Integration of Laboratory Data and Geophysical Logs for Permeability Prediction in the Barra Velha Formation, Santos Basin

*Note: Sub-titles are not captured for https://ieeexplore.ieee.org and should not be used

1st Filipe de Moura Antonio Cordeiro
Department of Geophysics
Federal University of Bahia)
Salvador, Brasil
filipemoura49@gmail.com

2nd Alexsandro G. Cerqueira
Department of Geophysics
Federal University of Bahia)
Salvador, Brasil
alexsandrocerqueira@ufba.br

***Abstract*—Understanding rock permeability – a measure of how easily fluids like oil and gas can flow through underground formations – is crucial for efficient hydrocarbon recovery. This is particularly challenging in heterogeneous carbonate reservoirs, such as those in the Barra Velha Formation in the Santos Basin. In this sense, this study proposes a robust machine learning-based approach to estimate rock permeability by integrating conventional well log data (geophysical measurements taken down a borehole) and laboratory core measurements (direct analysis of rock samples). Four regression algorithms were employed and compared: K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Support Vector Machines (SVM). The methodology included data preprocessing, with an emphasis on recovering missing well log data using XGBoost, followed by rigorous training and validation of the models using techniques such as holdout cross-validation. The results demonstrate that tree-based algorithms (Random Forest and XGBoost) provided the most accurate predictions, with good generalization in blind tests, a significant advancement considering the complexity of carbonate permeability. This research contributes to the enhancement of complex reservoir characterization, offering an efficient and economically feasible alternative to traditional permeability assessment methods.**

***Index Terms*—Permeability prediction, Carbonate Rocks, Pre-Salt, Santos Basin, Barra Velha Formation.**

I. Introduction

Borehole geophysics utilizes measurements taken inside wells to investigate the properties of the rock formations traversed. These measurements generate what are known as geophysical logs, which are continuous records of different physical properties (such as natural radioactivity, density, sound velocity, and electrical resistivity) as a function of depth. On the other hand, Petrophysics is the science that studies the physical and chemical properties of rocks, especially those relevant to the accumulation and flow of fluids, such as oil and gas. In this sense, one of the most important petrophysical properties is permeability, which characterizes the reservoir's fluid transmission capacity within its volume. This property is traditionally measured in the laboratory using rock samples (cores) or estimated by sensors through the borehole. However, the last method is more expensive and has some limitations.

The Santos Basin is extremely important for oil and gas exploration in Brazil [1]. With an area of approximately 350,000 km² and located along the coast of four Brazilian states, the Santos Basin is known for its oil and gas reserves in the pre-salt layer, which is situated below the seabed at a depth of about 5,000 meters. These reserves represent a national-scale economic potential that requires study. Understanding the permeability in the basins enables us to comprehend these reservoirs more effectively. In this context, this work aims to develop an efficient and cost-effective methodology for predicting permeability and identifying exploratory areas of interest using Machine Learning (ML) techniques based on basic geophysical logs.

Previous works explored the potential of machine learning techniques in other domains to predict additional geophysical properties. [2] showed that it is possible to use the supervised K-Nearest Neighbors (KNN) algorithm to estimate seismic velocity logs as a suitable alternative to replace conventional seismic logs as inputs for well-seismic correlation in a dataset from the Recôncavo Basin. Another example is the study conducted by [3], where machine learning algorithms were applied to estimate permeability in carbonate formations in Brazil. The results demonstrated the effectiveness of these techniques in permeability prediction, establishing a solid foundation for their application in the Santos Basin. Additionally, the study conducted by [4] on the prediction of Total Organic Carbon from geophysical well logs in the Santos Basin further contributes to this line of research, serving as a previous case study. These references, along with the aforementioned case study, form the basis for the methodology proposed in this study, thereby increasing the reliability and validation of the obtained results.

The approach developed aims to obtain reliable estimates of rock permeability in the Santos Basin, contributing to a better

Identify applicable funding agency here. If none, delete this.

understanding of the region's exploration potential. Moreover, the proposed methodology offers advantages in terms of efficiency and cost compared to traditional methods for measuring or estimating permeability. This enables a more comprehensive analysis and informed decision-making for oil and gas exploration in the Santos Basin. To achieve this objective, this study evaluates the application of a set of machine learning algorithms, including K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Support Vector Machines (SVM). These methods were selected due to their recognized ability to model complex and non-linear relationships, frequently encountered in geophysical and petrophysical data [5]–[7].

The focus area of this study was the Santos Sedimentary Basin, specifically the Búzios field. To accomplish this task, it was necessary to: i) perform preprocessing of geophysical data, as several geophysical logs, once we do not have all measurements of each property; ii) integrate laboratory and well data; iii) conduct training and validation of the supervised models; iv) perform prediction at the geophysical log scale in the wells used in the experiment.

