12th IFAC International Symposium on
Advanced Control of Chemical Processes
July 14-17, 2024. Toronto, Canada

Optimizing GOR Prediction in Oil Wells: Efficacy of Convolutional Neural Networks with Hybrid Data Integration ⋆

Juan F. de Amorim, Jean Panaioti Jordanou, Ubirajara F. Moreno *
Júlio Elias Normey Rico * **Bruno Ferreira Vieira** **

* *Departamento de Engenharia de Automação & Sistemas, Universidade Federal de Santa Catarina*

e-mails: juan.flor.amorim@gmail.com; jeanpanaioti@gmail.com; ubirajara.f.moreno@ufsc.br ** *CENPES-PETROBRAS, (e-mail: bfv@petrobras.com.br).*

Abstract: The accurate prediction of the Gas-Oil Ratio (GOR) is crucial in optimizing oil production and ensuring the longevity of oil wells. This study evaluates the effectiveness of Neural Networks in predicting GOR for oil wells using gas lift production. We demonstrate that Neural Networks, particularly when trained with a mix of real and simulated data, show great promise for precise and rapid GOR predictions.
To validate our approach, we analyzed data from real-world oil wells, employing Neural Networks for prediction. The results reveal that this method can predict GOR accurately and within a reasonable time frame, provided sufficient training data is available. Our findings offer valuable insights for oil well operators and engineers aiming to enhance their GOR prediction strategies.

Keywords: Identification, Neural Networks, Machine Learning, Oil and Gas Production.

1. INTRODUCTION

Oil wells are complex systems that, while being very economically important, are very hard to accurately model due to uncertainties associated with the process of oil extraction. In oil production, there are many variables susceptible to uncertainties that are important to assess the performance and behavior of the well. These uncertainties derive from that they are specifically related to reservoir variables, which are hard to measure and accurately model (Jahn et al., 2008). For instance, there are many fluid elevation mechanisms to help so called brown wells (Jahn et al., 2008). One such techniques is the Gas-lift, which is widely used in the oil and gas industry. In a gas-lift elevation system, gas is injected into above the production packer, to reduce the hydrostatic pressure and improve the flow of oil to the surface. Predicting the amount of gas needed for optimal gas-lift performance is one of the many challenging tasks involving production, as it requires taking into account various factors such as reservoir characteristics, fluid properties, and production rates, many of which being either uncertain or hard to model. An example of a crucial variable concerning oil production is the gas-oil ratio (GOR) of a well, which is essentially a metric to measure the gas inside the production fluid Jahn et al. (2008).

GOR is a measure of the amount of gas that is produced in relation to the amount of oil from the reservoir. The measure can be expressed either as a mass ratio between gas and oil, or a volumetric ratio, which is the one considered for this work. GOR, or Gas-Oil Ratio, serves as a metric to gauge the quantity of gas generated relative to the amount of oil extracted from the reservoir. This measurement can be articulated either as a mass ratio, denoting the relationship between gas and oil masses, or as a volumetric ratio, the latter being the focus of this study.

The GOR holds significance in assessing the production performance of wells and in the process of optimization calculations. Despite its importance, the GOR is subject to uncertainties due to its reliance on the properties of reservoir fluids, introducing potential inaccuracies in any physical model attempting to incorporate these properties (Jahn et al., 2008).

The inherent uncertainties provide an avenue for leveraging data-driven black-box modeling, commonly known as machine learning tools, to estimate GOR. The potency of machine learning tools lies in their capability to derive accurate models solely from input-to-output data (Bishop, 2006).

Static black-box models have found application in addressing challenges within the oil and gas industry, as evidenced by Thanh et al. (2020), where a neural network is employed to forecast CO2 recovery. Another noteworthy instance is presented in Sheikhoushaghi et al. (2022), where various

⋆ The authors acknowledge and thanks Cenpes/Petrobras, and CAPES-PrInt Grant 88887.717389/2022-00 for the financial support aswell as the brazilian agencies CNPq and CAPES that have partially funded the research under projects: CNPq 403949/2021-1, CNPq 406477/2022-1, CNPq 304032/2019-0 and CAPES/Print/Automação 4.0.

neural networks, including a convolutional one, serve as surrogate models for estimating oil production.

