Development and Evaluation of an Effective Machine Learning Model for Well Log Prediction: A Case Study of Sonic Log Prediction of Zircon Field Niger-Delta Nigeria

Chukwukelu Johnpaul Odiegwu[1*], Ifeanyi Augustine Chinwuko[1], Emmanuel Udenwa Aniwetalu[1], Maureen Chioma Umeh [1], Callista Nwanneka Igwebudu[1], Chinaza Janefrances Akpudo[1]

[1] Department of Applied Geophysics, Nnamdi Azikiwe University, Awka 5025, Nigeria

Corresponding author: cj.odiegwu@unizik.edu.ng

Abstract

The accurate prediction of sonic log data is critical for subsurface characterization and reservoir management in hydrocarbon exploration. Conventional methods of predicting missing well logs which often relied on interpolation techniques or empirical correlations are limited in their ability to capture the complex, nonlinear relationships that exist in subsurface formations. In this study we present a methodology for predicting a missing log. Three wells from the Zircon field in the Niger-Delta were used in the study: Well 7, 1, and 6 for training, validation and prediction phases respectively. The results of the preprocessing steps which involved outlier removal, missing value handling and filtering operation with Butterworth lowpass filter, were effective in improving the correlation among predictor variables and target Sonic log. Five models were initially used in the training and validation phases using the Sci-kit ML module in Python. The RF model was finally selected for the prediction phase having out-performed other models with a RMSE, MEDAE and R-SQUARED SCORE values of 2.5886μs/ft, 1.0380μs/ft, 0.9642 in the testing phase and 6.5588μs/ft, 3.6356μs/ft and 0.7695 in the validation phase respectively. A supplementary qualitative well correlation analysis performed using the training, validation and prediction wells gave satisfactory results based on similarities in the sonic log character and trend. The qualitative well correlation provided a crucial geological validation of the model's output. The findings of this research could significantly reduce the need for extensive well logging operations and provide a framework for integrating machine learning techniques into petroleum geoscience.

Keywords: Random forest, sci-kit learn, geophysics, geoscience, hydrocarbon exploration
MSC: 74Q15; 60G25

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 20, 2025; Revised: October 21, 2025; Accepted: October 24, 2025

Introduction

Well logging involves measurement of subsurface characteristics that are inputs in the computation of petrophysical parameters relevant to hydrocarbon accumulation and producibility such as p-wave velocity, porosity, permeability, water saturation, shale volume, hydrocarbon saturation etc. Well logs provide a comprehensive set of data that geoscientists use to make informed decisions about

reservoir development and management. The productivity of wells in hydrocarbon-bearing reservoirs depends on petrophysical properties which include lithology, porosity, water saturation, permeability, and saturation (Eke et al., [6]).

Sonic logs are indispensable for comprehensive reservoir characterization, providing critical data for porosity calculations, seismic-to-well ties, lithology discrimination, and the determination of rock mechanical properties (Alfaraj et al. [2]). However, the acquisition of reliable sonic log data is not always guaranteed and can be frequently compromised or entirely missing owing to several factors such as operational challenges, faulty instruments, economic constraints, legacy data etc.

Traditional methods of predicting missing well logs have often relied on interpolation techniques or empirical correlations based on other well logs. While these methods can provide approximate predictions, they are limited in their ability to capture the complex, nonlinear relationships that exist in subsurface formations. This is particularly true in regions like the Niger Delta, where geological complexity and heterogeneity make accurate log prediction challenging.

The intersection of machine learning and geoscience, particularly in the context of sonic log predictions, has witnessed significant explosion. The advent of machine learning (ML) offers a transformative approach to this problem. ML's ability to discern patterns and make predictions from large datasets is particularly suited to the multifaceted nature of geological data (Liu et al. [11]; Kouadio et al. [10]). Several studies have demonstrated the effectiveness of machine learning models in predicting well logs, particularly in cases where logs are incomplete or missing.

