# Data-Driven Hydrocarbon Production Forecasting Using Machine Learning Techniques

**Article** *in* International Journal of Computer Science and Information Security, · June 2020

CITATIONS
9

READS
1,039

**1 author:**

Mohammad Salam
Southern University and A&M College
**20** PUBLICATIONS **33** CITATIONS

SEE PROFILE

# Data-Driven Hydrocarbon Production Forecasting Using Machine Learning Techniques

Masoud Safari Zanjani, Mohammad Abdus Salam, and Osman Kandara
Department of Computer Science, Southern University, Baton Rouge, Louisiana, USA

***Abstract*— Data analytics utilizes advanced statistical and machine learning methods to find the concealed information and trends present in different types of datasets. These methods have recently shown great potential to solve the problems in the oil and gas industry. The ability to find insights from large datasets can make an oil company more profitable and successful.**

**The innovation of sophisticated artificial intelligence methods as well as new developments in powerful high-speed computing resources have made the machine learning techniques more powerful than ever. The oil and gas industry has been benefited from these algorithms and machine learning techniques have been applied to many petroleum engineering challenges.**

**Artificial Neural Network (ANN), Linear Regression (LR), and Support Vector Regression (SVR) were employed in this research to forecast the daily oil production using the Volve oil field production dataset. All three methods show great potential for hydrocarbon production forecasting. Results for well NO159-F-1C, however, indicate that ANN had the best performance compared to the other two methods. This doesn't mean that ANN is the superior method compared to LR and SVR in every situation. The performance of an algorithm must be examined for each specific case in order to select the best technique.**

***Index Terms*—Machine Learning, Hydrocarbon production forecast, ANN, LR, SVR.**

## I. Introduction

Data analytics is an evolving area which involves utilizing advanced statistical and machine learning methods to find out the concealed information and trends present in different types of datasets. The ability to find insights from large datasets can make an oil company more profitable and successful. Driving meaningful information from available exploration and production data can help companies to lower costs and higher efficiency. Machine learning is a tool which helps data scientists to drive such meaningful insights from raw data.

Demand for employing of new mathematical and computational methods in oil and gas industry is more than ever. In today's highly competitive environment there is a never-ending race between companies to cut the cost and increase the production efficiency. Oil and gas industry, like many other industries, has taken advantage of the recent artificial intelligence advancements. New development in modern computers has enabled mathematical theories to be more powerful than before. These theories are providing engineers and scientists with tools that intelligent machines can be developed by applying them. Machine learning techniques, as an application of artificial intelligence, has been greatly employed in exploration, production, and management of hydrocarbons.

Machine learning techniques have recently attracted interests in many areas, including mathematics, healthcare, economics, and engineering, among many others. This is due to innovation of sophisticated artificial intelligence methods as well as new developments in powerful high-speed computing resources. The oil and gas industry has been vastly benefited from these improvements. Machine learning techniques have been applied to many petroleum engineering challenges such as well logs analysis [1], prediction of bubble point pressure of crude oils [2], hydrocarbon production forecast [3], characterization of reservoir heterogeneity [4], prediction of thermodynamic properties of reservoir fluids [5], forecasting crude oil viscosity and solution GOR [6], and ultimate recovery estimation [7]. Artificial intelligence applications apply unconventional ways to connect input data to output which attracted the interest of engineers and scientists. Machine learning techniques help us analyze and forecast hydrocarbon production in highly complex systems where understanding the physical mechanisms are complicated [3].

In the world of data science, data visualization is essential to analyze massive amounts of data and make data-driven decisions. Data visualization is the graphical representation of datasets. This representation could be in the form of charts, graphs, and maps. Data visualization helps us to see and find trends, outliers, and patterns in data. Correlation pairs-plots and heat-maps are used in this study to visualize data and understand the trends and correlations among data features.

In this research, production data from Volve field, an oil field on the Norwegian continental shelf (NCS) [8], was analyzed using machine learning methods. The operator Equinor together with the Volve license partners, ExxonMobil and Bayerngas, has disclosed all subsurface and operating data from this oil field [9]. Artificial Neural Networks (ANN), Linear Regression (LR), and Epsilon-Support Vector Regression (ε-SVR) are the utilized machine learning techniques to predict daily oil production from well head pressure and well head temperature as input features.

A Python code was developed to apply ANN, LR, and SVR methods to the production data of four different wells from Volve field. Prediction results, then, were compared to the real data values. All three methods show a great potential for hydrocarbon production forecasting. Results for well NO159-F-1C, however, indicate that ANN had the best performance compare to other two methods. This doesn't mean that ANN is the superior method compare to LR and SVR in every situation.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The performance of an algorithm must be examined for each specific case in order to determine the best technique.

