Iraqi Journal of Science, 2026, Vol. 67, No. 4, pp: 2328- 2343
DOI: 10.24996/ijs.2026.67.4.34

ISSN: 0067-2904

Differential time (DT) Log Prediction based on Random Forest Machine Learning Model.

Ahmed I. Bijan[1*] , Ali M. Al-Rahim[2]
[1] *Ministry of Education, Diyala Education Directorate, Baquba, Diyala, Postcode-32001. Iraq.*
[2] *University of Baghdad, College of Science, Department of Geology, Baghdad, Iraq.*

Received: 10/ 11/2024 Accepted: 27/4/2025 Published: 30/ 4/2026

Abstract

A frequent challenge is the absence of sonic log data for various reasons, which calls for effective solutions. One practical approach to address this issue is the estimation or prediction of missing well logs, commonly referred to as soft logging, which can significantly lower exploration costs in the oil and gas industry.

Given the structural complexity and heterogeneity of geological reservoirs, there are often pronounced nonlinear relationships among different well logs. This study introduces a substitution method that leverages the Random Forest machine learning algorithm for sonic log prediction, to create a dependable model for predicting sonic logs using the available log data.

The target log type in this research is the Differential Time (DT) log, which indicates the velocity of wave propagation through a geological formation. The features used for training the model include resistivity, density (RHOB), porosity (NPHI), gamma-ray (GR), and the sonic log (DT).

This study incorporates ILD [Deep induction (Resistivity Deep)] logs and Micro-resistivity (MSFL) logs, with the DT log being blinded in well A-5 for both the training and testing phases, serving as the prediction target within the same well.

Experimental results indicate that the proposed method (Random Forest) provides a more accurate estimation of missing logs than traditional techniques, demonstrating notable performance. The effectiveness of the sonic log (DT) model prediction is largely attributed to the Random Forest Algorithm, as evidenced by reductions in MSE, RMSE, and MAE to 4.783, 2.187, and 1.351, respectively, alongside an increase in R^2 to 0.893. Furthermore, the correlation coefficient between the actual and predicted DT logs reached r = 0.99.

Keywords: Well Log, Sonic log (DT), Pre-process, Random Forest model, Evolution Model

تنبئ بالمجس الوقت التفاضلي حسب موديل الغابة العشوائية احد موديلات تعليم الالة

احمد ابراهيم بيجان[1]*, علي مكي الرحيم[2]
[1]وزارة التربية , مديرية العامة لتربية ديالى , بعقوبة , ديالى , عراق
[2]جامعة بغداد , كلية العلوم , قسم علوم الارض , بغداد , عراق

الخلاصة

يعد التسجيل الجيوفيزيائي أحد أهم تقنيات القياس لتطوير واستكشاف النفط والغاز، ويعد السجل الصوتي جانبًا مهمًا يوفر وصفًا تفصيليًا للخصائص تحت السطح المرتبطة بخزانات النفط والغاز . المشكلة التي تحدث

*Email: ahmed.I.bichan@gmail.com

بشكل متكرر هي عدم توفر بيانات السجل الصوتي لأسباب مختلفة تحتاج إلى حل فعال، في الممارسة العملية، يعد فقدان تقدير/تنبؤ سجلات الآبار أو التسجيل الناعم إحدى الطرق الفعالة لتوفير تكاليف التنقيب عن النفط/الغاز.

نظرًا للتعقيد الهيكلي وعدم تجانس الخزان الجيولوجي، يجب أن تكون هناك علاقات غير خطية قوية بين سجلات الآبار المختلفة، والنهج البديل المقترح في هذا البحث هو التنبؤ بالسجل الصوتي استنادًا إلى خوارزمية التعلم الآليRandom Forest ، باستخدام بيانات السجل المتاحة لبناء نموذج موثوق للتنبؤ بالسجل الصوتي.

في هذا البحث، نوع سجل DT المتوقع هو سجل الزمن التفاضلي(DT) ، وهو موجة السرعة التي تنتشر في التكوين. تشمل ميزات السجل المستخدمة للتدريب على السجل الصوتي (DT) وأشعة جاما (GR) والكثافة (RHOB)والمسامية (NPHI) وسجلات المقاومة (ILD) والمقاومة الدقيقة(MSFL) ، ويتم تعمية سجل DT في البئر B–5 في التدريب والاختبار وكهدف التنبؤ في نفس البئر.

أوضحت النتائج التجريبية أن الطريقة المقترحة يمكنها تقدير السجلات المفقودة بشكل أكثر دقة من الطرق التقليدية، والأداء واعد. دقة التنبؤ بنموذج السجل الصوتي (DT) استنادًا إلى الغابة العشوائية، كما ثبت ذلك من خلال انخفاض MSE وRMSE وMAE وزيادة R^2 و4.783 و2.187 و1.351 و0.893 على التوالي. وأخيرا، كان معامل الارتباط r بين سجل DT الفعلي والمتوقع هو 0.99.

1. Introduction

The most crucial technology for gathering information for the thorough description, assessment, and management of oil and gas reservoirs in the exploration and development sector is geophysical logging. Using well logs, geophysicists or geologists can obtain various important reservoir geological parameters such as porosity, permeability, oil-water saturation, lithology, and sedimentary micro-faces.

However, due to drilling conditions, instrument faults, differences in logging conditions, data loss due to improper storage, and the concern of exploration costs, there will inevitably be incomplete or even missing logs. Accurately estimating or predicting missing data is very attractive to geophysicists and geologists [1].

Sonic logs and other petrophysical logs such as GR, NPHI, RHOB, etc., evaluate lithology, reservoir properties, hydrocarbon carrier properties, and other important parameters in hydrocarbon resources. Sonic logs also contain important information about the formation, which is necessary for oil and gas exploration and production activities [2]. However, it is an essential component to compute other logs, such as shear velocity, Vp/Vs ratio, and reservoir characterization to discriminate lithology and fluids [3].

Unfortunately, shear logs are often either missing due to cost reduction in oil and gas drilling operations or unreliable due to cycle-skipping and poor borehole conditions. To address this challenge, numerous academics try to uncover the non-linear correlations between various well logs. The application of machine learning in the geophysical domain has experienced a significant improvement in recent years.

Several researchers ([4] , [5] , [6] and [7]) have shown the potential of machine learning algorithms in accurately predicting log values in various geoscience and petroleum engineering applications, making excellent machine learning to tackle big data in the oil industry such as petrophysical data, to perform machine learning and produce data-driven results.

Numerous researchers ([8], [9] [10] and [11]) have examined the Mishrif Formation,
which is categorized into four stratigraphic units: MA, CR, MB1, and MB2. According to [12].

The aim of the Random Forest model demonstrated promising results with a Root Mean Squared Error (MSE) of 4.783, a Root Mean Square Error (RMSE) of 2.187, a Mean Absolute Error of 1.351, and a Coefficient of Determination (R^2) of 0.893

2- Materials and Methods

Machine learning is the process of automatically acquiring knowledge from data and identifying existing patterns without the need for explicit programming or human intervention.

There are two primary types of learning: supervised and unsupervised. In supervised learning, especially in regression tasks, the model is trained on labelled data, which provides valuable information by indicating that the data corresponds to specific labels.

This research used 8719 observation data points of Mishrif Formation from ten wells for training and testing the Random Forest model. The feature selection process showed a statistically significant relationship between each input log and the shear log.

