A Comparative Study on Different Machine Learning Algorithms for Petroleum Production Forecasting

Lan Mai-Cao*, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam and **Hoa Truong-Khac**, PetroVietnam Exploration & Production Corporation – Integrated Technical Center (PVEP-ITC), Ho Chi Minh City, Vietnam

Abstract

In the recent years, machine learning and its subset, deep learning, have been quickly developed and applied with great success in various areas of petroleum engineering. Different machine learning algorithms for petroleum production forecast are studied in this work for efficiency comparison purpose. Historical production data from an oil well currently producing in the oilfield X, Southern Vietnam has been first pre-processed to construct eight different predictive models for production forecast on the oil well under consideration. The algorithms under consideration in this work are: (1) the classical machine learning algorithms, including random forest, gradient boosting, k-nearest neighbor, support vector regression; and (2) the deep learning algorithms, including multilayer perceptron, convolutional neural network, long short-term memory, and gated recurrent unit. The results from this comparative study show that in spite of their simplicity, some classical machine learning algorithms, especially the support vector regression shows its high efficiency in performing the prediction tests. In addition, it can be found from this work that pre-processing of the historical production data is crucial to the success of the application of artificial neural networks to production forecasting.

Introduction

Production forecast is an important task to perform economic evaluation and production optimization for oil and gas reservoirs. One of the most popular methods for production forecasting is decline curve analysis (DCA) thanks to its simplicity with the empirical observation of production decline and the basic assumption about the preserved trend of the rate decline curve in the future (Arps 1945). Since the physics associated with the hydrocarbon recovery process are not fully captured by DCA, prediction results of the method are usually not robust (Satter and Iqbal 2016). A more comprehensive approach to oil and gas production prediction is reservoir simulation. Although this approach yields more reliable prediction results, it requires much more data and effort as well as advanced domain knowledge for a valid reservoir model (Islam et al. 2016).

With the rapid development of computing technology and data analytics, machine learning (ML) and its special subset, deep learning (DL), have emerged as an advancement of data-driven approach based on artificial intelligence (Yucel et al 2020). In principle, the ML approach is well suited for problems which require advanced domain knowledge that is hard to gain but for which plenty of observed data is available. In recent years, various ML and DL applications have been applied with great success in different areas of

DOI:10.14800/IOGR.1205

Received May 10, 2022; revised June 20, 2022; accepted June 27, 2022.

* Corresponding author: maicaolan@hcmut.edu.vn

petroleum engineering such as oil production optimization (Shirangi 2012), drilling hydraulics prediction and optimization (Wang and Saeed 2015), ANN-based screening tool for CO_2 injection in naturally fractured reservoirs (Hamam and Ertekin 2018), interpretation of flow-rate, pressure and temperature data (Tian and Horne 2019), monitoring oi production rate (Khan et al. 2019), reservoir characterization and modeling (Lan and Le 2019; Lan and Hoa 2021), enhancement of production forecast (Doan and Vo 2021), to name just a few.

The objective of this work is to examine the ability of different machine learning algorithms to predict oil and gas production from historical data of a well currently producing oil from the field X (*due to the operator's requirements, the real name of the field of interest is not shared*) in Southern Vietnam. In particular, four classical machine learning algorithms, including random forest (RF), gradient boosting (GB), k-nearest neighbor (k-NN), and support vector regression (SVR) together with four advanced models, namely multilayer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), and gated recurrent unit (GRU) have been studied for efficiency comparison.

The remaining parts of this paper is organized as follows. Section 2 briefly presents the background and workflow to implement machine learning algorithms for petroleum production forecasting. Section 3 presents and discusses the results from this comparative study. Section 4 summarizes the main points of this study along with some concluding remarks on the performance of the different algorithms under consideration.