## II. Methodology

Machine learning algorithms find patterns in data and generate insight to make better decisions [10]. Machine learning is a technique that gives the computers the ability to learn from examples and improve their performance. This technique empowers computers to have decision-making ability by using advanced mathematical algorithms. Machine learning is a part of Artificial Intelligence (AI) which gives the human-like behavior to machines.

Many problems in oil and gas industry are categorized as continuous value problems or in statistical terms "regression" problems. In regression analysis typically the effects of variable(s) are estimated on another variable. These problems usually consist of a dependent or a set of dependent variables and one or more independent variable.

### A. Linear Regression

Linear Regression (LR) estimates the relationship between one or more independent variables and a dependent variable by minimizing the sum of the squares in the difference between the observed and predicted values [11]. In Linear Regression the relationship between a dependent variable and one or more independent variables is modeled by fitting a linear equation. Linear Regression fits a linear model with coefficients $w = (w_1, \dots, w_p)$ that will best minimize the residual sum of squares between the observed responses in the dataset and the responses predicted by the linear approximation [12]. Mathematically it solves a problem of the form: $\min_w \|X_w - y\|^2$

### B. Support Vector Regression (SVR)

In ε-SVR, the objective is to find a function *f(x)* which has at most ε deviation from the actually obtained targets $y_i$ for all the training data and at the same time is as flat as possible [13]. This method relies on defining the loss function that ignores errors which are situated within a certain distance from true value. This distance is ε and this type of function is often called epsilon-intensive loss function [14].

The free parameters in SVR model are C and epsilon. C is the penalty parameter of the error term and epsilon is the acceptable deviation from real values. For predicted points within a distance epsilon from the actual values, no penalty will be associated in the training loss function [15].

### C. Support Vector Regression (SVR)

An Artificial Neural Network (ANN) is a mathematical algorithm that tries to simulate the functionalities and structure of biological brains where neurons are highly connected and data is processes by learning from repetition of events [16]. These systems are generally capable of learning, machine learning, to perform a task using examples and without being programmed with task specific rules. ANN include many connected processing units which work together to process data and generate meaningful information from it. ANN can be used for various data science problems such as classification, prediction, and pattern recognition.

The basic building block of an artificial neural network is an artificial neuron which is a simple mathematical function. This function has three sets of rules: multiplication, summation and activation [17]. The inputs are weighted at the entrance of artificial neuron which means that every input value is multiplied by a weight. In the middle section of artificial neuron is summation function which sums all weighted inputs and bias. At the end, the result of summation passes through an activation function which is also called transfer function (Fig. 1) [17].

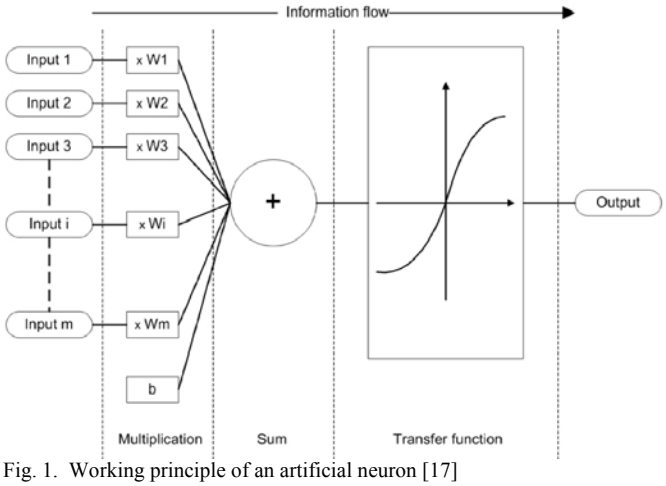



Fig. 1. Working principle of an artificial neuron [17]

Feedforward Neural Network is the simplest architecture of artificial Neural Network (Fig. 2). In this architecture, neurons (nodes) are arranged in layers and nodes from adjacent layers have connections (or edges) between them. These connections are weight associated.

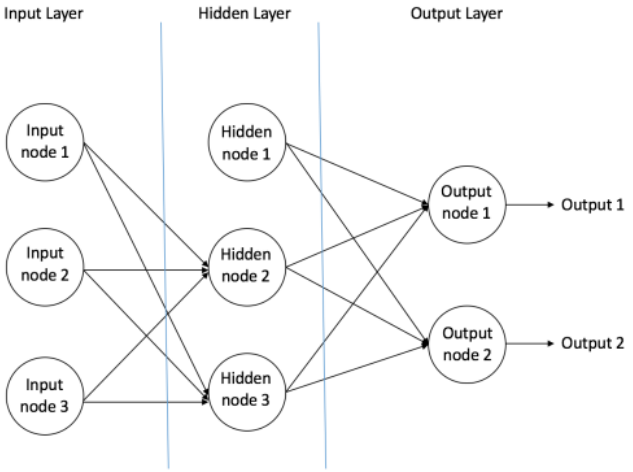



Fig. 2. An example of feedforward neural network [18].