In the context of this study, a neural network model is utilized to ascertain the Gas-Oil Ratio (GOR) by considering production well pressures and gas-lift injection. This approach draws inspiration from the methodology outlined in Junior and Moreno (2019), albeit in a reversed manner.

The scarcity or unavailability of field data in oil and gas production and reservoir systems poses a challenge. To address this issue in the realm of engineering and physical systems identification, one can turn to the family of methods known as Physics-informed machine learning (Karniadakis et al., 2021). The primary pillars of physics-informed machine learning involve data augmentation through simulated data, incorporating models with structures aligned with system physics, and applying regularization via the residue of physical equations.

The primary aim of this study is to formulate a technique to enhance Gas-Oil Ratio (GOR) estimation in gas-lifted oil wells. We assess and compare the efficacy of Neural Networks (NN) in addressing this specific issue. Despite their inherent complexity, neural networks demonstrate superior performance compared to less intricate models, particularly when a sufficient amount of data is available, as indicated by Bishop (2006). This advantage is evident in the comprehensive coverage of the operating range.

Our methodology involves the initial training of the Neural Network using the MARLIM simulation. In our case, data is generated from an oil well simulated in the MARLIM software. The model is trained based on these results, and the trained neural network is subsequently tested against its real-world counterpart in the MARLIM system Seman et al. (2020). Subsequently, real-world data is utilized as a test dataset for GOR prediction. The results highlight the effective performance of the Neural Network, achieving accurate GOR estimation under practical conditions. The contributions of this work goes twofold:

- Obtaining models from machine learning (namely NNs) to predict GOR from readily available well data, being able to provide a method to easily infer these parameters.
- Testing and Validating the approach of inserting simulated data into training to perform in the real world counterpart of the same well.

The work is distributed as follows: Section 2 describes the oil production process, Section 3 describes the NN model, Section 4 describes the experiments, Section 5 that describes a new branch of experiments that was realized using real world data, and Section 7 concludes the work.

2. BACKGROUND AND CONTEXT

An oil well is the tubing responsible for extracting the oil from the reservoir and placing it into the surface. Figure 1 is a schematic representation of a simple gas-lifted oil well. The production tubing, where the oil and gas fluid flow, is at the center, while the gas for the gas-lift is provided from an annulus involving the tubing. There are many ways for a well model to represent the fluid being produced, with the simplest being considering the oil production a linear

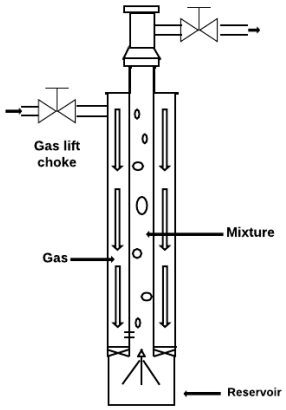

Figure 1. Schematic of a gas-lifted oil well.

relation of the well bottom-hole pressure P_{bh} (which is also referred to as the pressure measured at the Permanent Downhole Gauge (PDG)), and the pressure at the reservoir P_r:

$$\omega_{o,r} = PI(P_r - P_{bh}) \tag{1}$$

$$\omega_{g,r} = GOR\omega_{o,r} \tag{2}$$

$$\omega_{w,r} = \frac{BSW}{1 - BSW}\omega_{o,r} \tag{3}$$

where $\omega_{o,r}$ is the oil mass flow from the reservoir, $\omega_{g,r}$ is the mass gas flow from the reservoir, and $\omega_{w,r}$ the water mass flow from the reservoir. PI, the productivity index, BSW, the basic sediments and water quantity, and GOR are three important variables to measure the performance of a given oil well (Jahn et al., 2008). Given these equations, we can easily obtain GOR from the flows alone, however flows tend not to be easy to directly measure in oil wells, hence the need to develop a model from indirect measures instead. We consider the reservoir production equations as a principle from where to stipulate a function mapping for the calculation of GOR, based on lumped models such as the ones derived by Jahanshahi et al. (2012), which is roughly as follows:

