SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, Vol.04 No. 1 2025
DOI: 10.51984/sucp.v4i1.3994

المؤتمر العلمي الأول للتطبيقات الهندسية (ICEA'2024)
The First Scientific Conference on Engineering Applications (ICEA'2024)
Conference homepage: www.icea.ly

Metrics Of Models Evaluation for The Predictive Log Data and Vital Role of Machine Learning

Mohamed S. Elshbani , Saaida E. Al Nour*

Department of Electrical and Computer Engineering, Libyan Academy, Tripoli, Libya

Keywords:
Machin learning
Training Data
Predication
RMSE
MSE
MAE
R square
Confusion Matrix
ROC

ABSTRACT

Traditional methods of measuring well logs are expensive, error-prone and time-consuming, which has led to the development of machine learning models that can predict well logging based on well-log data. This study aims to determine the most effective metrics of model evaluation for predictive log by machine learning models for predicting of well logging based on available well-log data. The study covers a detailed explanation of the data-gathering and pre-processing techniques used. trained and evaluated based on their performance, namely linear regression, support vector machine (SVM), Neural Network (NN) and decision Trees (DT) models. The models were evaluated based on their Mean Squared Error, R squared, Mean Absolute Error and RMSE values, confusion matrix and ROC. Our results showed that the Decision Trees (DT) for MSE value of 10.86, achieving (RMSE) value of 3.29, (MAE) value of 2.225 and (R square) value of 0.92. These findings suggest that machine learning models can be a powerful tool for predicting of best training from well-log data, in particular, holds great promise for future modelling efforts in this area.

مقاييس تقييم النماذج لبيانات السجل التنبؤية والدور الحيوي للتعلم الآلي

محمداالشيباني و سعيدة النور*

قسم الهندسة الكهربائية و الحاسوب، الاكاديمية الليبية، طرابلس، ليبيا.

الكلمات المفتاحية:
تعلم الاله
تدريب البيانات
التنبؤ
RMSE
MSE
MAE
تربيع R
مصفوفة الارتباك
ROC

الملخص

الطرق التقليدية لقياس سجلات الآبار باهظة الثمن وعرضة للخطأ وتستغرق وقتًا طويلاً، مما أدى إلى تطوير نماذج التعلم الآلي التي يمكنها التنبؤ بتسجيل الآبار بناءً على بيانات سجلات الآبار. تهدف هذه الدراسة إلى تحديد المقاييس الأكثر فعالية لتقييم النموذج للسجل التنبؤي من خلال نماذج التعلم الآلي للتنبؤ بتسجيل الآبار بناءً على بيانات سجل الآبار المتاحة. وتغطي الدراسة شرحا مفصلا لتقنيات جمع البيانات والمعالجة المسبقة المستخدمة. تم تدريبهم وتقييمهم بناءً على أدائهم، أي الانحدار الخطي، وآلة ناقل الدعم (SVM)، والشبكة العصبية (NN)، ونماذج أشجار القرار (DT). تم تقييم النماذج بناءً على متوسط مربعات الخطأ ومربع R ومتوسط الخطأ المطلق وقيم RMSE ومصفوفة الارتباك وROC. أظهرت نتائجنا أن قيمة أشجار القرار (DT) لـ MSE كانت 10.86، محققة قيمة 3.29 (RMSE)، وقيمة 2.225 (MAE)، وقيمة 0.92 (R _Square). تشير هذه النتائج إلى أن نماذج التعلم الآلي يمكن أن تكون أداة قوية للتنبؤ بأفضل تدريب من بيانات سجل الآبار، على وجه الخصوص، تحمل وعدًا كبيرًا لجهود النمذجة المستقبلية في هذا المجال.

1. Introduction

Many real-world situations might benefit from the use of machine learning (ML) approaches due to recent advancements in ML and rising computational capacity [1]. With its abundance of data, the petroleum sector is well-positioned to take advantage of these strategies and provide major benefits. The measurements of physical attributes that are obtained during the drilling of exploratory boreholes and documented in well logs sequential recordings of features collected at regular depth increments usually make up the accessible data.