A feedforward neural network has three layers;

1. Input layer: no computation happens in this layer and input nodes just pass the data to the hidden nodes.
2. Hidden layer: nodes in this layer perform computations and transfer information from the input nodes to the output nodes. A network can have multiple hidden layers.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3. Output layer: nodes in this layer are responsible for computations and transferring information to outside of the network.

In a feedforward network, information moves from input layer toward output layer, one direction, and there are no cycles or loops in the network.

## III. RESULTS AND ANALYSIS

Forecasting hydrocarbon production accurately is a significant step in management and planning of a petroleum reservoir. In this research, three machine learning methods are employed to forecast oil production. The results and discussions are presented in this chapter.

### A. Studied Petroleum Field

To build a successful model it is necessary to have a right dataset to train and test the model. Finding the right data with the right format is usually a challenge for machine learning methods. Right data correlates with the target that is going to be predicted [19]. To build a successful model, the problem needs to be precisely understood first.

Data from Volve field on the Norwegian continental shelf (NCS) was used for this study [9]. Datasets were released by the operator, Equinor, on May 2018 [8]. One of the specific goals of the data release was to allow students to train on real datasets. This dataset is the most comprehensive and complete dataset ever gathered on the NCS. It covers data in regards of production data, well design, completion string design, seismic data, well logs (petrophysical and drilling), geological and stratigraphical data, static and dynamic models, surface and grid data [8].

The production dataset has been used for this study. The studied wells are NO159-F-1C, NO159-F-5AH, NO159-F-14H, and NO159-F-15D. Each well has the data on data recording date, average downhole pressure, average downhole temperature, average drill pipe tubing, average annulus pressure, average choke size, average well-head pressure, average well-head temperature, oil volume, gas volume, water volume, type of flow (production or injection), and well type (oil production or water injection). The production data was recorded on a daily basis.

To measure the production data, different sensors are usually installed in downhole and well-head. The recorded data, then, are transferred to monitoring stations to be stored and analyzed. The process is shown in Fig. 3 schematically.

### B. Data Visualization

Data visualization is essential to analyze massive amounts of data and make data-driven decisions. Data visualization helps us to see and find trends, outliers, and patterns in data.

Each parameter was plotted versus time (day). Fig. 4 shows the average well-head pressure, average well-head temperature, and oil production volume for well NO159-F-1C. There are 427 readings, one data reading per day, for this well.

### C. Feature Selection

Feature selection is an important concept is machine learning problems. The data features are used to train the machine learning model and have a significant influence on the performance of the model. Feature selection and data cleansing should be the first step on building the model.

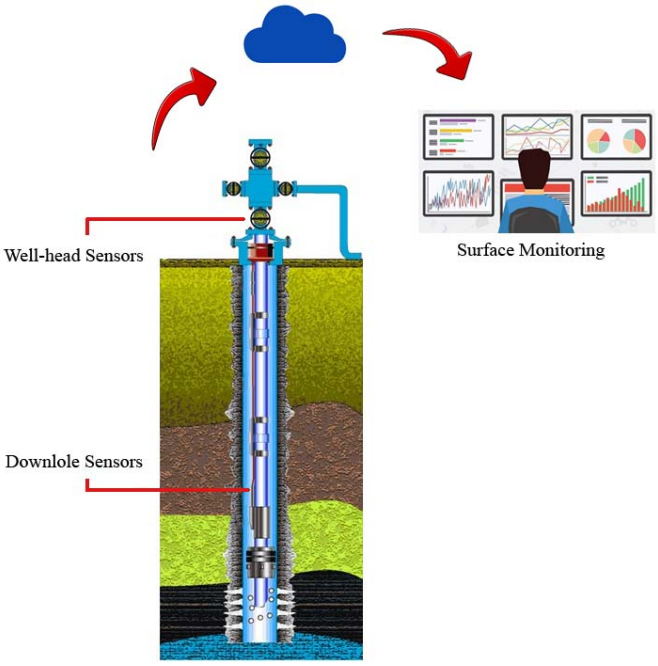



Fig. 3. Data recording process.

Feature selection is the process of choosing the features that contribute to the prediction value and are correlated to the desired output, either automatically or manually. Employing irrelevant features in the model may decrease the accuracy of the model and make the model learn from irrelevant parameters.

Highly correlated parameters, on the other hand, may have the shortcoming defect of not adding a new feature to the process of training the model [20]. Selecting highly correlated parameters as features may lead to reduction of model accuracy due to the lack of variation in the input data.

The correlation heat-maps were generated in this work to identify the highly correlated parameters. Pearson correlation coefficient is a measure of the linear correlation between two variables of X and Y. This coefficient has a value between +1 and -1, where +1 means total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

To make it more easily understandable visually, correlation results are presented in the form of heat-map. A heatmap is a graphical form of data representation which uses a system of color-coding to represent different values. The results for well NO159-F-1C is shown in Fig. 5. In this map a positive correlation is shown in red and a negative correlation is shown in blue.