Methodology

This section briefly presents the background of the machine learning algorithms used in this work. The classical ML algorithms under consideration include:

- k-nearest neighbors, k-NN (Cover and Hart 1967)
- Random forest, RF (Breiman 2001)
- Gradient boosting, GB (Friedman 2001)
- Support vector regression, SVR (Drucker et al. 1997)

The more advanced algorithms associated with the following deep neural networks are also studied in this work:

- Multilayer perceptron, MLP (Cheng and Titterington 1994)
- Convolutional neural network, CNN (LeCun 1989)
- Long short-term memory, LSTM (Hochreiter and Schmidhuber 1997)
- Gated recurrent unit, GRU (Chung et al. 2014)

The workflow for our comparative study on aforementioned algorithms in this study consists of the following steps:

Step 1: Data preparation

- Outlier detection: This task is to properly detect and handle outliers to ensure that our data is statistically significant.
- Data scaling: All data values are scaled into the fixed range to prevent the learning algorithms from being biased to greater magnitudes of the data, especially in those algorithms that leverage similarities between samples such as k-NN, SVR, etc. …
- Cross validation: The main concern in regression is the ability of the trained model to perform on unseen data. For a valid performance evaluation, the dataset is splitted into training, validating and testing subsets. In addition, the testing set is not used for model construction.
- Autoregression modeling: For time series data, the future value of a variable, y_t, can be predicted using a linear combination of its past values as follows (Paolella 2019).

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \epsilon_t \quad (1)$$

where β_i (i=0, 1, 2 ... p) are regression constants, y_{t-p} is the value of y at time (t-p) in the past, and ϵ_t is the noise.

Step 2: Model construction

In this step, several classical ML and DL networks are constructed with the training dataset through which the model parameters are tuned to minimize the loss functions. The validation set is used to adjust the hyper-parameters of the network during the training process.

Step 3: Model testing

The trained models are tested in this step with the blind test dataset for performance evaluation. Different performance metrics are calculated including the mean squared error (MSE), mean absolute error (MAE).

Results and Discussion

Three test cases have been performed with different classical machine learning and deep learning algorithms in this work. The objective of these test cases is to examine the performance efficiency of those algorithms in production forecasting on an oil well currently producing from the oilfield X in Southern Vietnam. The historical production data includes the flow rates of oil, gas, water and the bottom-hole pressures collected at more than a thousand points in time.

***Test Case 1**:*

Four classical machine learning algorithms including k-NN, RF, GB, SVR have been studied in this test case. All models have one single output which is the bottom-hole pressure difference between a particular time of interest and the initial time (ΔP). As can be seen from **Figure 1**, all models fit quite well with the training dataset (the left figure). For the test dataset, however, some algorithms including RF and GB provide the predicted values with high deviation from the measured data (the right figure). Among the others, SVR yields excellent results that match very well with the observed data throughout the full time range of interest. Especially, SVR shows its ability to capture stiff changes of pressure occurring in the middle of the testing period.