$$GOR = f(\omega_{g,gs}, P_r, P_{bh}, P_{wh}, PI, BSW, ...) \tag{4}$$

where $\omega_{g,gs}$ is the gas-lift source flow, and P_{wh} is the wellhead pressure. In production well systems, the bottom-hole pressure has a certain degree of dependence on the wellhead pressure, which is the reason the wellhead variables are present as inputs in this mapping. Also, other inputs are accepted, was Eqn. (1) to Eqn. (2) represent a simplified version of the flow dynamics. The function f could be physically modeled, however since well production models are very prone to uncertainty, we experiment the viability of data-driven approaches instead, by gathering data simulated from the steady state MARLIM model on a real well owned by Petrobras, see Seman et al. (2020) for more details.

There is a limited number of scientific articles specifically addressing the prediction/estimation of GOR. One example for GOR estimation is Hashemi Fath et al. (2020), where a MLP neural network is used to estimate solution GOR, with the difference being that the data is obtained from data obtained from analyzing the fluid. However, sim-

ilarities can be found in other oil well process predictions, such as Virtual Flow Metering (VFM) (Bikmukhametov and Jäschke, 2020) and various oil flow processes. In these related fields, regression-based models have often been demonstrated to be less accurate than Neural Networks (NNs), as in AL-Qutami et al. (2018) and Sandnes et al. (2021). These works compared NNs to Gradient Boosted Trees. The difference to our work is that we apply data-driven methodology to predict and estimate GOR instead of flow.

3. MODELING

In this section, we discuss the modeling techniques employed in our study to predict GOR in oil rigs. We explore the use of Convolutional Neural Networks (CNN) as our primary approach. We provide an overview of this method, including its mathematical foundations, parameter tuning, and model selection.

The data used in this study comprises two separate types of sources. The first type is simulated data generated by the Marlim simulator, while the second type is real-world data collected from 10 oil rigs.

We first created, separated, and indexed the simulated data-set before integrating its components into a vector.

$$\mathbf{x}_i^\top = [Whp_i \ GOR_i \ BSW_i \ PI_i \ qgl_i \ qo_i] \tag{5}$$

Where Whp is the Well's head pressure, qgl the flow of gas from the annulus to the tubing and qo being the total oil flow rate.

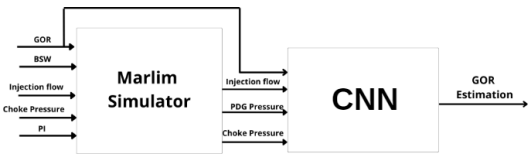

Figure 2. block diagram based on GOR estimation.

In Figure 2 it is shown which variables are possible to use as breakpoints in Marlim. Based on this, a model of inputs and outputs is built for the training and estimation of the CNN model.

3.1 Convolutional Neural Network (CNN)

The ability of 1D CNNs to learn complex relationships and mappings between input variables makes them an ideal choice for GOR prediction, given the intricate nature of the data and the numerous influencing factors.

A 1D convolutional layer uses multiple learnable filters or kernels to convolve with the input data and generate feature maps. These feature maps represent the presence of learned features at different spatial locations in the input. Mathematically, the 1D convolution operation is defined as follows:

$$y_i = \sum_m I_{(i-m)} K_m$$

where I is the input data, K is the kernel or filter, and y_i is the output feature map.