The features of the surrounding geology are inferred from the well logs using a variety of modelling methods. In the end, these help with commercial decision -making about the well's further development or the development of whole oil fields, depending on the hydrocarbon content estimate.

Well log interpretation is a time-consuming and costly process that transforms raw data into information that is useful to the commercial world. It calls for a large amount of human labor as well as a high level of skill and experience. Inaccurate hydrocarbon content estimation also has a big financial impact as it can lead to lost chances or high drilling expenses for a producing well with poor hydrocarbon return [2].

The well logs, which are inherently noisy and imperfect, include raw measurements taken with a variety of instruments. Petro-physicists were assigned the duty of correcting incorrect readings and guessing

*Corresponding author:

Article History : Received 25 March 2024 - Received in revised form 27 September 2024 - Accepted 15 October 2024

missing data in order to "condition," or clean, the well logs. Only once this has been completed can they proceed with the interpretation of the rock properties.

In this work, we determine the relationships among several physical characteristics measured within a borehole using supervised machine learning approaches. One of these attributes is then used as a dependent or target variable, while the other attributes are used as independent or input variables. Each well log record corresponds to a single observation, and we use a subset of these observations to train the models while keeping an evaluation-ready disjoint set. We only use data if the desired attribute has been recorded for training and evaluation purposes, utilizing the measured value as the ground truth. Petro-physicists will no longer need to label the data in advance, and our theory is that this will lead to more consistent data values and prevent human intervention.

2. Background

Using a collection of existing training data, supervised machine learning (ML) automatically determines a functional link between input and output variables. Each input comprises an ordered vector of values, referred to as independent variables or features, that characterize different aspects of the issue [3]. Through the evaluation of the learnt function across the input vector, the output values also referred to as dependent or target variables are predicted [5].

Regression is used for goals with a continuous, possibly infinite domain, whereas classification is used for targets where the target values constitute a (often small) finite set. This distinction is based on the domain of the target variables.

We used neural networks, Support Vector Machines and gradient tree boosting as two regression models to approximate the missing data in the logs. Both models are essentially different, even though they both take a vector of characteristics as input and output a goal value.

A neural network (NN) is made up of many layers, each of which has a variable number of neurons.

A neural network with one hidden layer is shown in Figure 1, and the network's exact design is determined by the issue it is meant to solve. Each neuron shown in Figure 2, the first input layer represents a single value from the input vector. In the training phase, these weights are learnt. The values of neurons in subsequent layers are computed as a weighted sum of their predecessors modified by an activation function. The last layer, called the output layer, shows the inferred value, which is derived from the vector of features supplied to the network, the activation functions (which specify a node's output based on its input), and the weights. At a neuron's output, the activation function serves as a decision-making body [8]. Based on the activation function, the neuron learns either linear or non-linear decision limits. Additionally, because of the cascading effect, it exerts a levelling impact on neuron output, preventing neurons' output from growing excessively big after several layers [6].

Neural networks can detect complex patterns in extremely non-linear data sets, such as well logs, since the activation function does not have to be linear to its parameters.

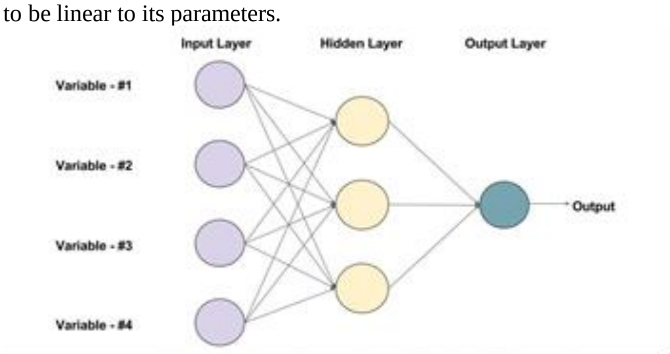

Fig. 1. Neural Network with one hidden layer (3 neurons)