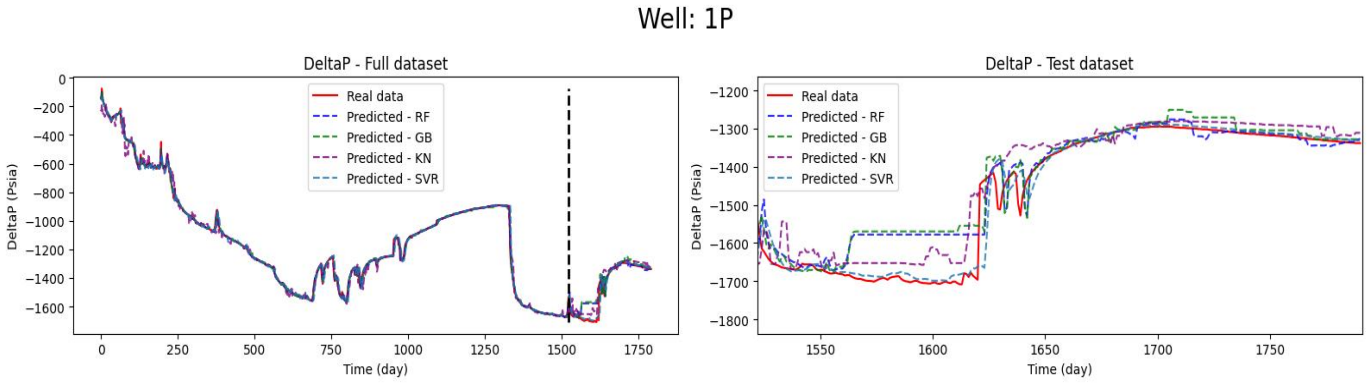
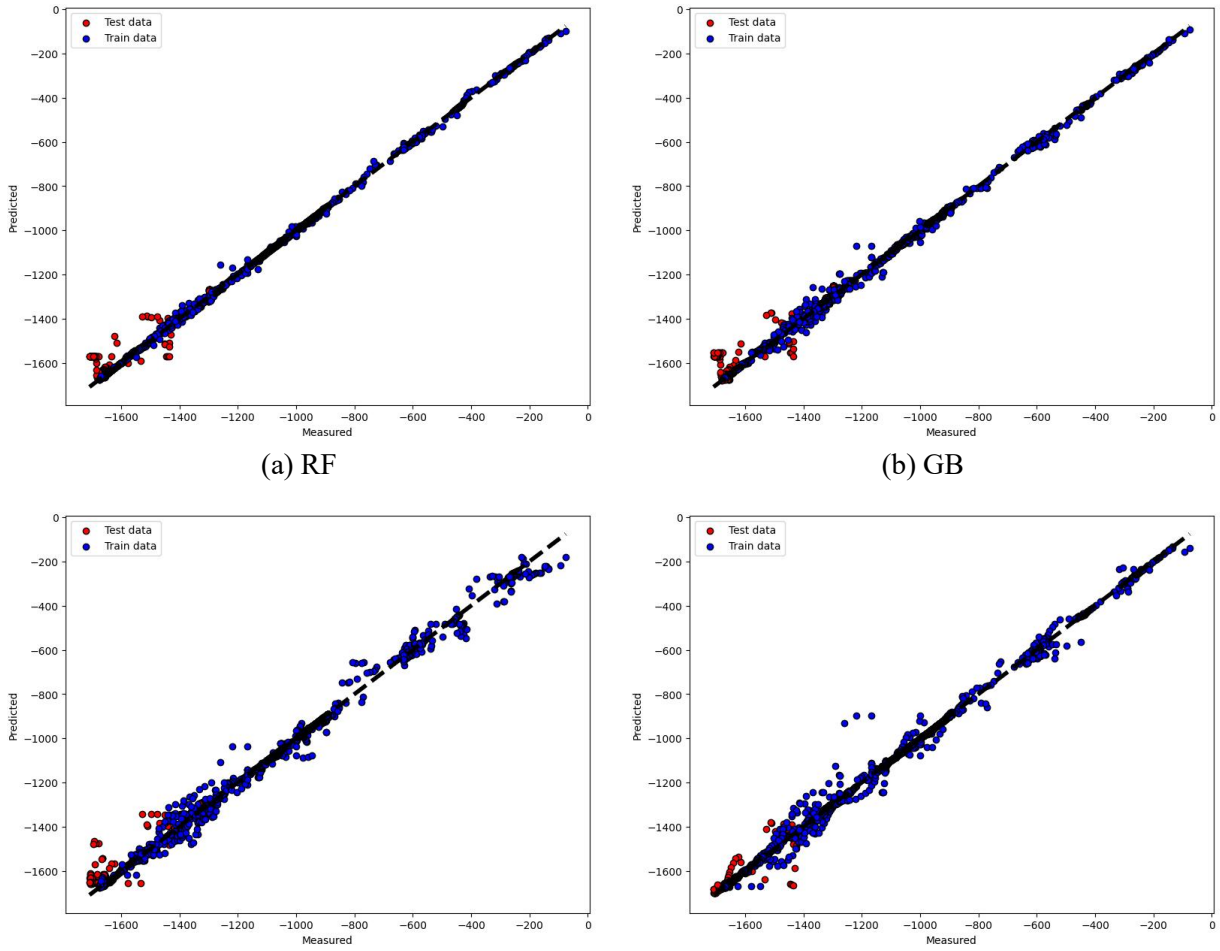


Figure 1—The results in the full time range (Left) ;The results in the test dataset interval (Right). The vertical dash line represents the time where the dataset is splitted into 2 subsets: the training dataset is located to the left of the line whereas the test dataset is located to its right.