A sample CNN architecture includes Conv1D, MaxPooling1D, Flatten, Dense layers, and a Dense output layer, and non-linear activation functions, such as Rectified Linear Unit (ReLU). They introduce non-linearity into the network, allowing the CNN to learn complex relationships between input variables. The ReLU activation function has the following definition:

$$f(x) = \max(0, x)$$

with x as the function input.

4. EXPERIMENTS AND RESULTS

4.1 Data collection and Methodology

For the initial training and testing phase, only the Marlim simulated data was used. The neural network (NN) model was trained on the same dataset with an 80/20 training-test split. This means that 80% of the data was used for training the model, while the remaining 20% was used to evaluate the model's performance.

4.2 Experiments with CNN

This section contains the details of the configuration processes with the Marlim software and the CNN. To obtain similar results of a real well model, we estimate some parameters based on inputs generated by Marlim. In this paper, BSW and GOR will be set up as the desired parameters for estimation. Thus, after obtaining the analysis of the results generated by Marlim, the CNN model was implemented with the GOR parameter as the output of the system. The inputs for the model are: well head pressure,injection flow rate, PDG, oil flow rate, PI and BSW.

The first experiments using the CNN estimator were set up in Marlim applying variations of GOR, choke pressure, gas injection flow rate, and productivity index, while keeping BSW values constant. For the CNN machine learning model, the well head pressure, injection flow, PDG pressure, oil flow rate, PI, and BSW were included as inputs, and the GOR parameter was classified as system output. In the initial tests, the potential of the CNN model was quickly recognized as it achieved high accuracy without extensive optimization. The model demonstrated a precision beyond 90%, which was a promising start. As the experiments progressed and the model was further fine-tuned, the R-squared (R^2) value reached an 0.9675 value, showcasing the model's excellent performance, as depicted in Figure 3.

For a new batch of experiments, we decided to explore the potential of the CNN model in predicting GOR with a varying BSW, based on the promising results obtained during the previous tests where BSW was kept constant. Predicting GOR accurately while considering varying BSW values or vice versa is a significant challenge

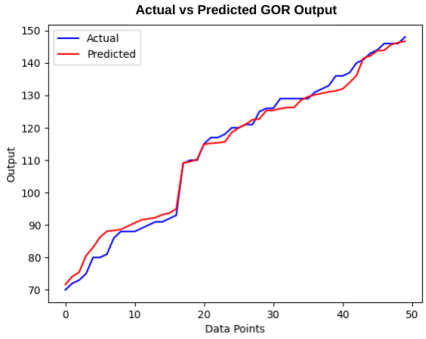

Figure 3. R-squared value of 0.9675 for the CNN model

and would demonstrate the robustness and adaptability of the model in real-world scenarios. We optimized the CNN model by exploring different hyperparameters and architectural changes, employing techniques like adjusting the learning rate, batch size, and layer configurations to find the optimal setup. After several iterations, the CNN's predictive capabilities improved significantly, maintaining high precision and outperforming initial results, underscoring the value of optimization.

Figure 4 showcases the initial accuracy of the CNN model with varying BSW values. This demonstrates the potential of the model in the early stages of the experiments.

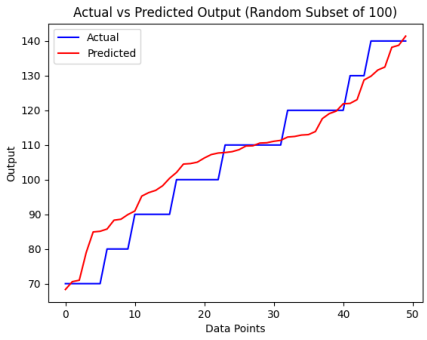

Figure 4. Initial accuracy of the CNN model with varying BSW

As the experiments progressed and the model was further optimized, we achieved a state-of-the-art CNN model that is able to reach an impressive R-squared (R^2) score of 0.994 while using Marlim for training and testing with varying BSW values, as shown in Figure 5.

The same neural network that is capable of predicting the GOR is also capable of predicting BSW with even greater precision, since that value is easier to determine.