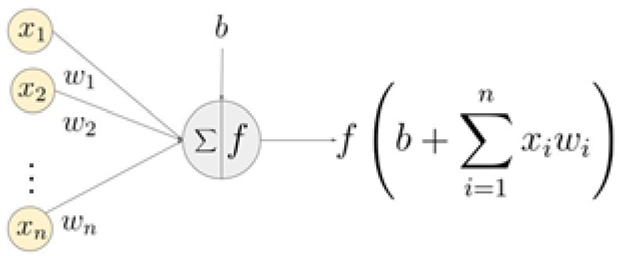

Fig. 2. A neuron showing the input (x1-xn), their corresponding weights(w1-wn), a bias (b) and the activation function f applied to the weighted sum of the input.

Gradient Tree Boosting (GB) is a term used to describe the widely used machine learning method called Gradient Boosting. Gradient Boosting is an ensemble learning technique that builds a powerful prediction model by combining several weak prediction models, often decision trees. Regression, classification, and ranking issues are just a few of the fields in which it has been effectively employed in machine learning. They frequently perform better than other conventional models and are strong algorithms for handling intricate forecasting jobs.

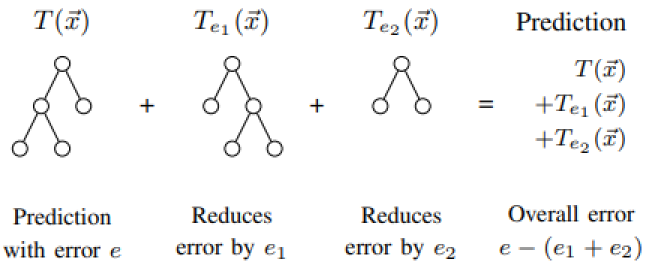

Fig. 3. The first tree (T) is a decision tree providing a rough estimation of the value to be predicted. The second tree (Te1) predicts and corrects the error of T, leading to a reduced overall error [4].

The accuracy of these regression decision trees is severely constrained since they can only reliably forecast the values that are represented by their leaves. Gradient Tree Boosting gets around this restriction by building an ensemble of trees, as shown in Figure 3, where each tree that follows after the first corrects the error of the one before it, improving the prediction until the error converges or the model begins to overfit.

3. Methodology

A massive data analysis and testing were undertaken, consistently, with the aid of metrics evaluation, testing and training, as well as, methods of machine learning. Furthermore, the raw data, have been elaborated and harvested, from well documented data, at Kaggle, data library, for data log. Alternatively, Excell sheets were utilized to represent data, after processing. Consequently, metrics of data models have been considered, for the Predictive purposes (RMSE, R_square, MAE, MSE).

Alternatively, machine learning techniques were applied in terms of decision trees (DT), support vector machines (SVM), neural networks (NN), and classification in terms of confusion matrix, in order to reach to realistic level of confidence. However, the methodology will be taken to gain insights analysis and more than adequate decisions ruling.