As an example, in the results shown in Fig. 6 it can be seen that BORE_OIL_VOL and BORE_GAS_VOL are highly correlated; Pearson correlation coefficient of 0.99. It means that including both of these parameters will not add a new feature to the model and these parameters have basically linearly correlated data.

The objective of this study is to predict the values of daily oil production. Well-head pressure and well-head temperature are selected as training features. These parameters are not highly correlated, correlation coefficient of -0.42, which means they will not add redundant feature to the model.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

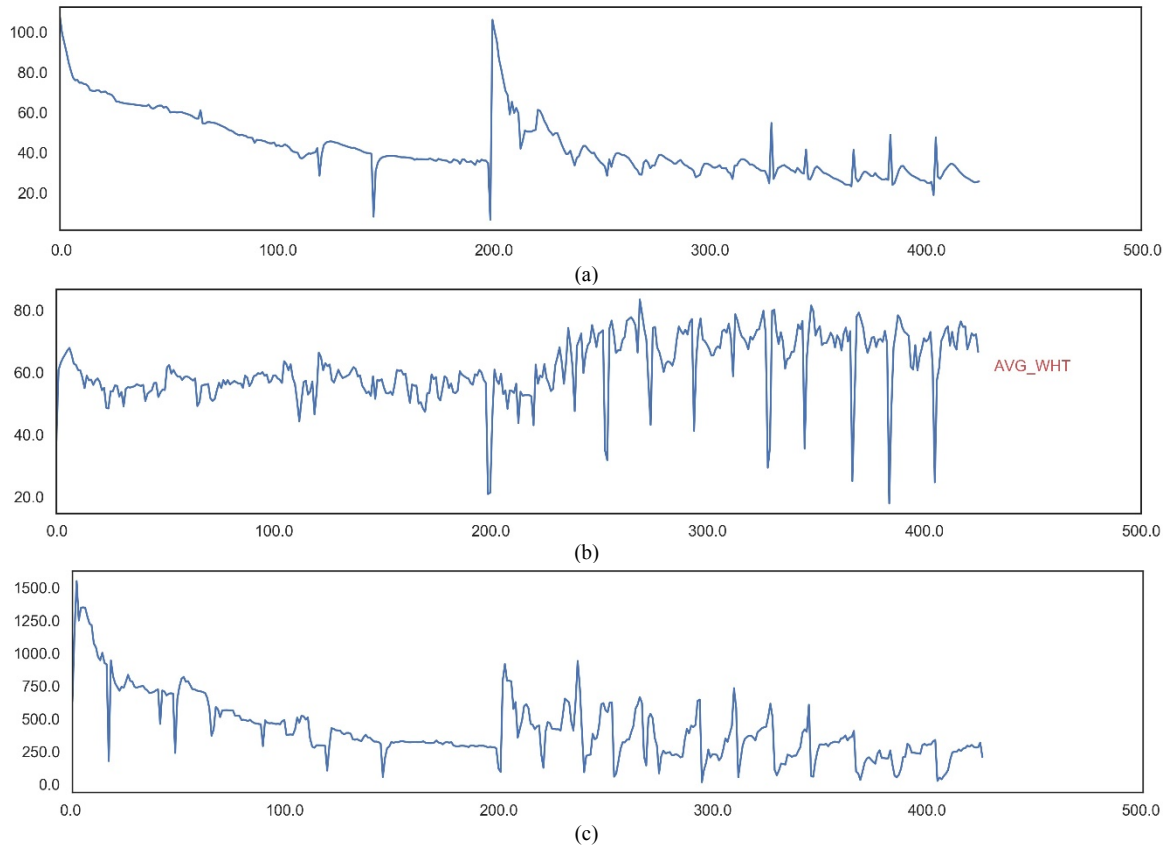



Fig. 4. Data plot versus time (day) for well NO159-F-1C; (a) average well-head pressure (b) average well-head temperature, (c) daily oil production volume.