II. Methods

Prediction of permeability in geophysical profiles is a complex and highly non-linear task, which makes it challenging to use parametric machine learning algorithms. Therefore, non-parametric algorithms were considered, which are machine learning methods that do not assume a specific distribution for the data or a fixed number of parameters. Instead, they allow the structure of the data to be learned directly from the data itself without making assumptions about the underlying functional form of the relationships. As a result, these algorithms are more flexible and adaptable to different types of data and can handle complex relationships. Examples of non-parametric algorithms include K-nearest neighbors (KNN), decision trees (DT), and random forests (RF).

A. Multiple Linear Regression

Multiple linear regression is a statistical technique used to model the relationship between a continuous dependent variable and two or more independent variables [7]. It is an extension of simple linear regression involving more than one independent variable. In multiple linear regression, the goal is to find a linear equation that best describes the relationship between the independent and dependent variables.

The equation for multiple linear regression is represented as follows (Eq. 1):

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_n \mathbf{x}_n + \epsilon \tag{1}$$

Where:

- $\mathbf{y}$ is the dependent variable to be predicted.
- $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_1}, ...\mathbf{x_n}]$ are the independent variables.
- β_0, β_1, β_2, ..., β_n are the regression coefficients representing the contribution of each independent variable to the dependent variable.
- ε is the error term that captures the variation not explained by the model.

B. K-Nearest Neighbors (kNN)

K-Nearest Neighbors (KNN) is a machine learning algorithm used for both classification and regression ([2], [7], [8]). In regression, KNN estimates the value of a response variable based on the values of the response variables of the nearest neighboring instances. The process of KNN in regression involves the following steps: first, the distance between each training instance and the test instance is calculated using a distance measure, such as Euclidean distance. Then, the "K" nearest training instances to the test instance are selected based on the smallest calculated distances.

After selecting the nearest neighbors, the predicted value for the test instance is calculated as the average or weighted average of the response variable values of the selected neighboring instances. Thus, KNN in regression uses the values of the response variables of the nearest neighboring instances to estimate the value of the response variable of the test instance.

C. Decision Tree

The decision tree is a widely used machine learning algorithm for solving classification and regression problems. It creates a tree structure where each internal node represents a test on an attribute, each branch represents the outcome of that test, and each leaf represents a class or an output value. The decision tree is built through recursive data partitioning based on the most important attributes, aiming to maximize purity or reduce uncertainty in the target classes [7], [9].

To evaluate the best parameters in the decision tree model, we used the information gain as the parameter of the tree split, which is given by the following equation:

$$IG(D, A) = Entropy(D) - \sum_{v=1}^{V} \frac{|D_v|}{|D|} \cdot Entropy(D_v) \tag{2}$$

where $IG(D, A)$ is the information gain when dividing the data D by attribute A, $Entropy(D)$ is the entropy of the data, V is the number of possible values for attribute A, D_v are the resulting data subsets from the split by A, $|D_v|$ is the number of samples in subset D_v, and $|D|$ is the total number of samples.

D. Random Forest

Random Forest is a machine learning algorithm that combines multiple decision trees for classification and regression tasks [7], [9]. Each tree is built independently, using a random sample of the training data and a random selection of attributes. The final prediction is obtained through majority voting (classification) or averaging (regression) of the trees. Random Forest performs well with large datasets, different types of attributes, and missing values. Additionally, it is less prone to overfitting compared to a single decision tree. In summary, Random Forest is a powerful and versatile approach in machine learning, capable of making accurate predictions in classification and regression problems.

III. Methodology

The experimental procedure was carried out by implementing Python code, utilizing the Scikit-learn library [5] to predict permeability measurements. These predictions were based on fundamental geophysical well logs. A visual representation of the adopted workflow can be seen in Figure 1. Each step of this process is detailed in the following subsections, providing an in-depth understanding of its execution and contribution to the research.

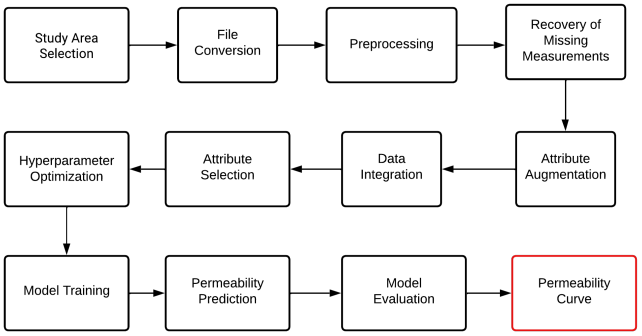

Fig. 1. Workflow employed in the permeability prediction.

A. Dataset

The data used for prediction in this study originate from the Búzios Field, located in the Barra Velha Formation (FBV), Santos Basin, approximately 200 kilometers off the coast of Arraial do Cabo, Rio de Janeiro, Brazil (see Figure 2). This dataset includes records from twelve wells, containing information such as gamma ray, transit time, density, porosity, photoelectric index, resistivity, and petrophysical permeability data associated with the FBV rocks. Data integration was performed between the geophysical logs and core samples, enabling the correlation of depths and the acquisition of permeability values at the same depths as the core samples.