In this study, Orange data mining software was used to train Random Forest models for predicting the sonic log, while Google Colab was utilized for graphical analysis.
Developing a machine learning model typically involves four key phases: data gathering, preprocessing, model training, and model evaluation [13]. Figure 1 illustrates the workflow of the regression analysis performed in this study using Orange Data Mining version 3.36.

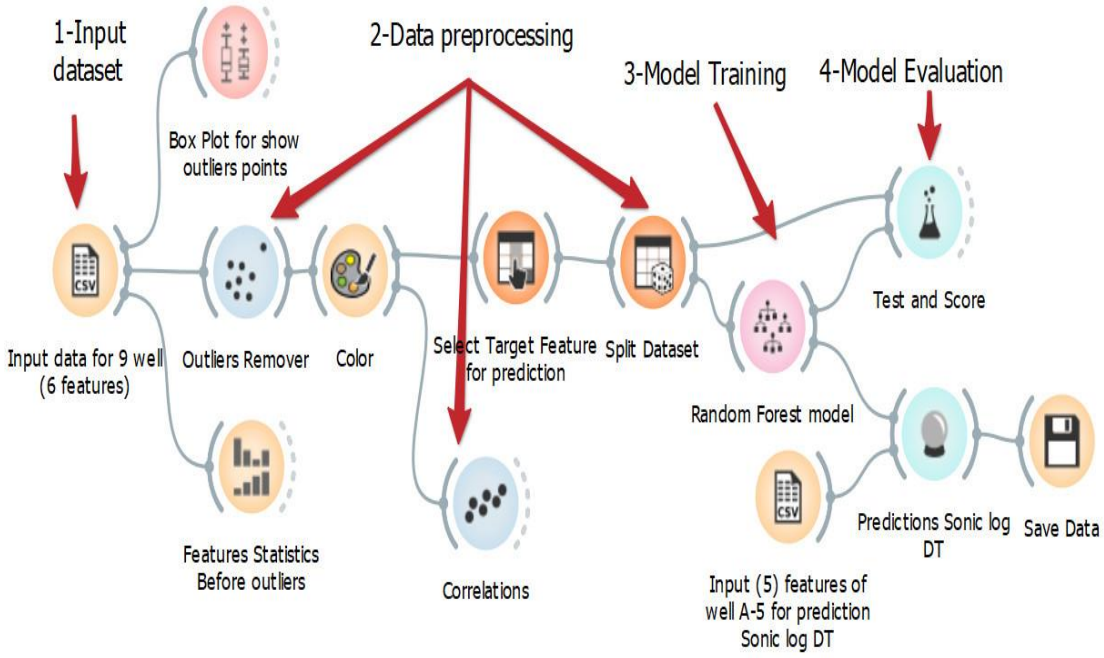

Figure 1: Shows the workflow of classification the study in Orange data mining softwareV3.36.

2-1 Data gathering

The dataset used for the machine learning models is an open-source dataset made available by the Basrah Oil Company. It is located in the XXXX field within the Basrah governorate in southern Iraq, as shown in (Figure 2). The dataset consists of 10 wells (A-3, A-4, A-5, A-15, A-17, A-18, A-19, A-34, A-39, and A-40), with a total of 8,719 observation

data points (samples). Each well contains six features corresponding to six wireline log measurements as shown in Figure 3.

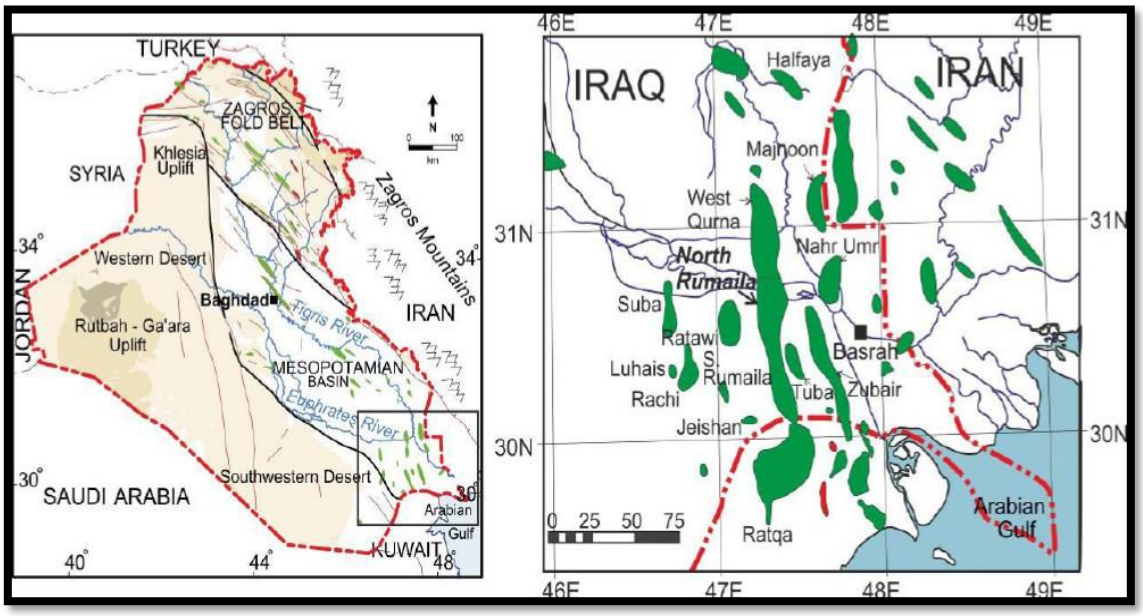

Figure 2: Studied area showing oil fields, southern Iraq, modified from [14].

Data points (samples). Each well contains six features corresponding to six wireline log measurements as shown in Figure 3.

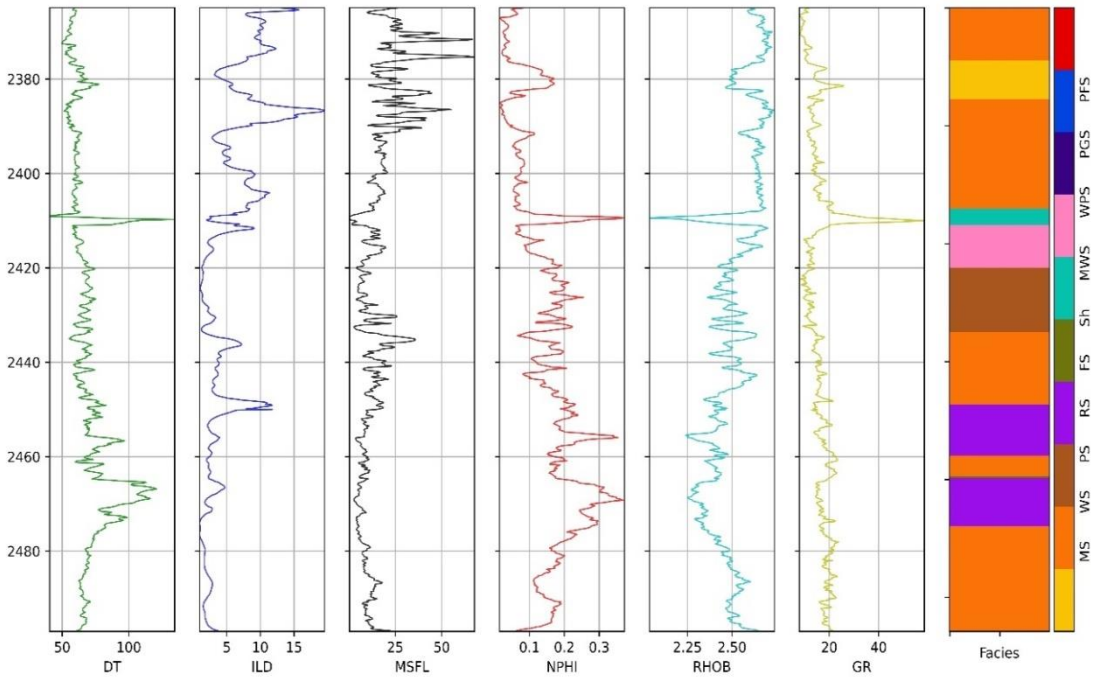

Figure 3: Shows all the logs and facies in well A-3 by Google Colab.