### D. *Feature Scaling*

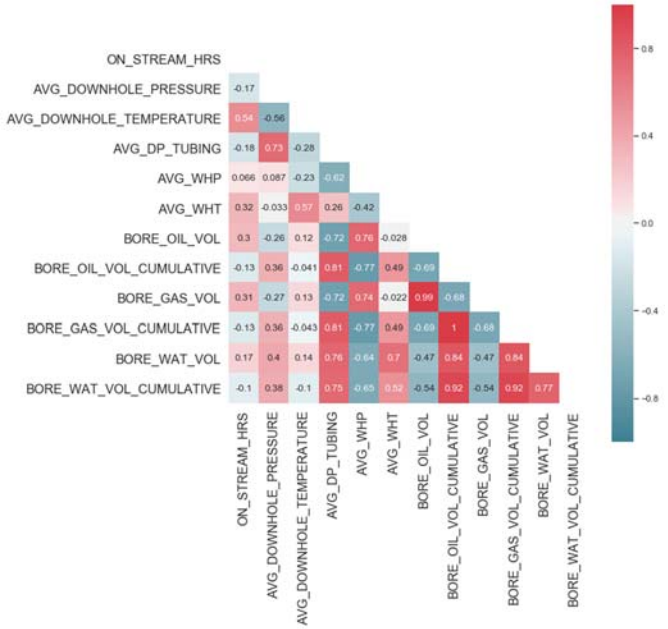



Fig. 5. Heat-map representation of correlation coefficient values between all data attributes for well NO159-F-1C.

Data values usually vary in range, magnitude, and units. This may cause a problem in machine learning algorithms as they use Euclidean distance between two data points in their computations. Feature scaling is used to normalize the range of independent variables or features of data to solve this problem.

The min-max scaling is used here to scale the range in [0,1]. The general formula for a min-max of [0,1] is given as:

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

### E. *Data Partitioning*

To build a machine learning algorithm, data is usually divided to two partition; Training dataset and Testing dataset. Training dataset is the part of the data that is used to train and fit the model. Test dataset, however, is used to evaluate the fitness of the model on the dataset. Data points in the training set are excluded from the test dataset.

In this research, 70 percent of data was used as training and remaining 30 percent was used as testing dataset. Both training and testing data subsets for the well NO159-F-1C have a pretty similar distribution compared to the total data distribution (Fig. 6). This guarantees an accurate model development procedure in the next phase of predictive model development.

### F. *Artificial Neural Network (ANN)*

The result of ANN prediction model is presented in this section. Artificial Neural Networks (ANN) are computational algorithms with the intention of simulating the behavior of biological brains. ANN prediction cross-plots of the total datasets, training dataset, and testing dataset are given in Fig. 8 for well NO159-F-1C. These graphs show predicted values for daily oil production versus the real values. The closer the points are to the 45º line (x = y), model has a better forecasting performance. Fig. 8, shows that a high number of data points falling along the 45º line, indicating a good agreement between predicted values and the real data values. Training data show a


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

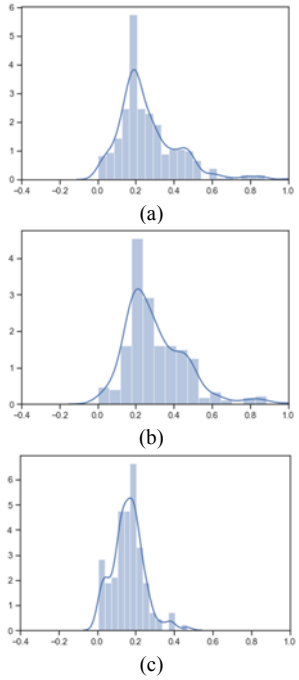


Fig. 6. Data distributions for well NO159-F-1C; (a) total dataset, (b) training dataset, and (c) testing dataset.

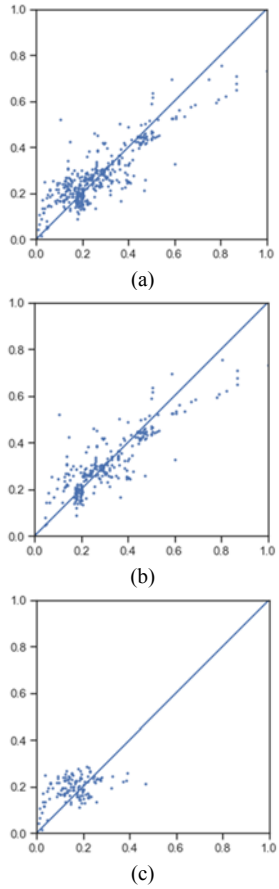


Fig. 7. ANN prediction cross-plots of (a) total datasets, (b) training dataset, and (c) testing dataset for well NO159-F-1C.

better agreement compared to testing dataset as the testing database is not presented to the model at the model building step.

A point-by-point comparison between the ANN predicted values versus the real data is shown in Fig. 8 for well NO159-F-1C. A good match can be observed between the forecasted values and real values.

### *G. Linear Regression (LR)*

In Linear Regression the relationship between a dependent variable and one or more independent variables is modeled by fitting a linear equation. LR prediction cross-plots of the total datasets, training dataset, and testing dataset are given in Fig. 10 for well NO159-F-1C. These graphs show predicted values for daily oil production versus the real values. Fig. 9, shows that a high number of data points falling along the 45º line, indicating a good agreement between predicted values and the real data values. Similar to ANN, training data show a better agreement compared to testing dataset as the testing database is not presented to the model at the model building step.