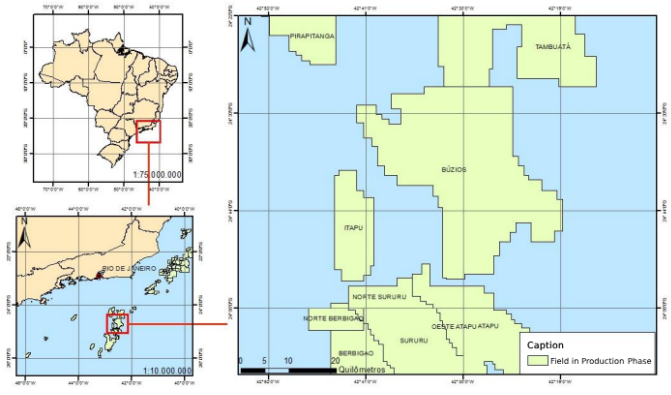

Fig. 2. Location map of the Búzios Field (Source: ANP).

B. Well-logging Preprocessing and Missing Data Recovery

The initial stage in analyzing a dataset involves editing and preprocessing, which in this study focuses on petrophysical information derived from well logs. This phase includes attribute standardization, consolidation of data from different wells, evaluation, and filtering of suspicious information, as well as the construction of categorical attributes. An important step was the identification and treatment of missing data. For instance, missing measurements were noted for AT10 and AT20 resistivity logs in several wells, and for DT and DTSM logs in others. Anomalous Gamma Ray (GR) values were also identified in well 3-BRSA-1195-RJS. Once these anomalous values were identified in the well logs, they were removed, and we performed a rapid regression to recover the rock properties at the well scale.

As the well log properties have different interval variations, we performed the normalization using the mean (μ) and the standard deviation (σ), as can be shown in Eq. 3.

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \tag{3}$$

For the recovery of missing essential petrophysical logs, such as gamma-ray, compressional slowness, and shear slowness (GR, DT, DTSM), the XGBoost algorithm was selected due to its demonstrated superior performance in previous tests for this purpose. Attribute selection for recovery models utilized a reverse selection method, and hyperparameter optimization was performed using Bayesian search. The recovery process prioritized logs with fewer missing values. Recovered GR logs for well 3-BRSA-1195-RJS and DT and DTSM for other wells showed low errors, providing high confidence in these imputed measures. AT10 and AT20 logs were not recovered due to their high correlation with the complete AT30 log and their infrequent selection in feature importance analyses. A median filter with a 9-sample window was applied to smooth certain logs, reducing noise and improving the reliability of subsequent modeling.

C. Data Integration and Feature Engineering

Integration between core and log data was achieved by comparing core depths with log measurement depths. Within one-meter intervals, the median of log measurements was related to corresponding core depths. To augment the dataset, logarithms of resistivity logs (AF90, AF60, AF30) and their differences were computed, as these showed a high correlation with permeability. Porosity estimated from density logs (PHID, Equation 4) and sonic logs (PHIS, Equation 5) were also calculated. The target values (permeability) were transformed to a base-10 logarithm to stabilize variance and linearize relationships, benefiting non-parametric models. The distribution of log-transformed permeability is shown in Figure 3.

$$\phi_D = \frac{\rho_m - \rho_B}{\rho_m - \rho_f} \tag{4}$$

$$\phi_S = \frac{\Delta t - \Delta t_m}{\Delta t_f - \Delta t_m} \tag{5}$$

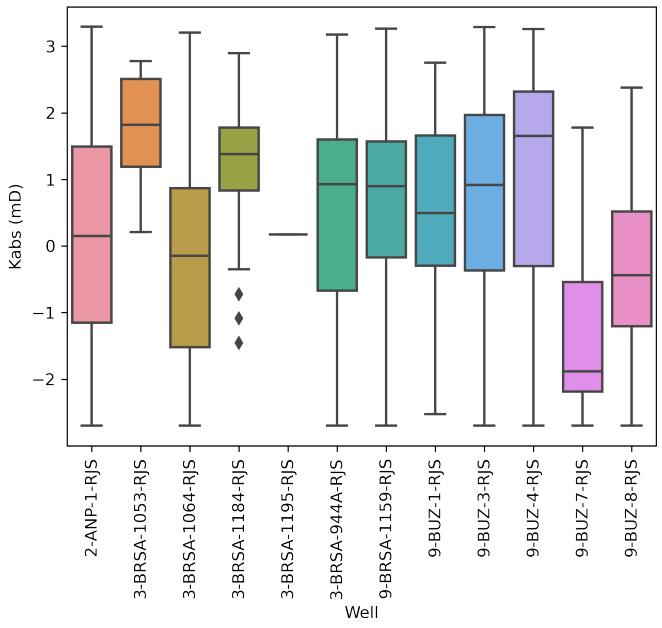

Fig. 3. Distribution of permeability samples for each well on a base-10 logarithmic scale.