1) DT: Acoustic Transit Time (P-Sonic) (uS/ft), which is the velocity measurement.
2) ILD: Deep induction (Resistivity Deep) (Ohm-m) is the resistivity measurement.

3) MSFL: Micro Spherically Focused Log (Oh-m) is the resistivity measurement for clay on the wall of a well.
4) NPHI: Neutron porosity index in petrophysics.
5) RHOB: Density log (gm/cm3).
6) GR: Gamma-ray (api).

(Figure 3) displays the compressional sonic (DT), Deep Induction (ILD), Micro Spherically Focused (MSFL), Neutron Porosity (NPHI), Bulk Density (RHOB), and Gamma Ray (GR) logs from the representative well A-3 in the training set. Nine wells were used for training, while one was reserved for a blind test of the data-driven model.

2-2 Data preprocessing

Data preprocessing, often referred to as raw data gathering, is a vital step in preparing data for machine learning training. This phase is necessary to clean the data, which ultimately improves the accuracy and effectiveness of the model.

In machine learning projects, encountering unclean data is quite common. Before any analysis can proceed, the data must be properly cleaned and formatted. Consequently, data preprocessing is a critical component of the overall workflow.

2-2-1 Normalization well log

Normalization is rescaling data from its original range so that the values fall within a specified range, typically between 0 and 1 or -1 and 1. This technique is particularly useful when the approximate upper and lower bounds of the data are known, there are few or no outliers present, and the data displays a nearly uniform distribution ([15] and [16]).

Cross plots are a widely utilized for visualizing the relationship between two properties as they vary with rock type. They are effective for identifying outliers or unfamiliar points, as illustrated in Figure 4. The cross-plot diagram features the histogram distribution of petrophysical logs (DT, ILD, MSFL, NPHI, RHOB, GR) along the diagonal, while the off-diagonal sections display every combination in a 2-D cross-plot of each log.

There are various methods for removing outliers, with one widely used approach involving the establishment of lower and upper boundaries defined by the interquartile range (IQR).

A data point is classified as an outlier if it falls outside the limits set by the first quartile (Q1) and the third quartile (Q3).

Specifically, a point is deemed an outlier if it is greater than Q3 or less than Q1 [17]. The lower and upper boundaries are computed using the following equations:

Lower boundary = Q1 – (1.5 × IQR) (1)

Upper boundary = Q3 + (1.5 × IQR) (2)

Where: IQR (Interquartile Range) = Q3 – Q1.

(Figure 5) illustrates the cross-plot after removing outliers from the dataset. You can see how the outlier point in (Figure 4) obscures the prevailing values, especially when compared to (Figure 5), where the prevailing values are clearly shown.

The relationships between the measurements and facies labels are not evident from the presented cross-plots, highlighting the potential benefits of employing machine learning techniques in this context.

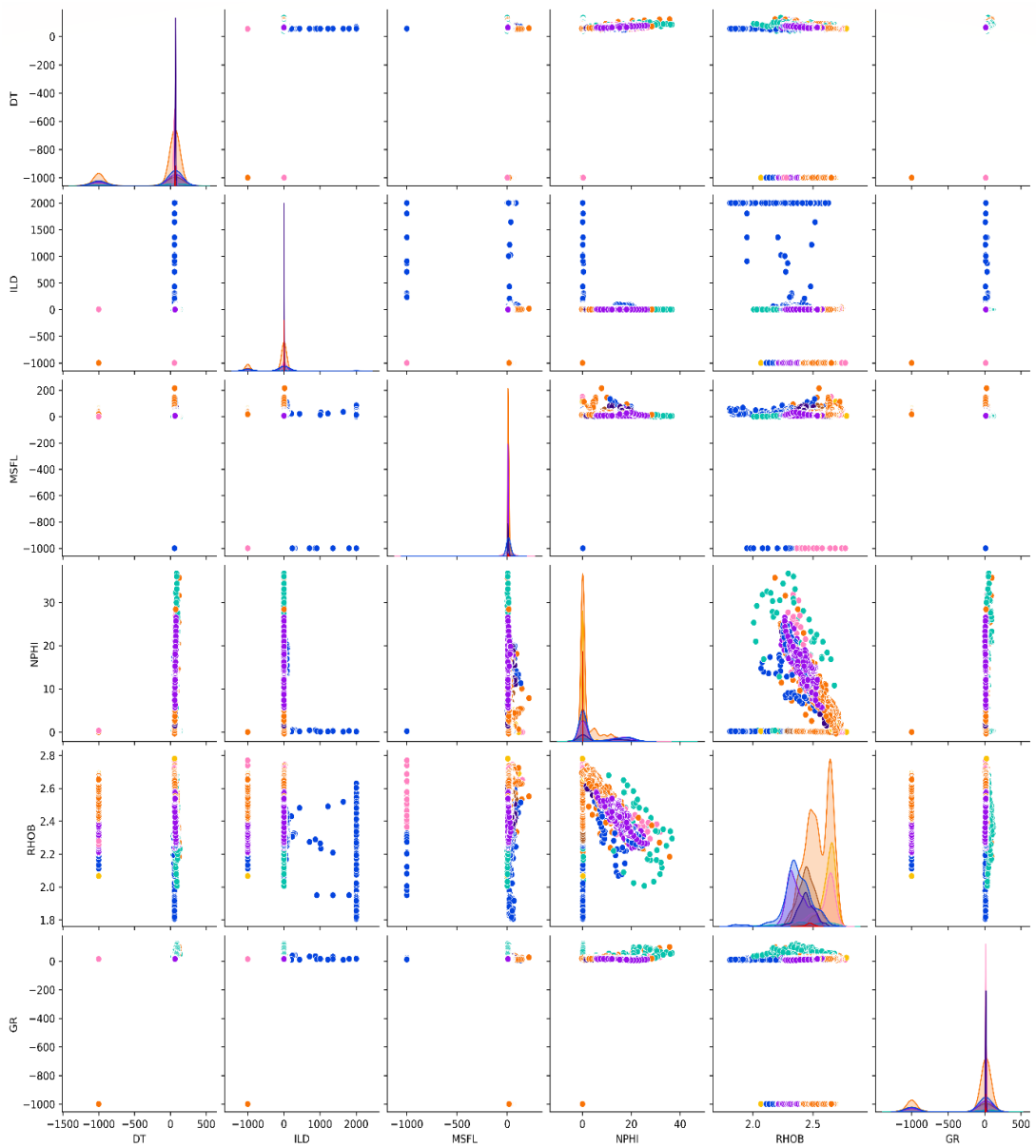

Figure 4: Cross-plot matrix generated using the Seaborn library before extracting outlier points, by Google Colab.