(a) RF (b) GB

(c) k-NN (d) SVR

Figure 2—Cross plots of the bottom-hole pressure from the four classical ML algorithms (Test Case 1). (a) Random Forest; (b) Gradient Boosting; (c) k-Nearest Neighbor; (d) Support Vector Regression.

The cross plots in **Figure 2** show how the predicted bottom-hole pressures deviate from between the measured and predicted values. As can be seen from the figure, the four algorithms under consideration yield very good match with the training dataset.

The performance metrics of the four classical ML algorithms are shown in **Table 1** for Test Case 1. As can be seen from **Table 1**, all algorithms yield quite good performance metrics, and the SVR algorithm shows its superior with the highest R^2 and lowest errors.

Table 1—Performance metrics of the four classical ML algorithms (Test Case 1)

Algorithm/Model	MSE (psia)	MAE (psia)	R^2
Random Forest (RF)	3894.56	40.11	0.87
Gradient Boosting (GB)	4500.84	44.68	0.85
k-Nearest Neighbor (k-NN)	2612.07	37.13	0.91
Support Vector Regression (SVR)	1194.33	15.84	0.96

***Test Case 2**:*

The objective of this test case is to check the model ability to perform time-series forecast with multiple network outputs. The same classical machine learning algorithms, i.e. RF, GB, k-NN, SVR have been studied in this test case with the two outputs which are the bottom-hole pressure and oil rate.

Compared to the single output case, good matches can also be observed in this test case with the training dataset whereas further deviation from the measured data can be noticed, especially with the predicted oil rate. **Figure 3** show that among the others, the SVR again is the most efficient algorithm for time-series forecasting. While the other algorithms tend to yield big errors at some points in time when high peaks occur, SVR shows its ability to capture stiff changes in bottom-hole pressure and oil rate.