D. Machine Learning Algorithms

This study investigated four supervised machine learning algorithms for permeability regression by integrating well and core data. The models were selected for their proven ability to model complex, non-linear relationships often found in geophysical and petrophysical data:

1) **K-Nearest Neighbors (KNN)**: A non-parametric method that predicts the value of a new instance based on the average of the values of its k closest neighbors in the feature space [8]. Proximity is typically measured using Euclidean distance.
2) **Random Forest (RF)**: An ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees for regression tasks [9], [10]. It is known for its robustness and ability to handle high-dimensional data.
3) **XGBoost (Extreme Gradient Boosting)**: An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable [6], [11]. It implements machine learning algorithms under the Gradient Boosting framework and has achieved state-of-the-art results on many problems.
4) **Support Vector Machines (SVM)** for Regression (SVR): SVR aims to find a function that deviates from y_i by a value no greater than ϵ for each training point x_i, and at the same time is as flat as possible [7], [12]. Kernel functions are used to map data into higher-dimensional spaces.

E. Model Training, Validation, and Optimization

In this work, we employed the Holdout cross-validation method to assess the generalization capability of the models, with the dataset randomly split into 80% for training and 20% for testing. While Holdout is simple, k-fold cross-validation can offer more robust estimates of model performance [7], [13]. Hyperparameter optimization was performed to minimize MAE. For KNN, an elbow method variant was used to determine the optimal k (Figure 4). For RF, XGBoost, and SVM, BayesSearchCV from Scikit-learn was employed to perform a more efficient and automated search for optimal hyperparameters. This approach involved over 50 iterations per algorithm, utilizing cross-validation to prevent overfitting.

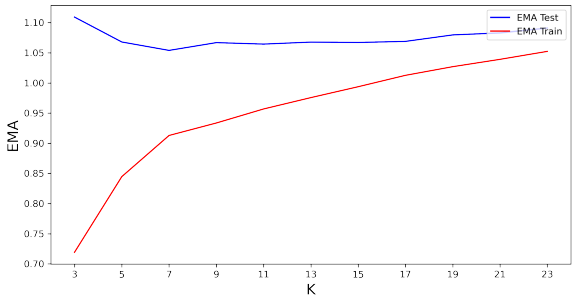

Fig. 4. Graph showing MAE for training and testing sets using the k-nearest neighbors algorithm.

Feature selection was guided by a decision tree-based method, with feature importances calculated to identify the most influential parameters (Figures 5 and 6). The final set of six most important features included Shear and Compressional Sonic Slowness (DTSM and DT), Gamma Ray (GR), Density Correction (DRHO), Photoelectric Factor (PEF), and Logarithm of Deep Resistivity (LogAT90).

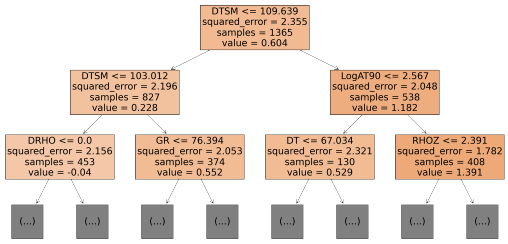

Fig. 5. The schematic image of the first two layers of the decision tree model used to select the best parameters.

F. Evaluation Metrics

The performance of the permeability prediction models was assessed using two primary metrics:

1) **Mean Absolute Error (MAE)**: Measures the average of the absolute differences between predicted ($\hat{y}_i$) and actual (y_i) values. It is robust to outliers.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

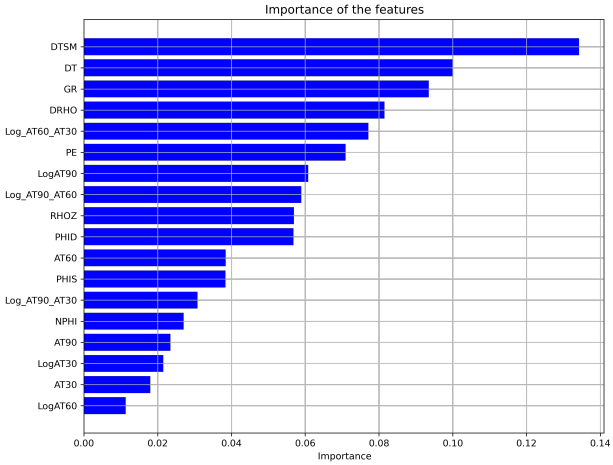

Fig. 6. Importance of each attribute selected by the decision tree method.

2) **Mean Squared Error (MSE)**: Measures the average of the squared differences between predicted and actual values, penalizing larger errors more significantly.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

Where n is the number of samples. These metrics provided a quantitative basis for comparing the accuracy of the different algorithms employed.

IV. Results and Discussion

The performance evaluation of the machine learning models in permeability prediction was conducted using the Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics for both the training and testing sets. The results obtained for the K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Support Vector Machine (SVM) algorithms are presented and discussed below.