Building upon the success of the optimized CNN model, we decided to further challenge its robustness and generalization capabilities by testing it with real-world measurements, while still training the model with Marlim-generated data. This experiment aimed to investigate the model's ability to adapt to real-world data, which may contain uncertainties and noise not present in the synthetic data generated by Marlim. In this part of the experiments, the model was put to the test against field measurements obtained from actual well performance data. It is important to note that the transition from synthetic data to real-world data may introduce discrepancies and affect the model's performance, as the complexity of field data can be higher.

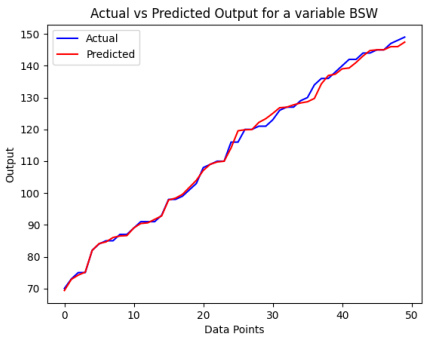

Figure 5. Predicted and real GOR values: R-squared score of 0.994

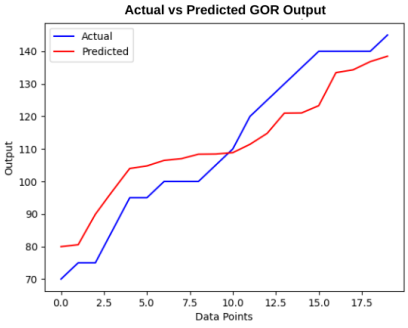

Figure 6. Performance metrics of the CNN model with real-world measurements without access to PI: R-squared score of 0.62

The results of the experiment with real world test data showed a decrease in the model's precision, dropping from an R-squared score of 0.98 to 0.62 when tested with real-world measurements. This reduction in performance can be attributed to not only the inherent complexities and uncertainties present in field data that are not captured in the Marlim-generated synthetic data, but also the fact that the PI was not available for training and testing, thus, reducing the availability of information to our neural network. Figure 6 illustrates the performance metrics of the CNN model when tested with real-world measurements. Although the precision has decreased in comparison to the results obtained with synthetic data, the model still demonstrates a reasonable level of accuracy, considering the complexity of real-world data. These results indicate that the CNN model could potentially be further improved and adapted for real-world applications, especially with access to more data and with additional fine-tuning and optimization to account for the unique characteristics of field data.

The experiments performed help to illustrate the effectiveness of the various optimization techniques employed and the overall success of the CNN model in predicting BSW and GOR, even when considering varying BSW values. The results highlight the potential of the CNN model as

a powerful tool for well performance estimation in the oil and gas industry.

5. ENHANCED GOR PREDICTION USING COMBINED REAL-WORLD AND SIMULATED DATA TRAINING FOR CNN

5.1 Introduction to the New Approach

5.1.0.1. Objective In the last chapter, we discussed the use of a neural network tested with Marlim data for real-life applications. In this chapter, we will dive deeper into this concept, employing both real-life data and Marlim simulated data, to more accurately predict the behavior of GOR in actual oil rigs.

5.1.0.2. Innovation Since the GOR predictions derived from Marlim-trained data for real oil well measurements were not as accurate as we believe achievable, this chapter explores how we simultaneously used real oil well data and Marlim simulated data to feed the same neural network.

5.1.0.3. Hypothesis The hypothesis driving this research is that a CNN trained on a combination of real-world data and simulated data from the Marlim model will yield a more accurate GOR prediction. This hypothesis stems from the idea that real-world data, with its inherent complexities and variabilities, coupled with the controlled and comprehensive nature of simulated data, can provide a more robust training environment for the CNN.

5.2 Methodology

5.2.0.1. Neural Network Structure As previously stated, the neural network structure we are using was specially designed for this task. It is a flattened, one-dimensional convolutional neural network.