A point-by-point comparison between the LR predicted values versus the real data is shown in Fig. 10 for well NO159-F-1C. A good match can be observed between the forecasted values and real values.

### *H. Support Vector Regression (SVR)*

In ε-SVR, the objective is to find a function $f(x)$ which has at most ε deviation from the actually obtained targets $y_i$ for all the training data and at the same time is as flat as possible [13]. SVR prediction cross-plots of the total datasets, training dataset, and testing dataset are given in Fig. 11 for well NO159-F-1C. These graphs show predicted values for daily oil production versus the real values. As it can be seen by comparing Fig. 11 to Fig. 9 and 7, data points fall closer to the 45º line in ANN and LR methods which means ANN and LR had a better performance compared to SVR in predicting daily oil production for well NO159-F-1C.

A point-by-point comparison between the SVR predicted values versus the real data is shown in Fig. 12 for well NO159-F-1C. With comparison of this graph to Fig. 11 and 9, it can be seen that ANN and LR comparing to SVR, were able to predict the daily oil production more accurately.

### *I. Models Comparison*

It is usually a challenge to pick the right machine learning algorithm for a specific problem. There are several statistical and practical ways to compare different methods. It is not probably a good idea to just compare the overall accuracy as there are more indicators that need to be investigated depending on the specific application.

To compare the results of this work visually, the prediction values from all three methods are drawn in one graph (Fig. 13). In this graph the total real values of daily oil production are compared to the corresponding predicted values by ANN, LR, and SVR. As it can be seen, ANN had a better performance comparing to other two, followed by LR.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

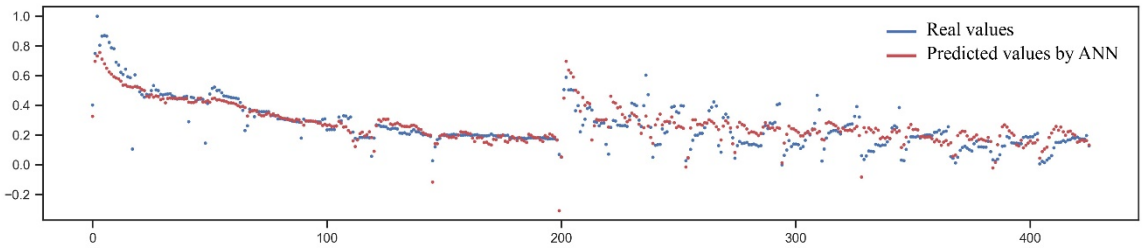
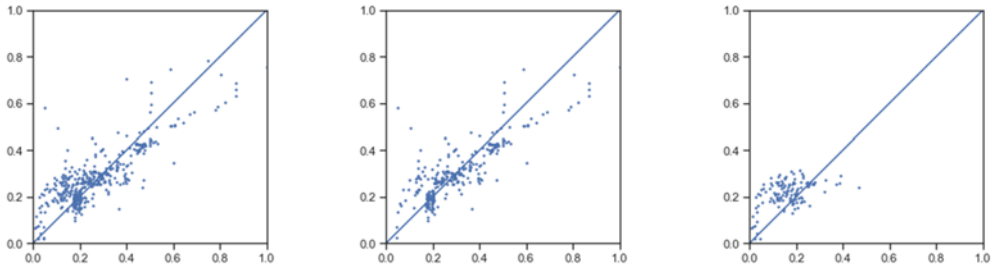



Fig. 8. Point-by-point comparison between the ANN predicted values versus the real data for well NO159-F-1C.



(a) (b) (c)

Fig. 9. LR prediction cross-plots of (a) total datasets, (b) training dataset, and (c) testing dataset for well NO159-F-1C.

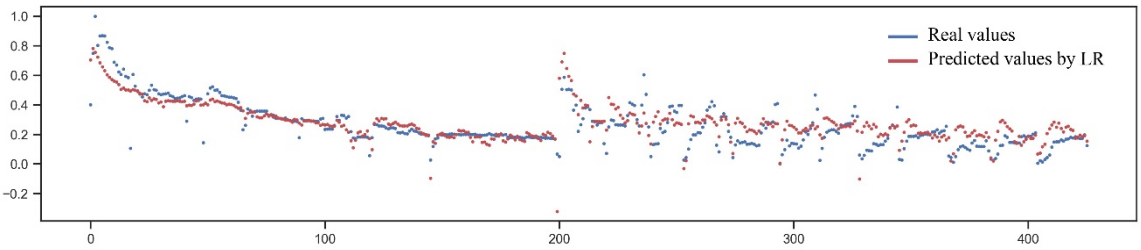
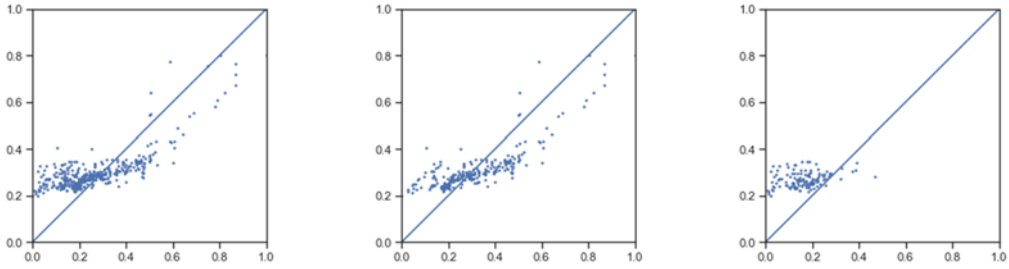