A. Performance of Models in Permeability Prediction

The quantitative results of the model performance are summarized in Table I. This table presents the MAE and MSE values for each algorithm applied to the training and testing sets without filtering the output values for a realistic range of permeability to carbonates. Table II details the model performance for a specific permeability range (0.1 mD to 1000 mD), which represents a range of great practical interest in the characterization of carbonate reservoirs, according to [14], [15].

It is observed that all algorithms demonstrated the ability to learn from the training data and generalize to the test set, however, with different levels of accuracy. Random Forest and XGBoost frequently exhibited the lowest errors, indicating a good capability to capture the complex relationships between geophysical logs and permeability, a common goal in petrophysical machine learning applications [3], [16]. SVM also showed competitive results, consistent with its robust performance in other regression tasks [12]. KNN, being a simpler model, may have shown slightly inferior performance in some scenarios, especially in cases of high heterogeneity, a known limitation of distance-based methods in complex geological settings [8].

Figure 7 displays the geophysical logs of well 9-BUZ-3-RJS, which was used in both the training and testing sets. From the fifth track onwards in this figure, the permeability predictions generated by the four algorithms (KNN, Random Forest, XGBoost, and SVM) are presented. Permeability measurements obtained from core samples are represented by circles and "x" marks, respectively, allowing for a direct visual comparison between predicted and actual values.

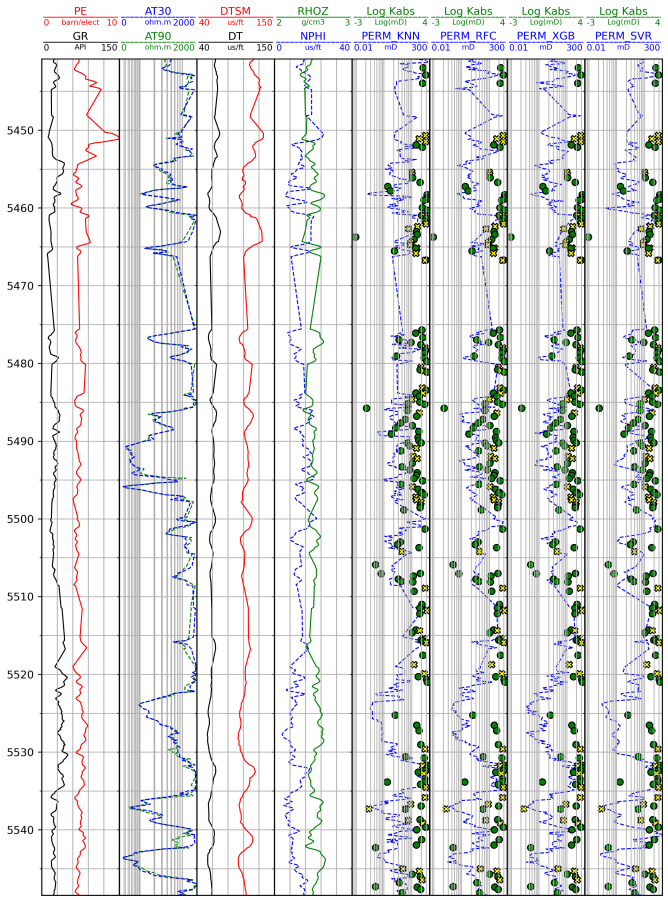

Fig. 7. Geophysical logs of well 9-BUZ-3-RJS, used in the training and testing sets. Starting from the fifth column, the predictions of the k-nearest neighbors, Random Forest, XGBoost, and Support Vector Machine algorithms are presented, along with core measurements represented by circles for training data and "x" for testing data.

B. Blind Test Analysis and Comparison with Core Values

For a more rigorous assessment of the model's generalization capability, a blind test was performed on well 3-BRSA-944A-RJS, whose data were not used during the training stage of any model. Figure 8 shows the permeability curves predicted by the different algorithms for this well. The central

TABLE I
RESULTS OF THE METRICS USED TO EVALUATE THE APPLIED MACHINE LEARNING MODELS ON THE TRAINING AND TESTING SETS. THE UNITS OF MAE AND MSE ARE, RESPECTIVELY, MD AND MD^2.

Models	MAE-Train	MSE-Train	MAE-Test	MSE-Test	MAE-Blind Test	MSE-Blind Test
k-Nearest Neighbors	±100.59	±76,914.00	±112.50	±93,732.50	±74.39	±216.07
Random Forest	±92.88	±68,715.86	±109.52	±92,577.86	±75.75	±223.85
XGBoost	±94.39	±70,443.54	±111.61	±93,836.10	±75.90	±224.45
Support Vector Machine	±113.24	±88,521.10	±120.44	±101,259.10	±75.91	±219.37

TABLE II
RESULTS OF THE METRICS USED TO EVALUATE THE MACHINE LEARNING MODELS FOR THE RANGE FROM 0.1 MD TO 1000 MD, APPLIED TO THE TRAINING AND TESTING SETS. THE UNITS OF MAE AND MSE ARE, RESPECTIVELY, MD AND MD^2.

