<td>Mean square error</td>
<td>0.14151</td>
<td></td>
<td>0.1445</td>
<td></td>
<td>0.14213</td>
</tr>
<tr>
<td>$N$</td>
<td>96</td>
<td></td>
<td>24</td>
<td></td>
<td>120</td>
</tr>
</tbody>
</table>

Fig. 10. Relative error distribution of the obtained outputs from LSSVM model versus corresponding pseudo skin factor data points.

Fig. 11. Comparison between predicted skin factor values by LSSVM model and actual skin factor when $L/2x_e = 1$, versus relevant dimensionless length ($L_D$).
particular system. Furthermore, Fig. 18 depicts the comparison between estimated skin factor values by PSO–ANN model and real skin factor in square drainage area ($X_e/Y_e = 1$) versus corresponding dimensionless length ($L_D$).

As shown in Fig. 18, results of PSO–ANN model aren’t followed real skin factor values with adequate level of precision and accuracy. Finally, the comparison between statistical criteria of the proposed LSSVM model and PSO–ANN model that contains mean squared errors (MSE), average absolute relative deviations (AARD), and determination coefficients ($R^2$) are summarized to Table 5. As reported in Table 5, suggested LSSVM approach is robust and effective in contrast with PSO–ANN model in estimation pseudo skin factor of horizontal wells in rectangular drainage area.

5.3. Leverage approach

Determining of the outlier of the mathematical approaches plays a vital role in applicability of the addressed approach in the considered issue. Recognition of outlier means that the data, which are very different from the main points in the data bank, are identified. Hence, it seems vital to assess the real data existing for pseudo skin factor data. Since uncertainties influence

Fig. 12. Comparison between predicted skin factor values by LSSVM model and actual skin factor in square drainage area ($X_e/Y_e = 1$) versus relevant dimensionless length ($L_D$).

Fig. 13. Relative importance of each input variables on the pseudo skin factor in box shape drainage area.

Fig. 14. Regression plot of the proposed PSO–ANN model versus actual pseudo skin factor.

Fig. 15. Comparison between actual pseudo skin factor and predicted values by LSSVM model versus relevant data index.

Fig. 16. Relative error distribution of the obtained outputs from PSO–ANN model versus corresponding pseudo skin factor data points.
the estimation capability of the evolved approaches. To gain this
goal, the approach of Leverage Value Statistics has been carried
out. The Graphical discovery of the doubtful data (or outliers) is
conducted through theme of the Williams plot owing to the
determined H values from the gained outcomes.

A detailed explanation of mathematical backgrounds and
computational procedure of this approach can be found in pre-
vious literature. The Williams plot is demonstrated in Fig. 19 for
the results implementing the vector machine and PSO–ANN
approaches. It can be concluded that the developed technique
statistically demonstrates high performance in terms of accuracy
and applicability range as a large percentage of the data are
placed in the intervals $H \in [0, 0.149]$ and $R \in [-3, 3]$. In addition, it

the all data considered for the modeling approach are present within the acceptable ranges.

6. Conclusions

In this paper, an LSSVM model for predicting pseudo skin
factor of horizontal wells in the rectangular drainage area has
been developed. Through this work massive horizontal well
productivities data bank's [1, 21] gathered from literature sur-
veys are faced to the LSSVM model to evolve and examine its
robustness. Following deductions can be drawn based on solu-
tions obtained from the LSSVM and PSO–ANN models:

1. Without any doubts at all which having adequate qualitative
and quantitative expertise about pseudo skin factor as
fundamental factor which plays the leading role in productivity of the horizontal wells. In order to gain economic and technical optimizing goals, a numerous number of conventional studies have been done to propose some models for precise calculation of this parameter, but some intrinsic limitations and constraints have triggered attentions to be drawn towards numerical intelligent models.