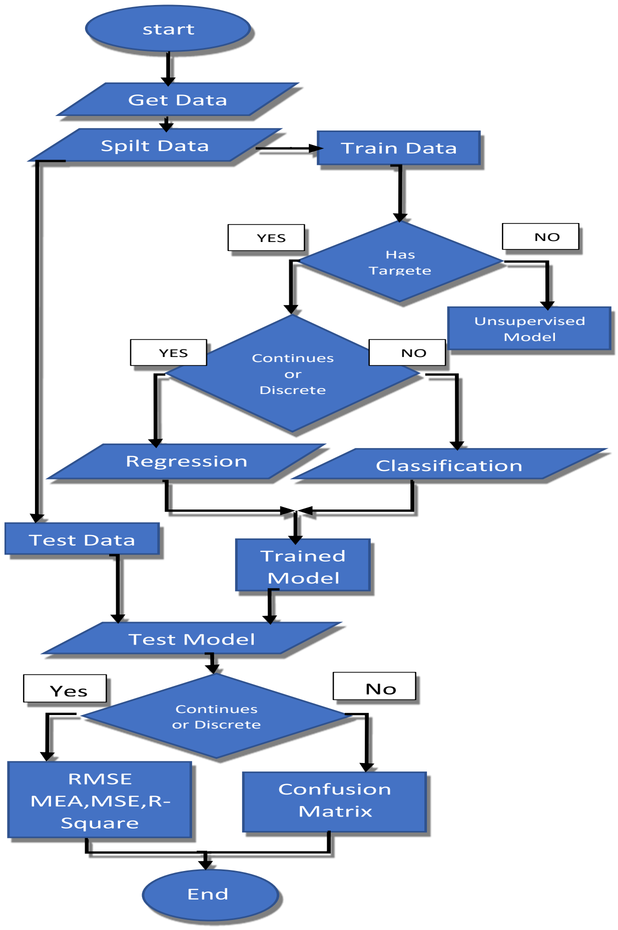

Fig. 4. model flowchart use supervised machine learning

4. Theoretical Background

Basic ideas and techniques form the theoretical basis for an extensive comparative analysis of several machine learning models in well log prediction and evaluation [4]. An important component of subsurface exploration is well logging, which procedures a variety of characteristics of rocks and fluids, as well as gamma-ray, resistivity, porosity, and acoustic logs [9]. By analyzing the physical and chemical characteristics of the materials that make up the Earth's subsurface, the field of petrophysics is important to the interpretation of well log data. Regression models such as support vector regression, decision tree regression, and linear regression are frequently used in supervised learning to make predictions, whereas classification models are used to handle discrete outcomes such as petrology classes [12]. Log transformations and depth-related features are two feature engineering strategies that improve the model's capacity to represent geological patterns.

Normalization, scaling, and filling in missing data are examples of data pre-processing procedures that guarantee the constancy of machine learning algorithms, various complexities in geological datasets in [4] are accommodated by model selection procedures such as decision trees, support vector machines, and neural networks. While classification metrics calculate accuracy, precision, recall, and F1-score for classification tasks, evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared assistance analyze prediction accuracy. By dividing the dataset into many folds, cross-validation techniques such as K-Fold Cross-Validation ensure robust model evaluation [11-12]. The models are further refined through ensemble approaches, hyperparameter tuning, and concerns for interpretability and integration of domain knowledge. Specialized strategies are needed to address imbalanced data, particularly in situations where certain lithologies are less common investigation.

5. Metrics for Model Evaluation

Define evaluation measures according to the type of task being predicted. Metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R Square (R^2) and Mean Absolute Error (MAE) are frequently employed for regression jobs [7-10-13].

Table 1. Definition of the measurements of MSE, R square, RMSE and MAE

Statistic	Description
RMSE	Root mean squared error. The RMSE is always positive and its units match the units of your response.
R-Squared	Coefficient of determination. R-squared is always smaller than 1 and usually larger than 0. It compares the trained model with the model where the response is constant and equals the mean of the training response. If your model is worse than this constant model, then R-Squared is negative.
MSE	Mean squared error. The MSE is the square of the RMSE
MAE	Mean absolute error. The MAE is always positive and similar to the RMSE, but less sensitive to outliers.

6. Results and Discussion

The regression leaner and classification trains models to predict data. we used training to search for the best regression model type, including linear regression models, regression trees, Gaussian process regression models, support vector machines,
ensembles of regression trees, and neural network regression models.

6.1. regression learner

The regression learner performs hyperparameter tuning by using Bayesian optimization. The goal of Bayesian optimization, and optimization in general, is to find a point that minimizes an objective function a point is a set of hyperparameter values, and the objective function is the loss function.