Tectonic and geologic framework of the study area

The study focuses on the Zircon field, located within the Niger Delta petroleum province of Nigeria. The Niger Delta, situated on the passive continental margin of the Gulf of Guinea in equatorial West Africa, is one of the world's most prolific hydrocarbon provinces, with intensive exploration and exploitation activities ongoing since the discovery of commercial oil in 1956 (Doust and Omatsola [7]). The basin's formation is linked to the Cenozoic development of a large, arcuate delta system built by the Niger and Benue Rivers, resulting in a thick sedimentary succession reaching up to 12 km2 This succession is broadly divided into three main diachronous lithostratigraphic units, from oldest to youngest: the Akata, Agbada, and Benin Formations (Ajaegwu et al., [1]).

The Akata Formation, Paleocene to Recent in age, consists primarily of dark grey, commonly fissile, marine shales, which are often sandy or silty. Intercalated within these shales are beds of turbidite sandstones, siltstones, and clays. A key characteristic of the Akata Formation is that it is typically overpressured, a result of rapid sediment burial by the overlying Agbada Formation, which inhibited normal dewatering and compaction (Turtle et al., [15]). Overlying the Akata Formation is the Eocene to Recent Agbada Formation, which is the major petroleum-bearing unit in the Niger Delta (Ajaegwu et al., [1]). It comprises paralic siliciclastics, representing the main deltaic sequence of alternating sandstones and shales, deposited in delta-front, delta-topset, and fluvial-deltaic environments. The youngest unit, the Oligocene to Recent Benin Formation, consists mainly of continental fluvial sands and gravels, with minor shale intercalations, reaching thicknesses of up to 2,100 meters (Ajaegwu et al., [1]).

The structural style of the Niger Delta is dominated by syn-sedimentary deformation, characterized by extensive gravity-driven faulting and folding, often facilitated by detachment on under-compacted, over-pressured shales of the Akata Formation. This has led to the formation of complex structural and stratigraphic traps, including roll-over anticlines associated with growth faults, fault closures, and subtle stratigraphic traps (Weber, [17,18]).