Models	MAE-Train	MSE-Train	MAE-Test	MSE-Test	MAE-Blind Test	MSE-Blind Test
k-Nearest Neighbors	±75.92	±29,510.64	±89.76	±34,018.96	±58.53	±126.66
Random Forest	±67.73	±24,500.85	±85.38	±31,703.26	±59.00	±133.92
XGBoost	±65.18	±23,651.88	±85.60	±31,749.11	±58.92	±134.25
Support Vector Machine	±87.90	±35,521.58	±88.07	±35,900.31	±59.32	±132.48

track highlights the points representing laboratory permeability measurements from cores, serving as a reference for evaluating the accuracy of predictions in a real-world application scenario, a crucial step in validating machine learning models in geosciences [17], [18].

Figure 9 complements this analysis, showing the complete geophysical logs for well 3-BRSA-944A-RJS. Similar to Figure 7, from the fifth track onwards, the predictions of the four algorithms are displayed, along with the core measurements (light blue circles). This visualization allows for an assessment of how the different models behave throughout the entire logged interval, identifying zones of greater or lesser agreement with the reference data.

C. Discussion of Permeability Results

The prediction of permeability in carbonate rocks, such as those of the Barra Velha Formation, is inherently complex due to the heterogeneity of the pore system, which can include intergranular, intragranular, moldic, and vuggy porosity, in addition to the influence of fractures [14], [19]. The results obtained demonstrate the potential of machine learning techniques to assist in this challenging task, aligning with a growing body of research applying AI to petrophysics [20], [21]. Tree-based models (Random Forest and XGBoost) proved particularly promising, which can be attributed to their ability to model non-linear interactions and automatically identify the most relevant features in the input data [6], [10].

The Mean Absolute Error (MAE), for example, obtained by the Random Forest model, is ±92.88 mD, although representing an average, indicating the order of magnitude of the uncertainty associated with the predictions. It is crucial to consider that permeability in carbonates can vary by several orders of magnitude [22]. Therefore, the interpretation of errors must be made in the context of the natural variability of the property and the specific objectives of the application (e.g., identification of higher or lower permeability zones, flow estimation [23]). The performance differences between the algorithms underscore the importance of model selection and hyperparameter optimization for each specific dataset. Factors such as the quality and quantity of input data, the representativeness of core samples, and the geological complexity of the formation [15], [24] directly influence model performance.

Figures 7, 8, and 9 are essential for the discussion, as they allow for a qualitative analysis of the adherence of the predicted curves to the core data and the identification of possible biases or limitations of the models in certain intervals or geological facies. For example, one can observe whether the models tend to overestimate or underestimate permeability in high- or low-permeability zones or whether they capture abrupt variations in the property accurately.

It is important to note that, although machine learning models offer an efficient alternative for continuous permeability estimation [25], they do not completely replace direct core measurements, which remain the most reliable data source for calibration and validation [26], [27]. The integration of these different sources of information is fundamental for a more robust reservoir characterization.

V. CONCLUSION

This study demonstrated the feasibility and potential of applying machine learning techniques for permeability prediction in complex carbonate reservoirs, such as those of the Barra Velha Formation in the Santos Basin. The integration of conventional geophysical well log data with laboratory permeability measurements, combined with the use of algorithms such as K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Support Vector Machine (SVM), enabled the generation of continuous permeability estimates along the analyzed wells.

The results indicate that tree-based models, particularly Random Forest, exhibited robust performance in permeability prediction, successfully capturing complex non-linear relationships between input attributes and the target property. This aligns with findings from other studies that apply similar ensemble methods to geoscientific problems. The step of recovering missing geophysical logs using XGBoost also proved effective, ensuring a complete dataset for training the