Table 2. Describe the measurements of MSE, R square, RMSE and MAE

	Regression Tree	Linear Regression	Wide Neural Network
MSE	11.2	53.6	12.3
R Square	0.91	0.59	0.91
RMSE	3.34	7.32	3.59

Table 3. Describe the measurements of optimizable MSE, R square, RMSE and MAE

	Optimizable Decision Tree	Optimizable SVM
MSE	10.8	49.6
R -Square	0.92	0.62
RMSE	3.29	7.04
MAE	2.225	4.121

In these models, the best validation in model of Decision Tree then we make to test of this model and the result show in table 4.

Table 4. Test results for testing model of decision tree

	Testing Model of Decision Tree
MSE (test	1.547
R square (test)	0.99
RMSE (test)	1.244
MAE (test)	0.818

6.2. CLASSIFICATION LEARNER

We used the same models that used in Regression Learner, and calculate Accuracy that was the best resulte of Accuracey equale to when apply the models of Decision Tree, plot confuion Matrix %99.9 and Receiver Operating Characteristic ROC as shown in figures**.**

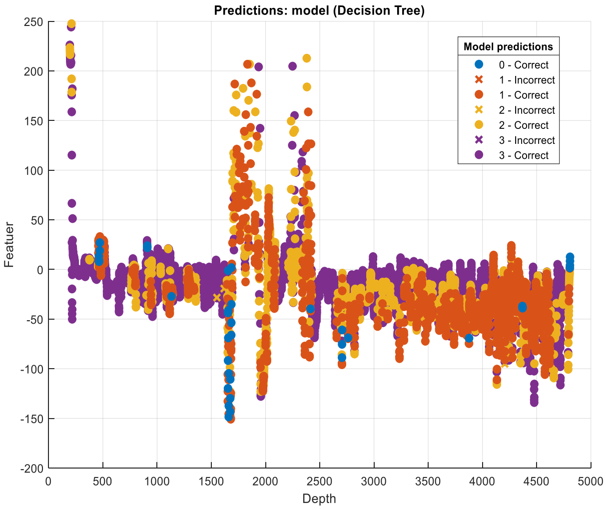

Fig. 5. Predication model for Data.

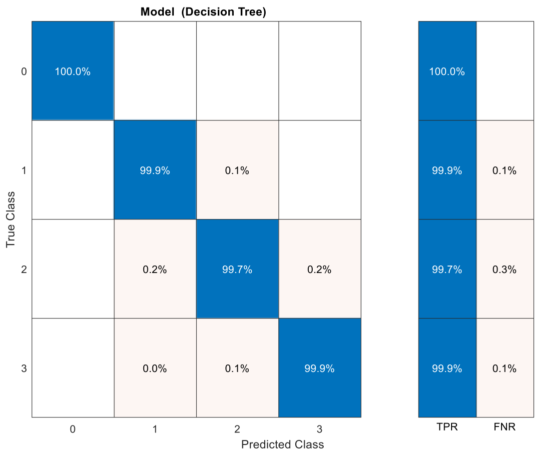

Fig. 6. Confuion matrix of Decision Tree.

where:

TPR= true positive rates. FNR=false negative rates.

PPV=positive predicative values. FDR=false discovery rates.