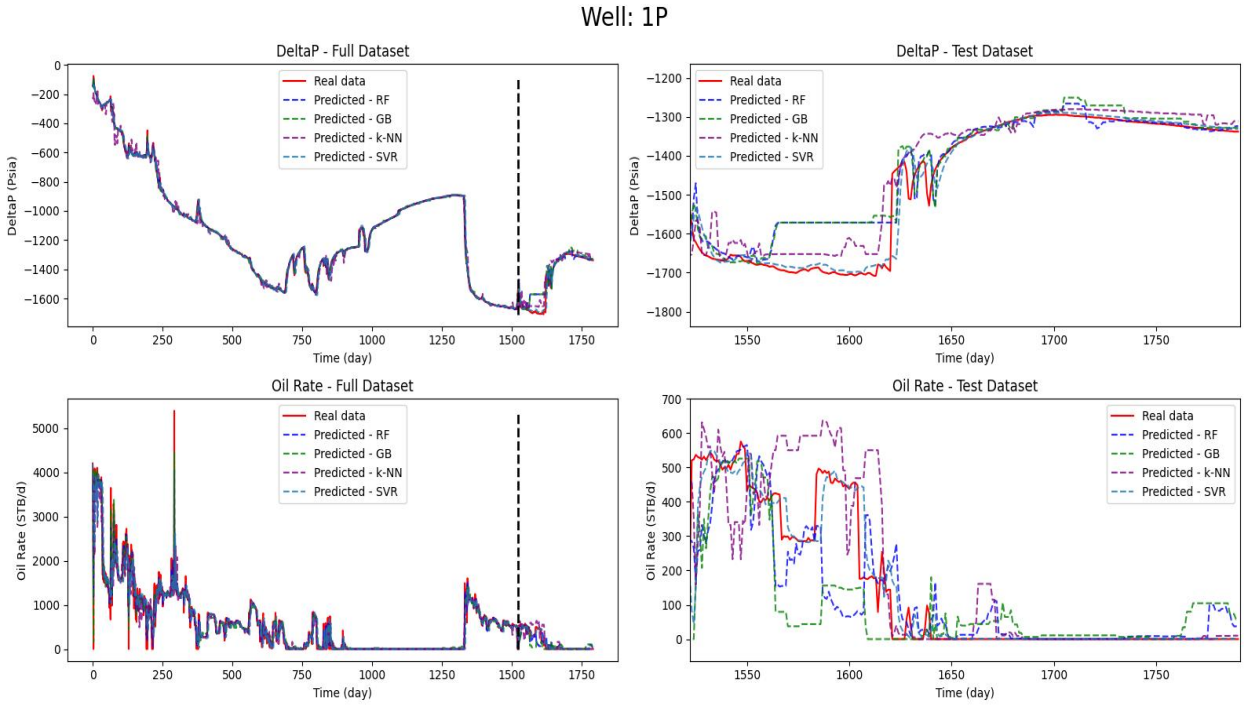


Figure 3—Test Case 2 with Classical ML Algorithms. ΔP vs Time - Full Dataset (Top Left); ΔP vs Time - Test Dataset (Top Right); Oil Rate vs Time – Full Dataset (Bottom Left); Oil Rate vs Time – Test Dataset (Bottom Right).

It can be seen from Figure 3 that the predicted values by the models with multiple outputs are not as good as those having single output. Since the loss function is formulated as the weighted average of the component losses, the loss function evaluation can be adjusted with the weight values. In principle, small weight is used for data with high uncertainty. Since the bottom-hole pressure is measured by a permanent downhole gauge with high reliability in our case, its weight is greater than that of the oil rate.

***Test Case 3**:*

Four deep learning models including Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been studied in this test case. All models have two outputs, the bottom-hole pressure difference (ΔP) and the oil rate of the well under consideration.

As can be seen from **Figure 4**, the predicted values are in good agreement with the measured data for most of the cases in a global sense. For the test dataset, however, deviations of the predicted values from the measured data are observed. In this test case, the predicted pressures are not much different among the four deep neural network models. On the other hand, greater oil rate differences between the models are observed.

It should be noted from Figure 4 that the deep neural networks studied in this test case do not show their ability to capture stiff changes, such as the two peaks in the pressure profile (the top-right figure) and some drop down in the oil rate profile (the bottom-right figure).