2-2-2 Correlations (Matrix Analysis)

A crucial step in the process is to evaluate the collinearity among all features (logs) that will be utilized in developing the predictive model. High collinearity can lead to confusion within the model, as these inputs may provide redundant information concerning the output ([18] and [19]).

To mitigate this issue, it is essential to filter out the highly correlated variables by selecting one variable from each pair of correlated variables.
Collinearity is assessed using a Pearson correlation coefficient heat map, as illustrated in Figure 6.

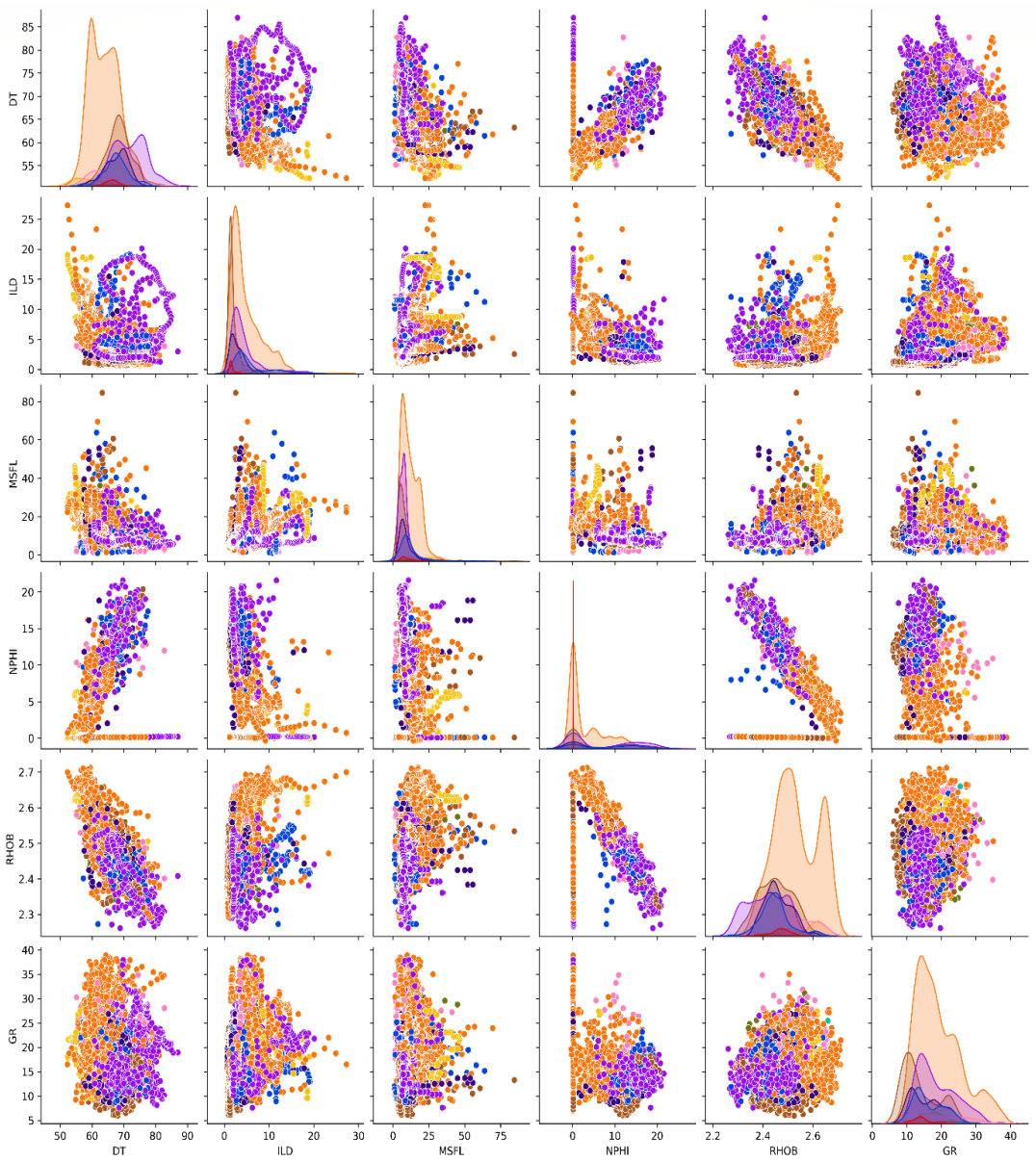

Figure 5: Cross-plot matrix generated with the Seaborn library after extracting outlier points, by Google Colab.

The Pearson correlation coefficients range from -1 to +1, where a negative value indicates an inverse relationship, a positive value indicates a direct relationship, and a zero value signifies no relationship. A high positive value implies that the features are strongly collinear.

2-2-3 Data split and cross-validation

Data partitioning is an essential step in machine learning, ensuring that models are evaluated on unseen data [20].

According to [21], the data is typically divided into two sets: a training set used to train the models and a testing set designated for final evaluation.

A common issue is overfitting to the training data, especially with small datasets, and cross-validation is a widely used technique to mitigate this problem.

In cross-validation, the training set is split into multiple folds. The machine learning estimator is trained on the remaining folds and progressively evaluated on each fold.
This method helps reduce the risk of overfitting by assessing the models across various partitions of the training data. At the end of this process, the model with the highest cross-validation score is selected to make predictions on the testing set for final evaluation.

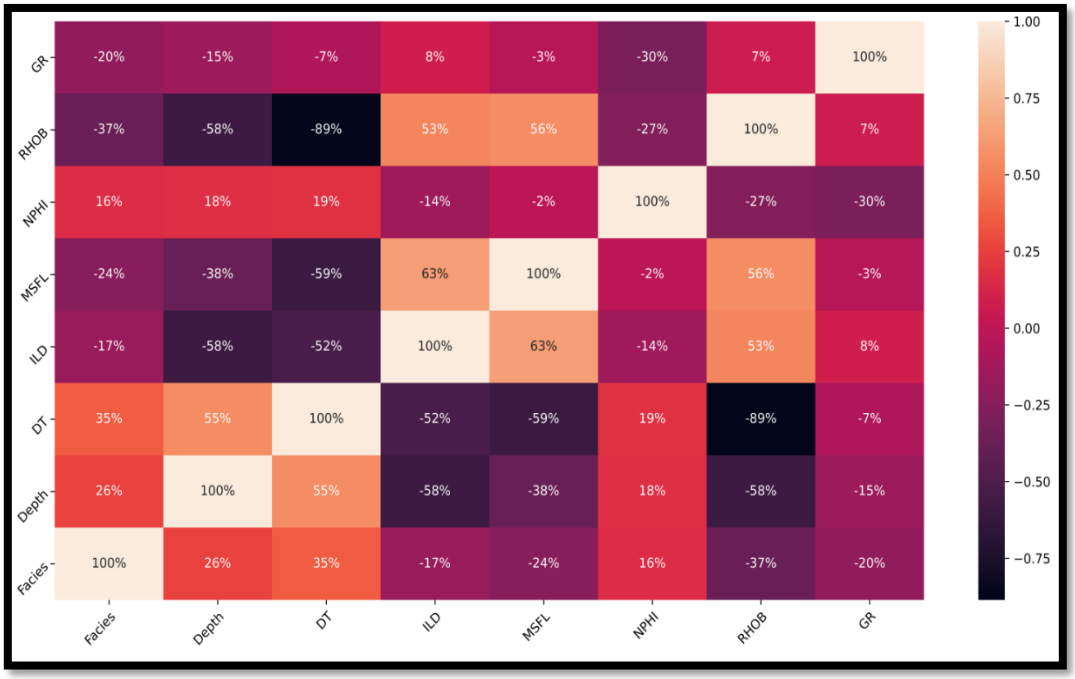

Figure 6: The Pearson's correlation coefficients of all features in the dataset, by Google Colab