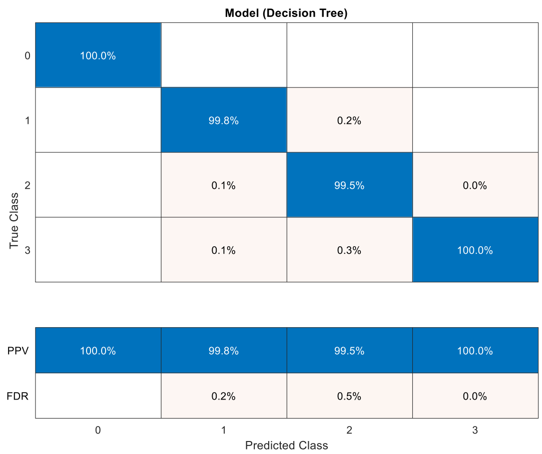

Fig. 7. Confuion matrix of Decision Tree

The **Area Under Curve** (AUC) value under a ROC curve, corresponds to the integral of a ROC curve true positive rate (TPR values) with respect to false positive rate (FPR) from FPR = 0 to FPR = 1. The AUC values are in the range from 0 to 1, and larger AUC values indicate better classifier performance. It used to compare classes and trained models to see if they perform differently in the receiver operating characteristic (ROC) curve.

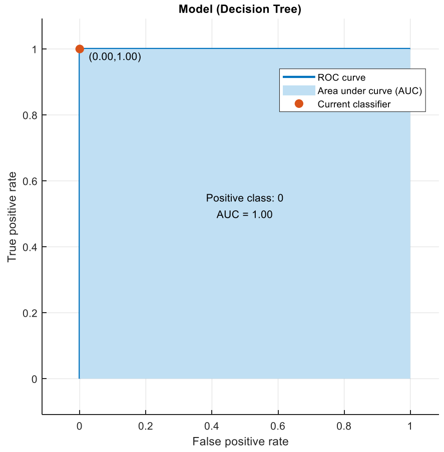

Fig. 8.. ROC of Decision Tree Model

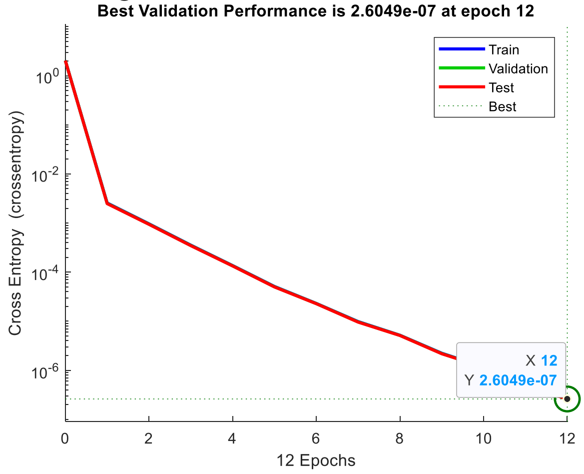

Fig. 9. Best validation performance.

W e calculate of Anomaly Detection using euclidean distance for scatter plot to visulize the clustermean and new sample as showen in follow figuer

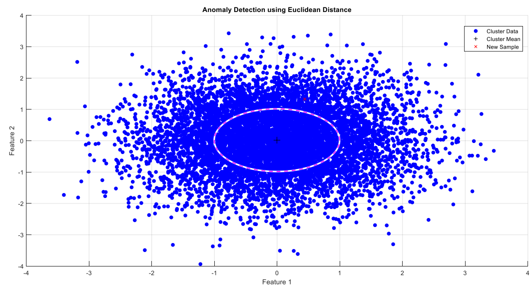

Fig. 10. Anomaly detection using Euclidean distance**.**

And calculate median absolute deviation (MAD) for normal and anomaly data for measure to Mahalanobis distance. Mahalanobis distance is a measure of the distance between a point and a distribution to visulize in follow figuer 11. It is useful in identifying outliers and in classification problems