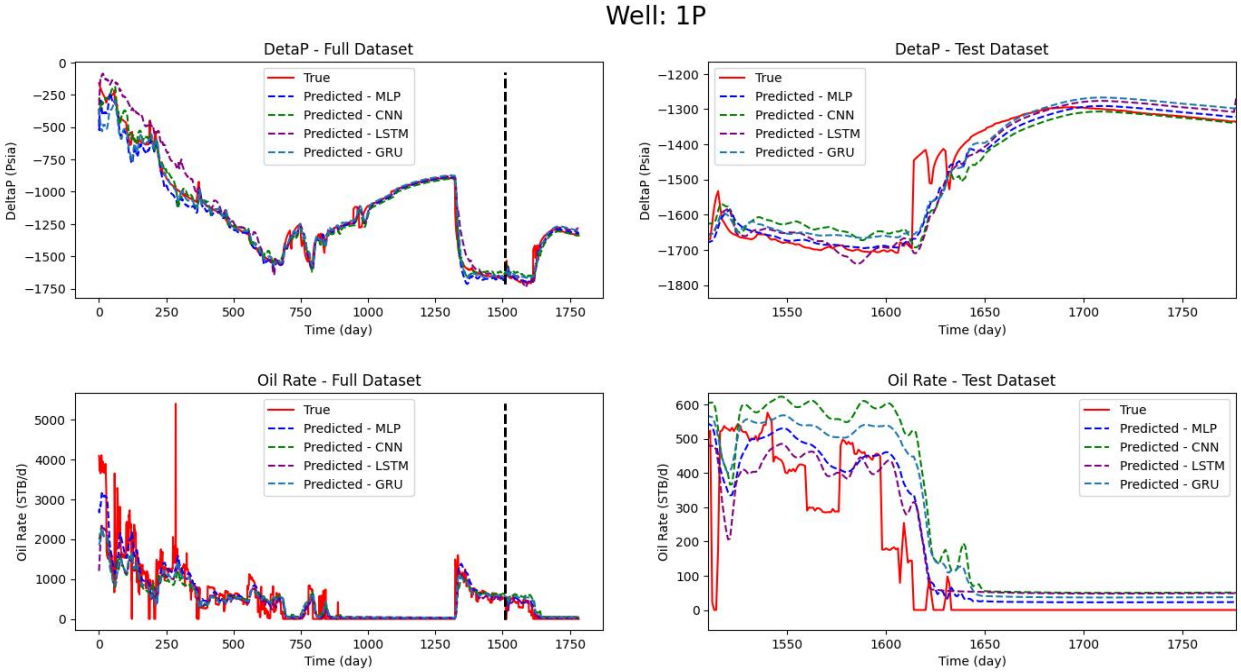


Figure 4—Test Case 3 with Deep Learning Algorithms: ΔP vs Time - Full Dataset (Top Left); ΔP vs Time - Test Dataset (Top Right); Oil Rate vs Time – Full Dataset (Bottom Left); Oil Rate vs Time – Test Dataset (Bottom Right).

Table 2—Performance metrics for the four deep neural networks (Test Case 3)

Algorithm/Model	MSE (psia)	MAE (psia)	R^2
Multilayer Perceptron (MLP)	2064.24	25.93	0.93
Convolutional Neural Network (CNN)	3620.19	41.63	0.88
Long Short-Term Memory (LSTM)	2914.20	33.21	0.90
Gated Recurrent Unit (GRU)	2481.41	37.29	0.92

Conclusions

This paper reports a comparative study on different machine learning algorithms for petroleum production forecasting. In particular, four classical machine learning models (k-Nearest Neighbor, Random Forest, Gradient Boosting, Support Vector Regression) and four deep neural networks (Multilayer Perceptron, Convolutional Neural Network, Long Short-Term Memory, and Gated Recurrent Unit) have been studied for their ability to perform prediction on a time-series data.

The results from this work show that unlike in image processing, voice recognition or object detection where complicated deep neural networks proved to be highly powerful, the classical machine learning approach with the Support Machine Regression is computationally efficient with good performance metrics and short computation time in all test cases performed in this work for petroleum production forecasting from historical production data of the oil well of interest.

Acknowledgement

We would like to thank Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for the support of time and facilities for this study.

Conflicting Interests

The authors declare that they have no conflicting interests.

Nomenclature

CNN: Convolutional Neural Network

DCA: Decline Curve Analysis

DL: Deep Learning

GB: Gradient Boosting

GRU: Gated Recurrent Unit

k-NN: k-Nearest Neighbor

LSTM: Long Short-Term Memory

ML: Machine Learning

RF: Random Forest

RNN: Recurrent Neural Network

SVR: Support Vector Regression

References

Arps, J.J. 1945. Analysis of Decline Curves. *Transactions of the AIME* **160**(1): 228–247.

Breiman, L. 2001. Random Forests. *Machine Learning* **45**(1): 5–32.

Cheng, B. and Titterington, D.M. 1994. Neural Networks: A Review from a Statistical Perspective. *Statistical Science* **9**(1): 2–54.

Chung, J., Gulcehre, C., Cho, K., et al. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv Preprint*: 1–9.

Cover, T.M. and Hart, P.E. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* **13**(1): 21–27.

Doan, T. and Vo, M.V. 2021. Using Machine Learning Techniques for Enhancing Production Forecast in North Malay Basin. *Improved Oil and Gas Recovery* **5**(1): 1–6.

Drucker, H., Burges, C.J.C. , Kaufman, L., et al. 1997. Support Vector Regression Machines. *Advances in Neural Information Processing Systems* **1**: 155–61.