Fig. 10. Point-by-point comparison between the LR predicted values versus the real data for well NO159-F-1C.



(a) (b) (c)

Fig. 11. SVR prediction cross-plots of (a) total datasets, (b) training dataset, and (c) testing dataset for well NO159-F-1C.

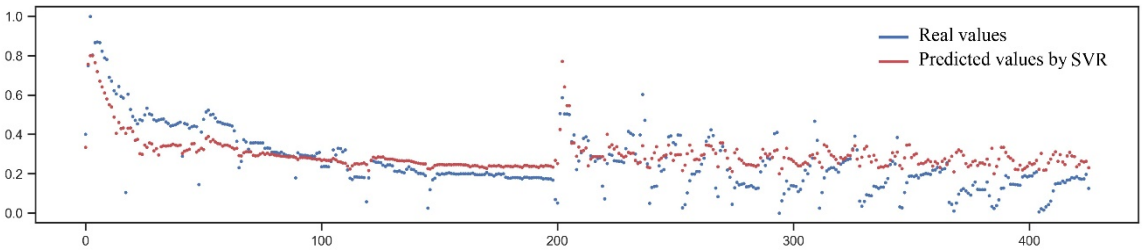



Fig. 12. Point-by-point comparison between the SVR predicted values versus the real data for well NO159-F-1C.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

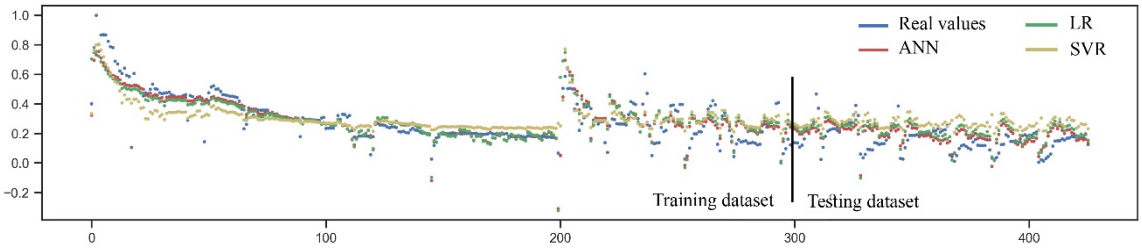



Fig. 13. Point-by-point comparison between the real data and predicted values by three methods of ANN, LR, and SVR for well NO159-F-1C.

As it was mentioned before, the objective of this project was to forecast the daily oil production of a petroleum field using machine learning techniques. To compare the performance of different utilized methods practically, the amount of the predicted oil production was calculated cumulatively. Cumulative produced oil is an important reservoir management parameter which shows the total amount of the oil that was produced from a specific well through a desired time period. To calculate this parameter, only the testing dataset was considered as it is new for the model and has not been used to train the model. The real values of the cumulative oil production were calculated using the testing portion of the real data and they have been compared to the predicted values in Fig. 14. As it can be seen here again, ANN shows the best performance and LR and SVR take the second and third place, respectively.

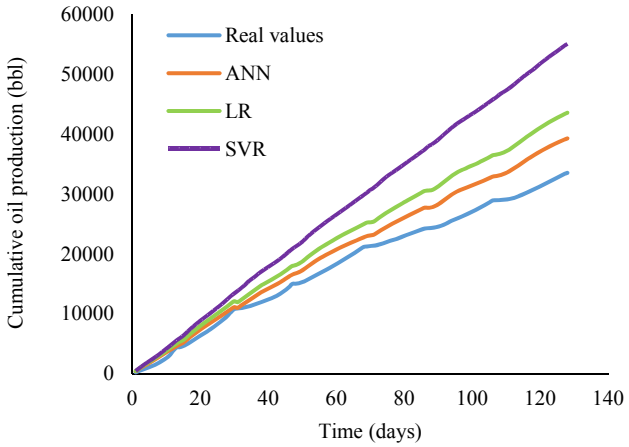



Fig. 14. Comparison of real cumulative oil production to the corresponding predicted values by three methods of ANN, LR, and SVR for well NO159-F-1C.