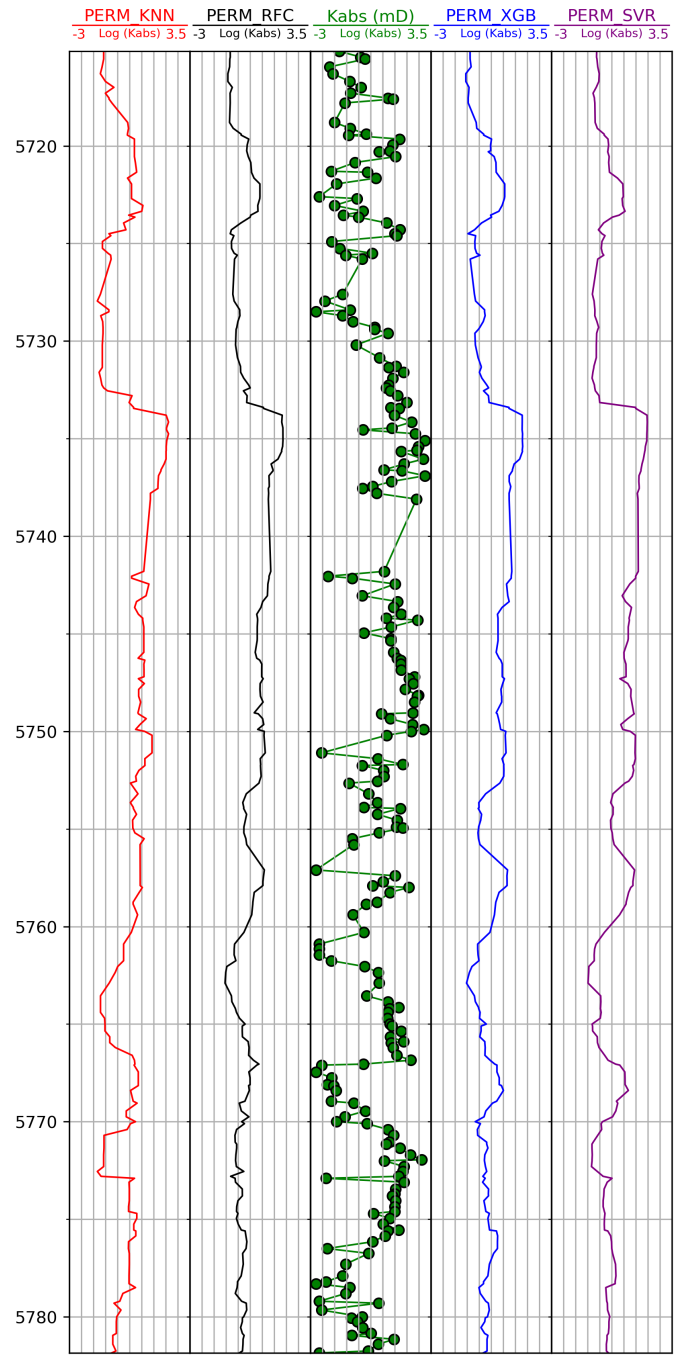

Fig. 8. The predicted permeability curves for different algorithms in the blind test, well 3-BRSA-944A-RJS, with the central column highlighting the points representing the laboratory core measurements.

permeability prediction models, a common challenge in well log analysis.

The evaluation of the models, using metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE), along with blind test analysis, provided a quantitative measure of the accuracy and generalization capability of the predictions. Although permeability prediction in carbonates remains a challenge due to the intrinsic heterogeneity of these rocks, the

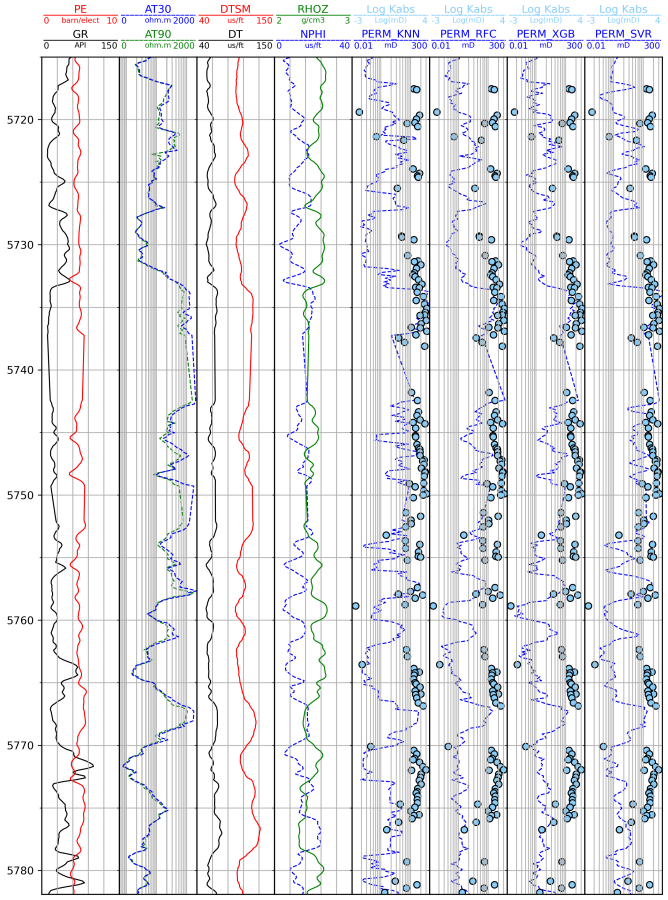

Fig. 9. Geophysical logs of well 3-BRSA-944A-RJS. Starting from the fifth column, the predictions of the *k*-nearest neighbors, Random Forest, XGBoost, and Support Vector Machine algorithms are presented, along with core measurements represented by light blue circles.

developed models represent a valuable tool to assist in reservoir characterization, complementing information obtained by traditional methods and contributing to the growing field of AI in petrophysics.

Acknowledgment

The authors would like to thank Petrobras for the F.C. Research scholarship, and the Instituto Nacional de Ciencia e Tecnologia de Geofisica do Petroleo (INCT-GP) for supporting this research. We would also like to acknowledge the support of the Postgraduate Program in Geophysics (PPGEOF) of the Federal University of Bahia (UFBA).