In this study, the dataset comprises 8,719 samples, with nine wells utilized for training and one well reserved for blind testing of the data-driven model.
Well A-5 contained 911 records, which were excluded from the total, resulting in 7,808 samples from well A-5 used for training to predict the sonic log.
Figure 7 illustrates the random split of the dataset into two groups: a training dataset consisting of 70% (5,856 samples) and a testing dataset comprising 30% (1,952 samples).

The training dataset was used to establish rules applicable to all features, while the testing dataset was utilized to make predictions and evaluate the accuracy of our machine learning model.

2-3 Random Forest model training

Random Forest (RF) is an ensemble machine learning technique that integrates multiple decision trees built upon the classification and regression tree model [22].
The core structure of the RF algorithm is depicted in Figure 8. For our analysis, we specifically use the training well logs to calculate D using the following equation:

$$D=\{(X\_i=[X\_ij\ j=1,2,\ldots\ldots,M],y\_i),i=1,2,\ldots\ldots,N\} \quad (3)$$

In this case, M stands for the number of input well logs, and N for the number of samples. Through the use of the bootstrap resampling technique, D is randomly transformed into n tree datasets with the same number as the original dataset.

Following this, *n tree* regression decision trees are constructed from these bootstrap datasets. For each bootstrap dataset, approximately one-third of the samples, known as out-of-bag (OOB) data, are not selected for training the corresponding decision tree.

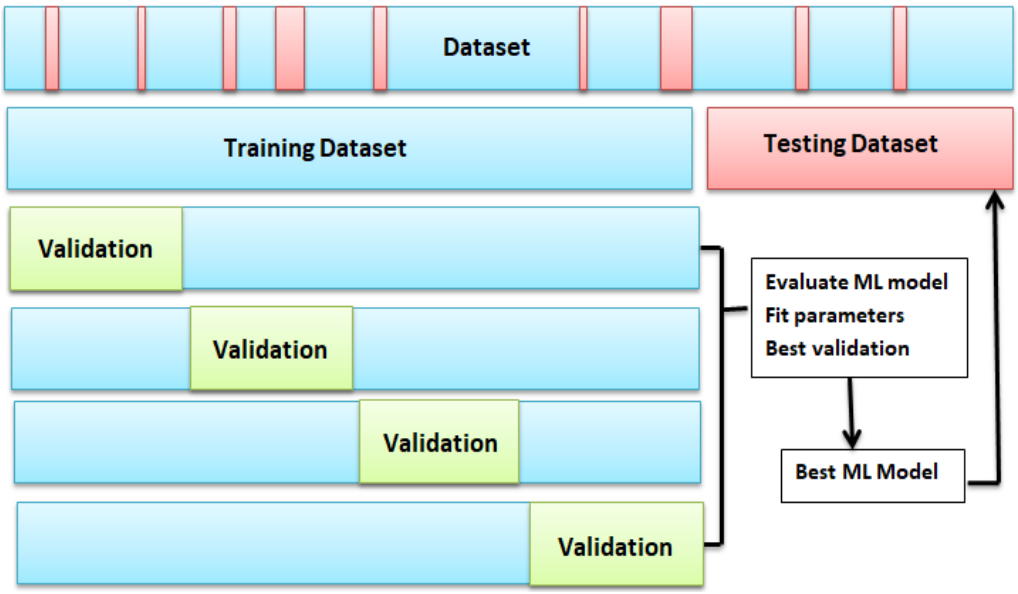

Figure 7: Illustration of the split of the dataset into two sets: the first set is for training, and the second set is for testing the machine learning model.

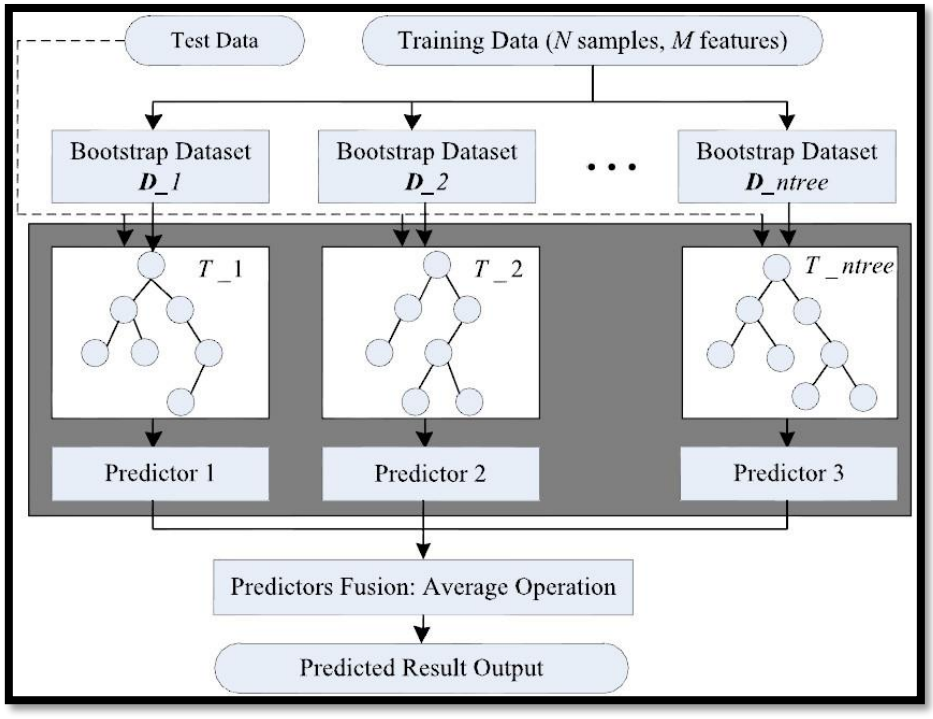

Figure 8: Structural diagram of random forests.

The out-of-bag (OOB) data is used to assess the performance of the regression decision tree. This built-in cross-validation mechanism allows the Random Forest (RF) to objectively estimate the generalization error without requiring external data.

As a result, the prediction error for a given regression decision tree can be evaluated using the OOB data [23].
Ultimately, with the OOB data acting as the optimization target, the RF model can be trained by progressively increasing its depth.
Generally, by implementing the bagging strategy, RF is less vulnerable to noise and outliers, thereby significantly mitigating the risk of overfitting.