5.2.0.2. Data Collection The data used was collected from real-world oil wells operated by Petrobras. Specifically, the data originates from the oil well RJS-710, spanning from January 17th to November 1st, 2021.

5.2.0.3. Training Process We divided the training and testing data, allocating 80% to Marlim simulated data and 20% to real oil well measurements. The testing was conducted exclusively with real oil well measurements.

5.3 Experiment 1: Combined Data Training

5.3.0.1. Experiment Setup We provided the previously mentioned neural network structure with a dataset comprising 80% Marlim simulated data and 20% real measurement data. The network architecture included five layers: two layers for input and output, with the remaining three being dense layers activated by ReLU functions.

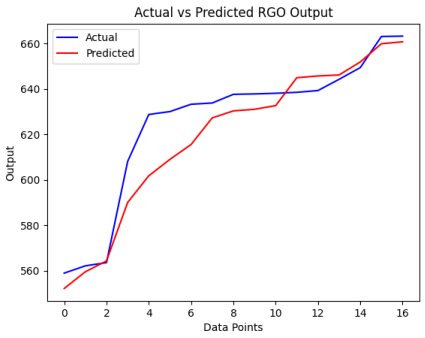

Figure 7. Graph of RGO predictions versus real measurements.

5.3.0.2. Performance Analysis As shown in Figure 7, our results were satisfactory, achieving an R^2 score of 0.689. This score is a significant improvement over the previous test, which used only Marlim training data and achieved an R^2 score of 0.62 (refer to Figure 6). The inclusion of real-world data evidently enhanced our model's prediction capabilities.

6. EXPERIMENT 2: REAL-WORLD DATA ONLY

Experiment Setup: Since using the real world data increased the predictive capabilities of our neural network, we decided that it was a good idea to test out a neural network that was only fed real world data, to see if its performance may surpass that of the Marlim simulated plus real world data.

Training Challenges: One of the big challenges that we faced during this training phase was over-fitting, since there wasn't that much real measurement data to train without needing to worry about that. So, to compensate, we implemented multiple measures like L1 (Lasso) regularization and dropout, which, through cross-validation with Marlim-based tests, proved to work pretty well.

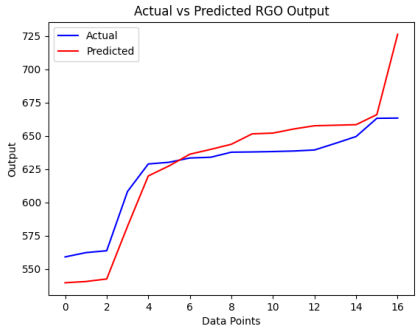

Figure 8. Graph of RGO predictions versus real world measurements using only real data for training

Performance Analysis: The R^2 score of 0.484 achieved in this experiment shows that a decent predictive capability can still be achieved using exclusively real world data, but, especially for our structure, it seems as if it is not the most optimal approach, since the R^2 score of 0.484 is considerably lower than the one that was reached while

using a mix of Marlim simulated data and real world measurements.

6.1 Analysis and Discussion

- Result Comparison: In comparing the results of the two approaches — one using combined Marlim and real-world data, and the other using only real-world data — we observed notable differences in performance. The combined data approach achieved an R^2 score of 0.689, significantly higher than the score of 0.484 achieved by the model trained solely on real-world data. This indicates a clear advantage in predictive accuracy when incorporating a blend of simulated and real data.
- Implications: These findings have important implications for GOR prediction in the oil and gas industry. The enhanced accuracy with combined data suggests that integrating diverse data sources can lead to more reliable and robust predictive models. This has potential applications in optimizing oil extraction processes and making more informed operational decisions.
- Benefits and Limitations: The benefits of using mixed data sources include improved model generalization and a richer training environment, leading to more accurate predictions. However, the limitations should not be overlooked. These include the potential for discrepancies between simulated and real data, challenges in data integration, and the need for large and diverse data sets to fully realize the advantages of this approach.
- Theoretical Insights: The superior performance of the combined data approach can be attributed to several factors. First, simulated data from Marlim provides a comprehensive and controlled environment for the CNN to learn from, while real-world data introduces the model to the complexities and variabilities of actual oil wells. This combination likely enables the network to develop a more nuanced understanding of GOR dynamics, leading to improved prediction accuracy.