2. Based on the employed statistical indexes the monitoring of the pseudo skin factor of horizontal wells made by developed LSSVM approach result the closest arrangement with the real data among the artificial neural network approaches.

3. The summary is that performing the evolved approach causes reaching a high degree of performance and precision in terms of calculating the pseudo skin factor of horizontal wells that has not already been conceivable via using conventional schemes.

4. The bottom line of this paper is that the LSSVM model for determining the pseudo skin factor of horizontal wells in box-shaped drainage area is easy-to-use model for petroleum engineers. Moreover, the aforementioned model can couple with commercial reservoir simulators to improve accuracy and decrease run time. Finally, the LSSVM model proposed in this paper can aim the petroleum engineers to determine optimum well location with the concept of reverse engineering.

Nomenclature

Abbreviations

AARD average absolute relative deviations
ANN artificial neural network
ANOV analysis of variance
BP back propagation
EOS equation of state
GA genetic algorithm
HGAPSO hybrid genetic algorithm and particle swarm optimization
ICA imperial competitive algorithm
LSSVM least square support vector machine
MAE mean absolute error (MAE)
MSE mean squared error (MSE)
PSO particle swarm optimization
R² coefficient of determination
RBF radial basis function
UPSO unified particle swarm optimization

Variables

\( \bar{y} \) the average of the predicted data
\( \bar{y}^t \) the average of the actual data
A drainage area of horizontal well, ft²
A extension of drainage volume of horizontal well in x-direction, ft
\( A_t \) area open to flow, ft²
b extension of drainage volume of horizontal well in y-direction, ft
\( b_j \) bias
\( B_o \) oil formation volume factor, rb/STB
C unit conversion factor
c cognition component
c2 social components
\( C_l \) geometric factor in Babu and Odeh’s model
c1 total compressibility, psi⁻¹
D non-Darcy flow coefficient, day/Mscf

e error = Actual – Model output
h reservoir thickness, ft
\( I_{ui} \) permeability anisotropy, dimensionless
J productivity index, STB/day/psi
\( K \) effective permeability, md
\( k_d \) effective permeability in damaged zone, md
\( k_h \) horizontal permeability, md
\( k_v \) vertical permeability, md
\( k_x \) permeability in x-direction, md
\( k_y \) permeability in y-direction, md
\( k_z \) permeability in z-direction, md
L horizontal wellbore length, ft
\( L_{1/2} \) half-length of horizontal wellbore, ft
\( L_D \) dimensionless length
N the total number of data points
\( o_j \) output
p reservoir pressure, psi
\( p_0 \) reference pressure, psi
\( p_{av} \) average reservoir pressure, psi
\( p_e \) reservoir pressure at boundary, psi
\( p_i \) initial pressure, psi
\( p_{wf} \) bottom-hole flowing pressure, psi
q0 oil production rate, STB/day
\( r_1^2 \) and \( r_2^2 \) two random numbers
\( r_{db1} \) radius of damaged zone in horizontal direction, ft
\( r_e \) wellbore radius, ft
s skin factor, dimensionless
\( S_j \) sum of interconnection weights
\( S_R \) partial penetration skin factor, dimensionless
t time, h
\( v_j \) velocity of particle i
\( W_{ii} \) interconnection weights in network model
\( x_0 \) x coordinate of center of well, ft
\( x_i \) position of particle i
\( x_e \) half of drainage area side, ft
\( y_i^f \) the ith output of the model
\( y_i^t \) the actual at the sampling point i
\( z_0 \) z coordinate of center of well, ft
\( z_w \) distance of wellbore from the lower boundary, ft

Greek letters

\( \mu \) viscosity, cp
\( \mu_o \) oil viscosity, cp
\( \phi \) porosity, dimensionless
\( \gamma \) regularization parameter
\( \delta \) absolute relative error
\( \rho \) density, lbm/ft³
\( \sigma^2 \) RBF parameter
\( \varphi \) activation function
\( \omega \) the inertia weight

References