References

[1] T. M. d. Castro, "Avaliação dos reservatórios carbonáticos do pré-sal no campo de búzios, bacia de santos," 2019.

[2] M. Barbosa, V. Carneiro, and A. Cerqueira, "Seismic well tie using geophysical logs obtained from k-nearest neighbor regression algorithm," *Brazilian Journal of Geophysics*, vol. 40, 03 2022.

[3] C. d. N. Natalino and H. P. E. Almeida, "Aplicação das técnicas de redes neurais e lógica difusa na estimativa da permeabilidade em formações carbonáticas usando dados de perfilagem de poços e ressonância magnética nuclear (rmn)," 2021.

[4] C. A. C. d. Purificação, "Predição de carbono orgânico total a partir de perfis geofísicos de poços da bacia de santos," 2021.

[5] S. Raschka and V. Mirjalili, *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, (New York, NY, USA), pp. 785–794, ACM, 2016.

[7] G. James, D. Witten, T. Hastie, R. Tibshirani, *et al.*, *An Introduction to Statistical Learning with Applications in R*, vol. 112. Springer Science and Business Media, 7th ed., 2013. ISBN: 978-1-4614-7137-7.

[8] F. Nwanganga and M. Chapple, "k-nearest neighbors," pp. 221–249, 04 2020.

[9] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," in *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, pp. 81–227, Now Publishers, Inc., 2012.

[10] H. Singh, "Understanding random forests." https://medium.com/@harshdeepsingh_35448/understanding-random-forests-aa0ccecdbbbb. Accessed on: May 15, 2023.

[11] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Chem. Biol. Drug Des.*, vol. 297, pp. 273–297, 01 2009.

[13] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of Database Systems*, vol. 532–538, pp. 532–538, 01 2009.

[14] F. J. Lucia, C. Kerans, and J. W. Jennings Jr, "Carbonate reservoir characterization," *Journal of petroleum technology*, vol. 55, no. 06, pp. 70–72, 2003.

[15] V. P. Wright and A. J. Barnett, "An abiotic model for the development of textures in some south atlantic early cretaceous lacustrine carbonates," *Geological Society, London, Special Publications*, vol. 418, no. 1, pp. 209–219, 2015.

[16] L. Zhang and C. Zhan, "Machine learning in rock facies classification: An application of xgboost," pp. 1371–1374, 05 2017.

[17] Z. Huang, J. Shimeld, M. Williamson, and J. Katsube, "Permeability prediction with artificial neural network modeling in the venture gas field, offshore eastern canada," *Geophysics*, vol. 61, pp. 422–436, 03 1996.

[18] Z. Zhong, T. Carr, X. Wu, and G. Wang, "Application of a convolutional neural network (cnn) in permeability prediction: A case study in the jacksonburg-stringtown oil field, west virginia, usa," *GEOPHYSICS*, vol. 84, pp. 1–46, 08 2019.

[19] J. H. Schön, *Physical properties of rocks: Fundamentals and principles of petrophysics*. Elsevier, 2015.

[20] R. Farmanov, F. Feldmann, E. Mathew, M. Tembely, E. Al-Shalabi P.E., W. Alameri, S. Masalmeh, and A. Alsumaiti, "Application of machine learning for estimating petrophysical properties of carbonate rocks using nmr core measurements," 01 2023.

[21] X. Tong, L. Yan, and K. Xiang, "A prediction method of compacted rock hydraulic permeability based on the mgemtip model," *Minerals*, vol. 13, p. 281, 02 2023.

[22] A. J. Katz and A. H. Thompson, "Quantitative prediction of permeability in porous rock," *Phys. Rev. B*, vol. 34, pp. 8179–8181, Dec 1986.

[23] A. Weller and L. Slater, "Permeability estimation from induced polarization: an evaluation of geophysical length scales using an effective hydraulic radius concept," *Near surface geophysics*, vol. 17, no. 6, pp. 581–594, 2019.

[24] J. L. P. Moreira, C. V. Madeira, J. A. Gil, M. A. P. Machado, *et al.*, "bacia de santos," *Boletim de Geociencias da PETROBRAS*, vol. 15, no. 2, pp. 531–549, 2007.

[25] Z. Guan, X. Tang, B. Ran, S. Guo, J. Zhang, K. Du, and T. Jia, "Machine-learning-based automatic well-log completion and generation: Examples from the ordos basin, china," *Interpretation*, vol. 10, pp. 1–35, 05 2022.

[26] D. Tiab and E. Donaldson, "Petrophysics: Theory and practice of measuring reservoir rock and fluid transport properties: Second edition," *Gulf Prof.*, p. 1008, 12 2003.

[27] M. H. M. H. Rider, *The geological interpretation of well logs / Malcolm Rider.* Caithness: Whittles Pub., second edition. ed., 1996.