## IV. Conclusion

Oil production prediction is an important input for making decisions in an oil company. This parameter can be used for estimating reserves, optimizing production operations, business planning, and investment scenario evaluation. Production forecast is conventionally done by empirical equations. In the recent unconventional resources, however, production prediction is more challenging because of the extremely low permeability of the bed rock. Several data-driven techniques were studied here as a potential solution for oil production forecasting problem.

Artificial Neural Network (ANN), Linear Regression (LR), and Support Vector Regression (SVR) were the employed machine learning techniques to forecast the daily oil production. Prediction cross-plots and point-to-point comparison were presented for each method. To compare these three methods practically, the predicted cumulative oil production by each method was calculated and compared to the real cumulative oil production for testing dataset.

All three methods show a great potential for hydrocarbon production forecasting. Results for well NO159-F-1C, however, indicate that ANN had the best performance compare to other two methods. LR was more successfully predicted the production values compare to SVR. Although, ANN showed a better performance in this case study, it doesn't mean that it is the superior method compare to LR and SVR. Performance of machine learning methods depends on the studied dataset and the problem characteristics greatly and therefore; performance of an algorithm must be examined for each specific dataset and problem in order to select the best technique.

## References

[1] C. Zhou, X.-L. Wu and J.-A. Cheng, "Determining reservoir properties in reservoir studies using a fuzzy neural network," in *SPE Annual Technical Conference*, Houston, Texas, 1993.

[2] A. Hashemi-Fath, A. Pouranfard and P. Foroughizadeh , "Development of an artificial neural network model for prediction of bubble point pressure of crude oils," *Petroleum,* vol. 4, p. 281e291, 2018.

[3] P. Panja, R. Velasco, M. Pathak and M. Deo, "Application of artificial intelligence to forecast hydrocarbon production from shales," *Petroleum,* vol. 4, pp. 75-89, 2018.

[4] S. Mohaghegh, R. Arefi and S. Ameri, "A Methodological Approach For Reservoir Heterogeneity Characterization Using Artificial Neural Networks," in *SPE Annual Technical Conference & Exhibition*, New Orleans, 1994.

[5] J. Nagi, T. S. Kiong and S. K. Ahmed, "Prediction of PVT Properties In Crude Oil Systems Using Support Vector Machines," in *3rd International Conference on Energy and Environment*, Malacca, Malaysia, 2009.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[6] M. Oloso, A. Khoukhi , A. Abdulazeez and M. Elshafei, "Prediction of Crude Oil Viscosity and Gas/Oil Ratio Curves Using Recent Advances to Neural Networks," in *SPE/EAGE Reservoir Characterization and Simulation Conference*, Abu Dhabi, UAE, 2009.

[7] L. Jin, "Machine Learning Aided Production Data Analysis For Estimated Ultimate Recovery Forecasting," *M.S. thesis, Texas A&M University,* 2018.

[8] "Volve Data Village," Equinor, 18 10 2018. [Online]. Available: https://data.equinor.com/dataset/Volve. [Accessed 2019].

[9] "Disclosing all Volve data," Equinor, 14 6 2018. [Online]. Available: https://www.equinor.com/en/news/14jun2018-disclosing-volve-data.html. [Accessed 2019].

[10] "Introducing Machine Learning," MathWorks, 2019.

[11] "Ordinary Least Squares Regression," 04 10 2019. [Online]. Available: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/ordinary-least-squares-regression. [Accessed 2019].

[12] scikit-learn, "Generalized Linear Models," scikit-learn developers, 2007. [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html. [Accessed 2019].

[13] A. J. Smola and B. Scholkopf, "A Tutorial on Support Vector Regression," ESPRIT Working Group in Neural and Computational Learning, 1998.

[14] "Support Vector Machine Regression," [Online]. Available: http://kernelsvm.tripod.com/. [Accessed 2019].

[15] scikit.learn, "sklearn.svm.SVR," scikit-learn developers, 2007. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html. [Accessed 2019].

[16] S. e. Z. Lashari, A. Takbiri-Borujeni, E. Fathi, T. Sun, R. Rahmani and M. Khazaeli, "Drilling performance monitoring and optimization: a data-driven approach," *Journal of Petroleum Exploration and Production Technology,* 2019.

[17] A. Krenker, J. Bešter and A. Kos, "Introduction to the Artificial Neural Networks," in *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, InTech, 2011, p. 362.

[18] ujjwalkarn, "A Quick Introduction to Neural Networks," 9 8 2016. [Online]. Available: https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/. [Accessed 2019].

[19] C. Nicholson, "A.I. Wiki," Skymind, [Online]. Available: https://skymind.ai/wiki/datasets-ml. [Accessed 2019].

[20] M. Kuhn and K. Johnson, Applied Predictive Modeling, New York: Springer Science+Business Media, 2013.