2-4 Model evaluation

Evaluating the model using appropriate metrics is crucial to ensure that your model delivers accurate predictions. Evaluation metrics provide quantitative measures that assess the accuracy of a model's predictions. In the context of regression, these metrics evaluate the model's effectiveness.

The Scikit-learn library offers various evaluation metrics, with strengths and limitations, to determine how well a model fits the data. Before exploring the key evaluation metrics, it is important to understand the concept of "residuals" in relation to regression model evaluation.
It is unrealistic to expect a model to predict the exact value of a continuous variable in a regression scenario. Instead, a regression model can only predict values above or below the actual values.

Thus, the model's accuracy can only be assessed through residuals, representing the difference between the actual and predicted values. Residuals can be considered distances; the closer a residual is to zero, the better our model is at making accurate predictions.
There are four common error metrics used to evaluate and report the performance of a regression model:

1- **Mean Squared Error (MSE):** The Mean Squared Error (MSE) is a widely used metric in statistics and machine learning. It measures the average of the squares of the discrepancies between the actual values and the predicted values in a dataset. MSE is frequently employed in regression problems to assess the performance of predictive models [24]. MSE, which is calculated using an equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2 \quad (4)$$

Where: I = Number of samples in dataset, y = Actual value , $\hat{y}$ = Predicted value

2- **Root Mean Squared Error (RMSE):** RMSE is a metric used to assess the presence of large errors or deviations in a model's predictions. It helps identify whether the model has overestimated predictions (where the predicted values are significantly higher than the actual values) or underestimated predictions (where the predicted values are lower than the actual values) [24].
RMSE quantifies how closely the model's predictions align with the true values, with lower RMSE values indicating better model performance. If large errors are a concern, RMSE is a particularly effective metric to use.
Since the residuals are squared in the calculation, significant overestimations or underestimations of predictions will lead to larger error values. As a result, RMSE provides valuable insight into the model's performance, especially in detecting substantial deviations from the actual values [24]. RMSE, which is calculated using the equation:

$$RMSE = \sqrt{MSE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2} \quad (5)$$

If large errors are a concern, RMSE is a useful metric to be considered. The squaring of residuals in its calculation means that any significant overestimations or underestimations in predictions will lead to larger error values.
Consequently, RMSE offers important insights into the model's performance, especially when it comes to identifying substantial deviations from actual values [24].

3- **Mean Absolute Error (MAE):** The Mean Absolute Error (MAE) is a widely used metric in statistics and machine learning. It quantifies the average absolute differences between actual and predicted values in a dataset [24].
MAE offers a clear interpretation of model accuracy, reflecting the average magnitude of errors without considering their direction.
It is important to note that low MSE, RMSE, and MAE values indicate that the model is making accurate predictions. In contrast, higher MAE values imply that the model is underperforming in its predictions. MAE, which is calculated using an equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (6)$$

4- **Coefficient of Determination (R^2):** The R^2 score is a metric that assesses the performance of the model. Unlike loss metrics that indicate the magnitude of errors in absolute terms, R^2 provides insight into the proportion of variance in the dependent variable that the independent variables in the model can explain.
A higher R^2 value indicates better model performance in terms of fitting the data. R^2 which is calculated using an equation:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (7)$$

RSS = sum of squares of residuals $RSS = \sum(y_i - \hat{y}_i)^2$
TSS= total sum of squares $TSS = \sum(y_i - \overline{y})^2$
Unlike MAE and MSE, which are context-dependent, the R^2 score is independent of the specific context of the data. This characteristic makes R^2 a useful baseline for comparing models, offering insights that other metrics may not provide.
Similarly, a fixed threshold of 0.5 is commonly employed for decision-making in classification tasks. Essentially, R^2 assesses how much better the regression line performs compared to a mean line.
So, how should the R^2 score be interpreted? An R^2 score of zero indicates that the regression line is identical to the mean line, implying that 1 minus 1 equals zero. In this scenario, the two lines overlap, reflecting the model's poorest performance, as it fails to utilize information from the output variable.
On the other hand, an R^2 score of 1 signifies that the regression line perfectly predicts the data, resulting in a division term of zero. However, obtaining a perfect R^2 score is unrealistic in real-world situations.
Therefore, we can infer that as the regression line nears perfection, the R^2 score approaches one, indicating enhanced model performance. Typically, the R^2 score will range between zero and one; for instance, an R^2 score of 0.8 means that your model explains 80% of the variance in the data [25].
Table 1 shows the results of performance measures, MAE, MSE, RMSE and R2 for the predictions.

Table 1: shows performance measures values of MAE, MSE,RMSE and R^2.

Model	MSE	RMSE	MAE	R^2
Random Forest	4.783	2.187	1.351	0.893

3- Results and Discussion

The random forest model is beneficial in datasets with more than five features. Our dataset has six features, so the random forest is used.

In this study, a random forest model was developed, recognized as one of the most widely adopted techniques in machine learning. The training utilized data from nine wells, with the tenth well designated for blind testing.

Cross plots for all features within the dataset are necessary to illustrate an unclear distribution pattern. The data points are clustered and stacked within a limited range, primarily due to outlier values, which can compromise the effectiveness of both training and testing, ultimately resulting in inaccurate predictions.

Figure 9 presents the actual DT log (shown by the black line) alongside the predicted DT log (depicted in red) for Well B-5. Although the prediction results are generally favourable, some qualitative discrepancies are evident at specific depths, particularly between 2377–2380 meters and 2428–2438 meters.

These discrepancies underscore the necessity for both quantitative and qualitative data in Well A-5 at certain depths, suggesting that the accuracy of the algorithm may not be sufficient for reliable predictions of DT logs in this well at those specific depths.

Identifying a considerable number of outliers in this study necessitated the removal of a significant portion of the data, ultimately impacting the DT log prediction model.

Implementing the outlier elimination technique is a critical step in the data preprocessing phase, resulting in a marked decrease in the dataset size. Initially, the sample comprised 14,747 records; following outlier removal, the dataset was reduced to 8,719 samples.

The machine learning model exhibited strong performance metrics during the evaluation phase, including MSE, RMSE, MAE, and R^2. The model demonstrated significant sonic log prediction (DT) potential, achieving MSE = 4.783, RMSE = 2.187, MAE = 1.351, and R^2 = 0.893.

The correlation coefficient between the actual and predicted DT logs was r = 0.99, as shown in Figure 10. Overall, the quantitative analysis illustrated in Figures 10 and 11 indicates that the predictions are accurate, with both logs closely aligned.