7. CONCLUSION

Through the application of a single dimension flattened Convolutional Neural Network, we have successfully created an algorithm capable of predicting Gas-Oil Ratio in oil wells, particularly, those which we have modeled in Marlim. Our experiments have convincingly shown that the CNN, when trained with a combination of simulated and real-world data, outperforms other models that were trained exclusively on either type of data. This is a significant finding, as it demonstrates the CNN's robustness and adaptability in hadling diverse data sources.

The integration of Marlim simulated data with real-world measurements has emerged as a particularly effective strategy. This approach not only improves prediction accuracy but also enriches the model's learning environment, enabling it to capture the complexities and variabilities inherent in real-world oil well data. Our results indicate that such mixed data training is of essence, especially given the often limited availability of comprehensive real-world data in this domain.

In light of these promising results, our future research will concentrate on developing a context-based neural network. This new direction aims to extend our methodology to wells that do not have existing Marlim models. By incorporating available data and contextual information, we intend to design a neural network capable of accurately predicting well behavior in a broader range of scenarios. This model will build upon the insights gained from our current work with the CNN and mixed data training. Ultimately, our goal is to create a versatile and effective tool for predicting the BSW and GOR of real-world oil wells, enhancing operational decision-making in the oil and gas industry.

REFERENCES

AL-Qutami, T.A., Ibrahim, R.B., Ismail, I., and Ishak, M.A. (2018). Virtual multiphase flow metering using diverse neural network ensemble and adaptive simulated annealing. *Expert Syst. Appl.*, 93, 72–85.

Bikmukhametov, T. and Jäschke, J. (2020). First principles and machine learning virtual flow metering: A literature review. *Journal of Petroleum Science and Engineering*, 184, 106487.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Hashemi Fath, A., Madanifar, F., and Abbasi, M. (2020). Implementation of multilayer perceptron (mlp) and radial basis function (rbf) neural networks to predict solution gas-oil ratio of crude oil systems. *Petroleum*, 6(1), 80–91.

Jahanshahi, E., Skogestad, S., and Hansen, H. (2012). Control structure design for stabilizing unstable gas-lift oil wells. *IFAC Proceedings Volumes*, 45(15), 93–100.

Jahn, F., Cook, M., and Graham, M. (2008). *Hydrocarbon Exploration and Production*. Developments in Petroleum Science 55. Elsevier, 2nd ed edition.

Junior, D.H. and Moreno, U.F. (2019). Estimação de pressão de fundo de poço utilizando svr e ukf. In *Congresso Brasileiro de Automática-CBA*, volume 1.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 422–440.

Sandnes, A.T., Grimstad, B., and Kolbjørnsen, O. (2021). Multi-task learning for virtual flow metering. *Knowledge-Based Systems*, 232, 107458.

Seman, L.O., Miyatake, L.K., Camponogara, E., Giuliani, C.M., and Vieira, B.F. (2020). Derivative-free parameter tuning for a well multiphase flow simulator. *Journal of Petroleum Science and Engineering*, 192, 107288.

Sheikhoushaghi, A., Gharaei, N.Y., and Nikoofard, A. (2022). Application of rough neural network to forecast oil production rate of an oil field in a comparative study. *Journal of Petroleum Science and Engineering*, 209, 109935. doi:https://doi.org/10.1016/j.petrol.2021.109935.

Thanh, H.V., Sugai, Y., and Sasaki, K. (2020). Application of artificial neural network for predicting the performance of CO2 enhanced oil recovery and storage in residual oil zones. *Scientific Reports*, 10(1).