Friedman, J.H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29**(5): 1189–1232.

Hamam, H. and Ertekin, T. 2018. A Generalized Varying Oil Compositions and Relative Permeability Screening Tool for Continuous Carbon Dioxide Injection in Naturally Fractured Reservoirs. Paper SPE-192194-MS presented at the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, 23-26 April.

Hochreiter, S. and Schmidhuber, S. 1997. Long Short-Term Memory. *Neural Computation* **9**(8): 1735–80.

Islam, M.R., Mousavizadegan, S.H., Mustafiz, S., et al. 2016. *Advanced Petroleum Reservoir Simulation*, second edition. Beverly, MA, USA: Scivener Publishing.

Khan, M.R, Alnuaim, S., Tariq, Z., et al. 2019. Machine Learning Application for Oil Rate Prediction in Artificial Gas Lift Wells. Paper SPE-194713-MS presented at the SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 21-23 March.

LeCun, Y. 1989. Generalization and Network Design Strategies. Paper presented at the Connectionism in Perspective, Elsevier.

Lan, M.C. and Le, C. 2019. A Self-Organizing Map, Machine Learning Approach to Lithofacies Classification. *International Journal of Simulation: Systems, Science & Technology* **19**(3): 1-16.

Lan, M.C. and Hoa, T.K. 2021. Reservoir Property Modeling with Seismic Attributes and Artificial Neural Network. *Science & Technology Development Journal-Engineering and Technology* **4**(3): 61-69 (Vietnamese).

Paolella, M.S. 2019. *Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH*. New York City: John Wiley & Sons Inc.

Satter, A. and Iqbal, G.M. 2016. Decline Curve Analysis for Conventional and Unconventional Reservoirs. *Reservoir Engineering*: 211–232.

Shirangi, M.G. 2012. Applying Machine Learning Algorithms to Oil Reservoir Production Optimization. Paper presented at the Stanford Machine Learning Conference, Stanford, California, USA, 12 December.

Tian, C. and Horne, R.H. 2019. Applying Machine-Learning Techniques to Interpret Flow-Rate, Pressure, and Temperature Data from Permanent Downhole Gauges. *SPE Reservoir Evaluation and Engineering* **22**(2): 386–401. SPE-174034-PA.

Wang, Y. and Saeed S. 2015. Drilling Hydraulics Optimization Using Neural Networks. Paper SPE-173420-MS presented at the SPE Digital Energy Conference and Exhibition, The Woodlands, Texas, USA, 3-5 March.

Yucel, M., Bekda, G., Nigdeli, S.M. 2020. Review and Applications of Machine Learning and Artificial Intelligence in Engineering: Overview for Machine Learning and AI. In *Artificial Intelligence and Machine Learning Applications in Civil, Mechanical, and Industrial Engineering,* ed. Bekdas, G., Nigdeli, S.M., and Yucel M., Chap. 1, 1-12. Hershey, Pennsylvania, USA: IGI Global.

Mai-Cao Lan received B.E. degree (1991) in Mechanical Engineering from HCMUT, Vietnam, M.E. degree (1998) in Systems Engineering from Royal Melbourne Institute of Technology, Australia, and Ph.D. degree (2009) in Computational Mechanics from the University of Southern Queensland, Australia. Currently, Dr. Lan is the Head of the Department of Drilling & Production Engineering, Faculty of Geology and Petroleum Engineering, Ho Chi Minh City University of Technology, Vietnam. His current research mainly focuses on enhanced/improved oil recovery (EOR/IOR), integrated production modeling, flow assurance and machine learning applications in petroleum industry.

Truong-Khac Hoa graduated with a bachelor's degree from the University of Natural Sciences, Ho Chi Minh City, majoring in Applied Mathematics in 2002. Completed the Master's program at Ho Chi Minh City University of Technology, majoring in Petroleum Engineering in July 2019. MSc. Truong Khac Hoa is currently working at the Integrated Technical Center of PetroVietnam Exploration & Production Corporation - PVEP.