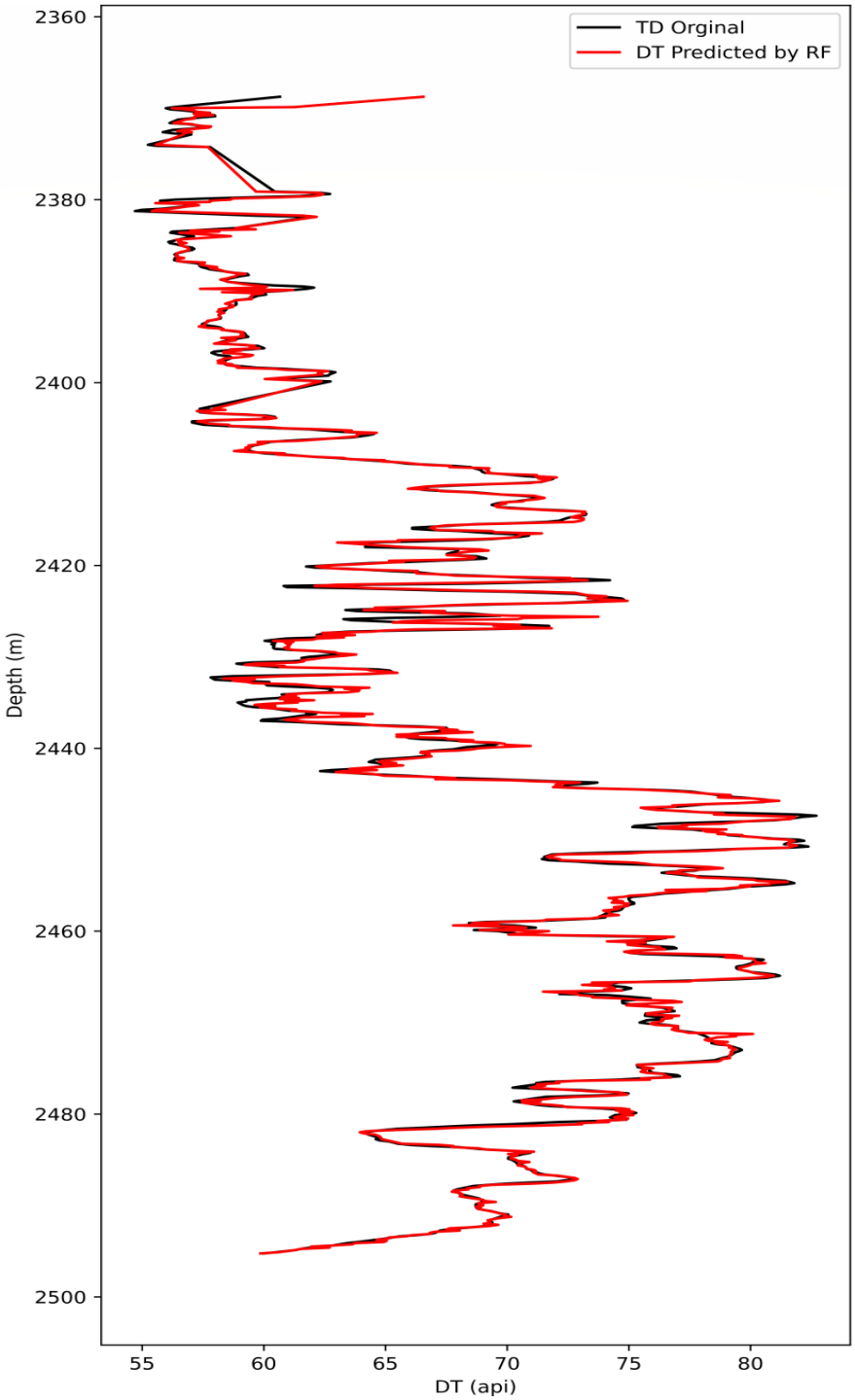

Figure 9: Shows the actual DT log represented as a black line, and the predicted DT log by the random forest model represented by a red line in well A-5 illustrated by Google Colab.

4- Conclusion

We have developed a data-driven machine learning model for predicting DT logs using data from the XXXXX in Basra Province, southern Iraq.

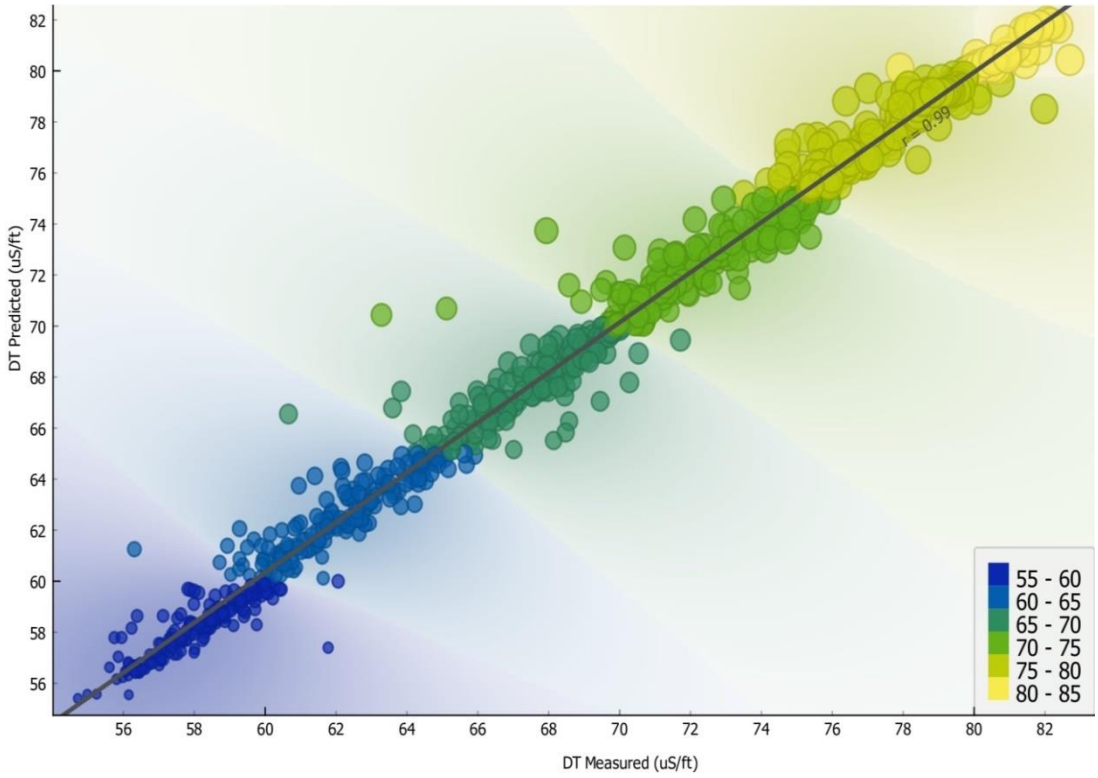

Figure 10: Scatter plot with the actual and predicted DT log in well A-5 and the correlation coefficient was 0.99 illustrated by orange data mining .

The Random Forest model produced reliable results on both the testing dataset and a blind well test.

The effectiveness of the proposed method for DT log prediction was assessed, revealing enhanced accuracy as reflected by metrics such as MSE, RMSE, MAE, and R^2. Nonetheless, the training of the machine learning algorithms faced challenges due to data quality and quantity limitations, which likely contributed to inaccuracies in DT log predictions, particularly at specific depths within the wellbore.

The findings suggest that alternative strategies for managing missing information are essential to enhance model accuracy. Deletion techniques to remove outliers from the dataset have drawbacks, including a reduced sample size that negatively impacts the quality and quantity of the available data.

This, in turn, affects the training of machine learning models and contributes to the inaccuracies observed in DT log predictions at certain depths.
Overall, the results of this study highlight the significant potential of machine learning in DT log prediction. They also underscore the need for continued methodological advancements in data processing and improvements in algorithm efficiency for future geophysical applications.