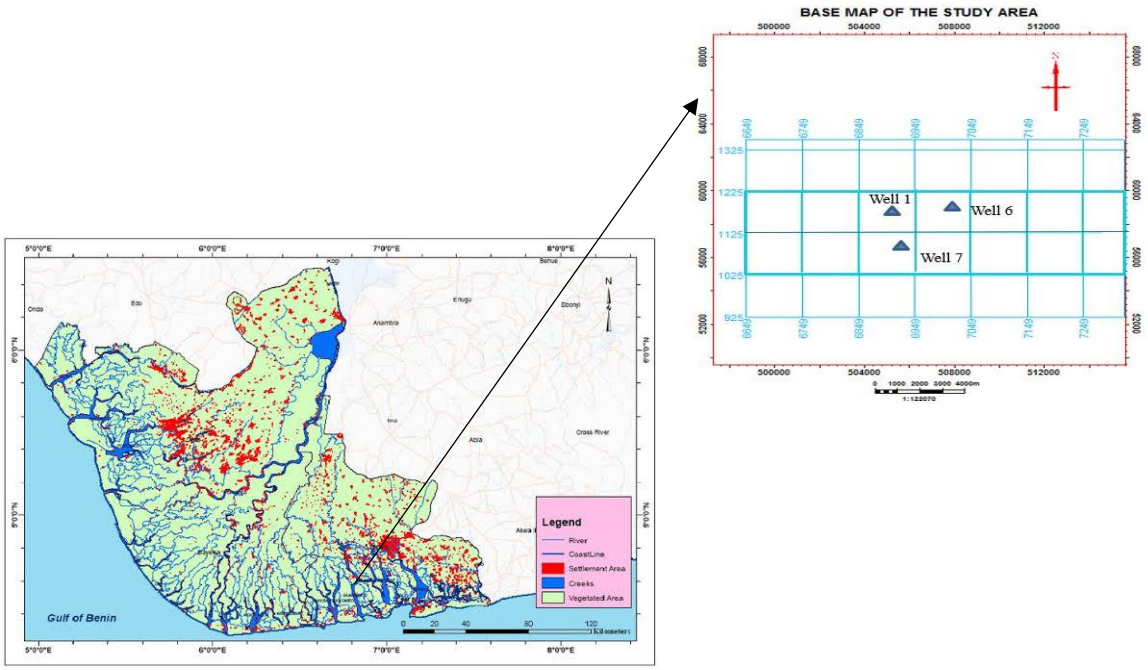

Figure 1: Map of the study area

Materials and Methods

Materials

The dataset for this study comprises well log data from three wells within the Zircon field. For each well, a suite of conventional logs was available, these include Gamma Ray (GR) in API units, Sonic log in microseconds per foot (µs/ft), Bulk Density (RHOB) in grams per cubic centimeter (g/cm³), Caliper (CALI) in inches, and Resistivity (LLD9) in Ohm-meter (Ω-m). The distance between Well 1 and Well 6 is 1.10km while the distance between Well 1 and Well 7 is 1.04 km and the distance between Well 6 and Well 7 is 1.20km

The three wells were partitioned for the machine learning workflow as follows:

Well-7 and Well-1: These wells had a complete suite of logs, including the target sonic log. Data from these wells were used for training and validating the machine learning models respectively.

Well-6: This well lacked the sonic log and was therefore designated as the prediction or blind test well, where the optimized ML model would be applied to generate a synthetic sonic log.

A summary of the well log data utilized is presented in Table 1.

Table 1: Summary of available well logs for the study.

Well	Available Logs	Depth Interval (ft)	Purpose
Well-7	GR, SONIC, RHOB, CALI, LLD9	3434.0 - 9926.5	Training
Well-1	GR, SONIC, RHOB, CALI, LLD9	4010.0 -9969.5	Validation
Well-6	GR, RHOB, CALI, (SONIC missing)	4018.5 -9918.0	Prediction

The entire data processing, model development, and evaluation workflow were implemented using the Python programming language. The following key open-source Python libraries were instrumental:

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

Pandas: Utilized for efficient data handling, including loading well log data into DataFrame structures, data manipulation, and initial statistical exploration (McKinney, [12]).

NumPy: Employed for fundamental numerical computations, particularly for operations on arrays which form the basis of log data handling and ML model inputs (Harris et al., 2020).

Matplotlib and Seaborn: These libraries were used for creating interactive visualizations, including log plots, histograms for data distribution analysis, boxplots for outlier detection, and crossplots for examining relationships between different logs (Hunter, [9]; Waskom, [16]).

Scikit-learn (sklearn): This comprehensive machine learning library was the cornerstone for the ML aspects of the study. It provided tools for data preprocessing, model implementation, and performance evaluation.

Methodology

The methodology used in this study follows a step-by-step approach for exploratory data analysis, data preprocessing, model training, validation, and prediction as shown in the diagram below.