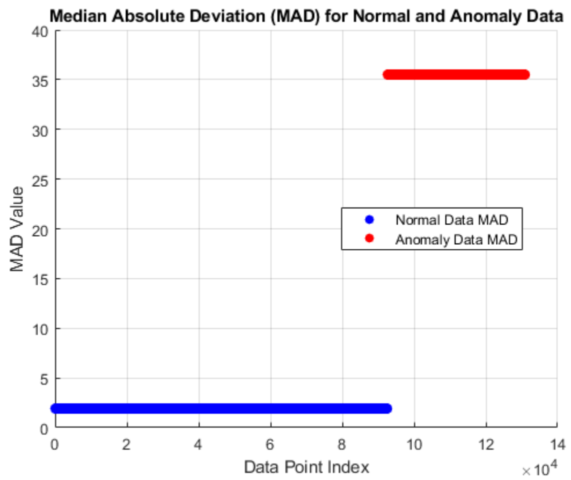

Fig. 11. Anomaly detection using Mahalanobis distance**.**

7. Conclusion

Finally, in an effort to get beyond the drawbacks of conventional measuring techniques including expense, inaccuracy, and time consumption, this study investigated the use of machine learning models for well logging prediction based on well-log data. The study concentrated on assessing the effectiveness of several machine learning models, including decision trees (DT), support vector machines (SVM), Neural Network (NN) and linear regression. The decision trees (DT) model beat the other models during the assessment process, as evidenced by its mean squared error (MSE) value of 10.86, root mean square error (RMSE) value of 3.29, mean absolute error (MAE) value of 2.225, and R-squared value of 0.92.

The best resulte of Accuracey in classification learner equale to when applay the models of Decision Tree %99.9**.**

Based on the available well-log data, these findings demonstrate how effectively the DT model predicts well logging. The study's conclusions demonstrate the potential of machine learning models as effective instruments for forecasting well logging and determining which well-log data makes for the best training and good predication. The industry may gain from increased well logging forecast accuracy, efficiency, and cost-effectiveness by utilizing these models, which will eventually enhance decision-making procedures and maximize resource exploration and production.

References

[1] Joshi, Dev, et al. "Prediction of sonic log and correlation of lithology by comparing geophysical well log data using machine learning principles." *GeoJournal* (2021): 1-22.

[2] Talebkeikhah, Mohsen, Zahra Sadeghtabaghi, and Mehdi Shabani. "A comparison of machine learning approaches for prediction of permeability using well log data in the hydrocarbon reservoirs." *Journal of Human, Earth, and Future* 2.2 (2021): 82-99.

[3] Ahmadi, Mohammad Ali, and Zhangxing Chen. "Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs." *Petroleum* 5.3 (2019): 271-284.

[4] Nordloh, Vito Alexander, Anna Roubícková, and Nick Brown. "Machine learning for gas and oil exploration." *arXiv preprint arXiv:2010.04186* (2020).

[5] Rahmati, Mehdi, Ghasem Zargar, and Abbas Ayati Zadeh Tanha. "Predicting density log from well log using machine learning techniques and heuristic optimization algorithm: A comparative study." *Petroleum Research* (2024).

[6] Ullah, Jar, et al. "A multidisciplinary approach to facies evaluation at regional level using well log analysis, machine learning, and statistical methods." *Geomechanics and Geophysics for Geo-Energy and Geo-Resources* 9.1 (2023): 152.

[7] Käppel, Martin, Stefan Jablonski, and Stefan Schönig. "Evaluating predictive business process monitoring approaches on small event logs." International Conference on the Quality of Information and Communications Technology. Cham: Springer International Publishing, 2021.

[8] Zhang, Dongxiao, C. H. E. N. Yuntian, and M. E. N. G. Jin. "Synthetic well logs generation via Recurrent Neural Networks." *Petroleum Exploration and Development* 45.4 (2018): 629-639.

[9] Struyk, C., et al. "LAS Version 2.0: A Digital Standard for Logs, Update February 2017." *Update* (2017).

[10] Chen, Stanley F., Douglas Beeferman, and Roni Rosenfeld. "Evaluation metrics for language models." (1998).

[11] Gąsior, Irena, and Anna Przelaskowska. "Estimating thermal conductivity from core and well log data." *Acta Geophysica* 62 (2014): 785-801.

[12] Rashidi, Hooman H., et al. "Common statistical concepts in the supervised Machine Learning arena." *Frontiers in Oncology* 13 (2023): 1130229.