References

[1] H. Jian, L. Chenghui, G. Zhimin, M. Haiwei, "Integration of deep neural networks and ensemble learning machines for missing well logs estimation", *Journal of Flow Measurement and Instrumentation*, Elsevier, vol.73, no.101748, 2020, https://doi.org/10.1016/j.flowmeasinst.2020.101748.

[2] D. Onalo, S. Adedigba, O. Oloruntobi, Khan, F., L. A. James, S. Butt, "Data-driven model for shear wave transit time prediction for formation evaluation", *Journal of Petroleum Exploration and Production Technology*, vol.10, pp. 1429–1447, 2020, https://doi.org/10.1007/s13202-020-00843-2.

[3] J. Pendrel, and S H. chouten, "Facies—The drivers for modern inversions", *The Leading Edge*, vol.39, no.2, pp. 102-109, 2020, https://doi.org/10.1190/tle39020102.1.

[4] A. K. A. Mohammed, M. K. Dhaidan, "Prediction of Well Logs Data and Estimation of Petrophysical Parameters of Mishrif Formation, Nasiriya Field, South of Iraq Using Artificial Neural Network (ANN)", *Iraqi Journal of Science*vol.64, no.1, pp. 253-268, , 2023, DOI: 10.24996/ijs.2023.64.1.24.

[5] A. Al Ghaithi, and M. Prasad, "Machine learning with artificial neural networks for shear log predictions in the volve field norwegian north sea", *Society of Exploration Geophysicists*, 2020, doi: 10.1190/segam2020-3427540.1.

[6] F. F. Melo, V. Perez, C. S. Lee, "Shear sonic log prediction with Deep Neural Networks: an example from Gulf of Mexico", *Conference Geoconvention-June 20-22*, University of Calgary-Canada, 2022.

[7] A. Hakam, W. Utama, S. A. Garini, O. A. Jabar, A. Nurdien, F. Indsni, Y. Rosandi, "Sonic Log Prediction Based on Extreme Gradient Boosting (XGBoost) Machine Learning Algorithm by Using Well Log Data", *BIO Web of Conferences*, 2024, https://doi.org/10.1051/bioconf/20248909003.

[8] B. A. Al-Baldawi, and M. E. Nassir, "Evaluation of Petrophysical Characteristics of Carbonate Mishrif Reservoir in Ahdeb oil Field, Central Iraq", *Iraqi Journal of Science*, vol. 60, no.2, pp.321-329, 2019, DOI: 10.24996/ijs.2019.60.2.12.

[9] A. S. Al-Banna, N. A. Nassir, G. H. Al-Sharaa, "A Comparison of Mishrif Formation Characteristics in Kumait and Dujaila Oil Fields, Southern Iraq, Using Seismic Inversion Method and Petrophysical Properties Analysis", *Iraqi Journal of Science*, vol. 61, no. 12, pp.3294-3307, 2020,DOI: 10.24996/ijs.2020.61.12.17.

[10] T. A. Mahdi, A. M. A. Aqrawi, A. D. Horbury, G. H. Sherwani, "Sedimentological characterization of the mid-Cretaceous Mishrif reservoir in southern Mesopotamian Basin, Iraq", *GeoArabia*, v.18, no.1, pp.139-174, 2013, DOI: 10.2113/geoarabia1801139.

[11] A. A. M. Aqrawi, J. C. A. D. Goff, Horbury, F. N. Sadooni, "The Petroleum Geology of Iraq", 1st Edition, *Scientific Press*", p. 424, 2010.

[12] P. R. Sharland, R. Archer, D. M. Casey, R. B. Davies, S. Hall, A. Heward, A. Horbury, M.D. Simmon, "Arabian Plate Sequence Stratigraphy", *Geo Arabia*, pp.371, 2001, DOI:10.2113/geoarabia0901199.

[13] N. Silaparasetty, "Machine Learning Concepts with Python and the Jupiter Notebook Environment: Using TensorFlow 2.0", *Distributed by Springer Science + Business Media New York*, 233 , 2020, https://doi.org/10.1007/978-1-4842-5967-2.

[14] A. Al-Khafaji, M. H. Hakimi, I. El-Khedr, A. A. Najaf, H. Al Faifid, A. Lashin, "Organic geochemistry of oil seeps from the Abu-Jir Fault Zone in the Al-Anbar Governorate, western Iraq: Implications for early-mature sulfur-rich source rock", *Journal of Petroleum Science and Engineering*, vol.184, pp.106584, 2020, https://doi.org/10.1016/j.petrol.2019.106584.

[15] Y.,N. Pandey, A. Rastogi, S. Kainkaryam, S. Bhattacharya, L. Saputelli, "Machine learning in the oil and gas industry", 1st Edition, *Apress Berkeley*, 2020, p. 300, eBook ISBN 978-1-4842-6094-4, https://doi.org/10.1007/978-1-4842-6094-4

[16] F. Neli, "Python Data Analytics", 2nd Edition, *Apress Berkeley*, 2018, p. 576, eBook ISBN 978-1-4842-3913-1 DOI:https://doi.org/10.1007/978-1-4842-3913-1

[17] H. Belyadi, A. Haghighat, , "Machine Learning Guide for Oil and Gas Using Python: A Step-by-Step Breakdown with Data, Algorithms, Codes, and Applications", *imprint by Gulf Professional Publishing, Published by Elsevier*, 2021, p. 462, ISBN 978-0-12-821929-4 DOI:https://doi.org/10.1016/C2019-0-03617-5

[18] M. Petrelli, , "Machine Learning for Earth Sciences Using Python to Solve Geological Problems",1st Edition, *published by the Springer cham*, 2023, p. 209, https://doi.org/10.1007/978-3-031-35114-3.

[19] P. Deital, H. Deital, , "Python for programmers with introductory AI case studies", *Publisher(s): Pearson*, 2019, p. 641, ISBN 9780135231364 DOI: 10.24996/ijs.2022.63.2.16.

[20] H. V. Thanh, A. Zamanyad, M. Safaei-Farouji, U. Ashraf, Z. Hemeng, "Application of hybrid artificial intelligent models to predict deliverability of underground natural gas storage sites", *Renew Energy, Elsevier* .vol.200, pp.169–184, 2022, https://doi.org/10.1016/j.renene.2022.09.132.

[21] M. Alghazal and D. Krinis, "A novel approach of using feature- based machine learning models to expand coverage of oil saturation from dielectric logs", *Conference: SPE Europec featured at 82nd EAGE Conference and Exhibition*, Amsterdam, Netherlands, 2021, https://doi.org/10.2118/205162-MS.

[22] L. Briman, J. Friedman, R.A.Olshen, C.J Stone. 2017, Classification and Regression Trees, 1st Edition, eBook Published in New York, eBook: ISBN9781315139470, pp.368, https://doi.org/10.1201/9781315139470.

[23] L. Breiman, "Bagging Predictors, Machine Learning", *Article research in springer* , vol.24, pp.123-140, 1996.

[24] A. Geron, "*Hands-on Machine Learning with Scikit-Learn, Keras, and Tensor Flow Concepts, Tools, and Techniques to Build Intelligent Systems"*, 2nd Edition, Published by O'Reilly Media, 2019, pp.848,

[25] J. M. Wooldridge," *Introductory Econometrics: A Modern Approach"*, 5th Edition, Published by south-western cengage learning, eBook:ISBN- 1-111-53104-8, 978-1-111-53104-1, 2013, pp.878.