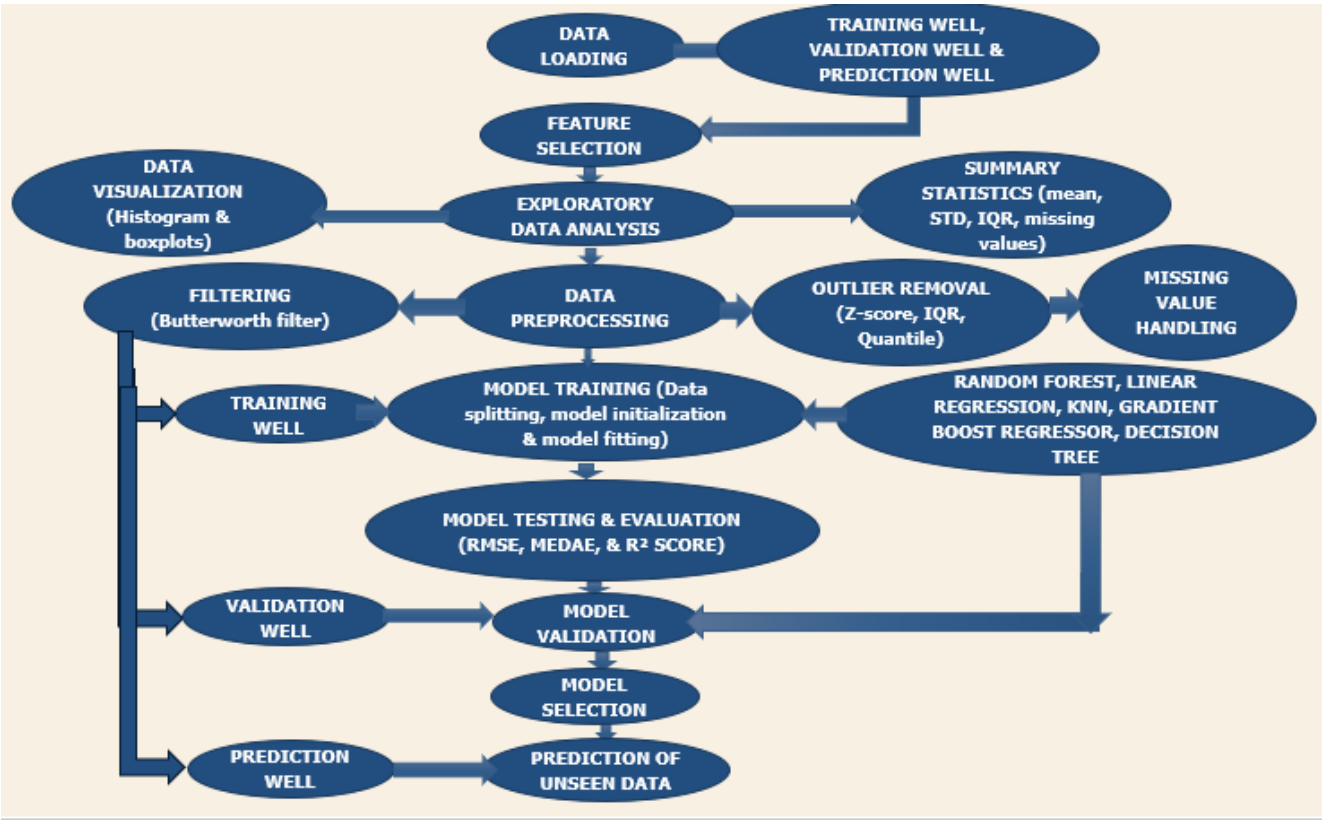

Figure 2: Workflow for the study.

Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution and relationships among the well logs. EDA in this study included the following components.

Summary Statistics

Calculation of descriptive statistics (mean, median, standard deviation, minimum, maximum, and quartiles) for each log curve in all the wells provided a quantitative understanding of the data distributions, typical value ranges, and potential presence of outliers prior to explicit outlier treatment. Histograms were also generated to visualize these distributions. A correlation coefficient was also generated to examine the strength of relationships among the well logs using heatmap.

Missing value handling

Well log data can often contain missing values or gaps due to various operational and technical issues like tool failure, cost considerations or sections where logging was not possible. Missing values can impede reservoir studies and if not properly handled can lead to biased machine learning models and increase uncertainty in results. Missing values were handled through a combination of imputation techniques and, in some cases, by removing incomplete records if they were determined to be non-informative.

Outlier detection and removal

Outliers are data points that significantly deviate from the overall pattern of the data set. They can distort the results and adversely impact the performance of the models leading to incorrect interpretations. Boxplots were generated for each input log (GR, RHOB, NPHI, CALI) and the target log (SONIC) for all the wells to visually identify potential outliers.

The choice of the outlier detection and removal method was informed by the nature of data distribution as revealed by the histogram plots. The specific method and thresholds for outlier treatment were chosen on a per-log basis. Three commonly outlier removal methods were used:

a. Z-Score method
b. Interquartile Range (IQR)
c. Quantile method

Z-Score method

Also called Standard score often used for dataset with a normal distribution. A threshold is set for the upper and lower limits within which the data points are expected to lie. The upper and lower limit were set as follows:

$$Upper_{limit} = \bar{x} + 3\sigma \quad (1)$$

$$Lower_{limit} = \bar{x} - 3\sigma \quad (2)$$

Where:
$\bar{x}$ = Mean of data distribution
σ = Standard deviation

Any data outside of this range is considered an outlier and is replaced or capped by the value of either the upper limit or lower limit.

Inter-quartile range (IQR) method

Normally used for skewed distribution. The IQR is a statistical measure used to assess variability within a dataset. It divides the data into quartiles, capturing the middle 50% of the observations. Specifically, it focuses on the range between the first quartile (Q1) and the third quartile (Q3). The upper and lower boundaries were also set as follows:

$$Upper_{limit} = Q3 + 1.5IQR \quad (3)$$

$$Lower_{limit} = Q1 - 1.5IQR \quad (4)$$

Where Q1 and Q3 are the 25th and 75th percentile of the dataset, respectively. IQR represents the inter-quartile range and is given by Q3−Q1.

Outliers were defined as points falling below the lower limit or above the upper limits. These points were removed or capped at these boundary values.

The Quantile or Percentile-based approach

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

In some cases, particularly for logs with extreme values at the tails such as the LLD9 log, fixed percentile-based capping was used to mitigate the influence of extreme outliers without excessive data removal. For this study the commonly used thresholds of 10th percentile (0.1 Quantile) and 90th percentile (0.9 Quantile) were used for the lower and upper limits respectively.

Data filtering

Low-pass filtering is a signal processing technique that allows low-frequency components of a signal to pass through while attenuating high-frequency components. In other words, it smooths out rapid variations or noise in the data while preserving the overall trend.

Filtering operation using Butterworth low pass filter was used in removing noise in the well log for better visualization and interpretability. This was implemented using the Scipy library in Python. Filter parameters were carefully chosen in other not to over filter the data and remove genuine geological variations. The filtered logs were validated against the raw logs and geological context by direct comparison and correlation.

Data normalization

The next preprocessing flow involved log standardization or optimization to remove systematic errors so that reliable results could be obtained from the machine prediction. This was achieved using the standardization equation (5) (Codd, [5]) below implemented in the Python programming language using the StandardScaler() function.

$$z = \frac{x_i - \bar{x}}{\sigma} \tag{5}$$

Results and Discussions

Summary statistics

All wells cover a significant depth range, generally from around 3400-4000 ft to approximately 9900 ft. This indicates that the wells are exploring a substantial stratigraphic column within the Niger Delta basin. The summary statistics of the three wells are displayed in tables 2,3 and 4.

Table 2: Summary statistics of training well (Well-7)

STATISTIC	DEPTH	CAL	LLD9	RHOB	GR	SONIC
Count	12986	12960	12939	12960	12939	12939
Mean	6680.25	12.4982	27.0358	2.1604	56.5479	111.8373
STD	1874.4398	1.0339	43.2074	0.1022	23.9306	13.7102
Min	3434	11.3906	0.0696	1.434	26.702	27.8
25%	5057.125	12	1.48825	2.1106	36.375	102.9
50%	6680.25	12.1406	8.0234	2.1497	45.0523	110.9
75%	8303.375	12.3906	31.9575	2.2046	81.3605	120.4
Max	9926.5	21.125	1876.614	2.5701	120.1977	187.2

Table 3: Summary statistics of validation well (Well-1)

STATISTIC	DEPTH	CAL	LLD9	RHOB	GR	SONIC
Count	11920	11920	11920	11920	11920	11920
Mean	6989.75	12.1416	13.5535	2.2042	58.6579	109.9341
STD	1720.5760	1.5955	14.9909	0.0797	22.2230	12.7340
Min	4010	10.313042	-1.594149	2.06769	26.368276	87.552851
25%	5499.875	10.841853	1.666707	2.139616	39.375248	99.222328
50%	6989.75	11.591681	6.815229	2.186813	52.164395	107.30806
75%	8479.625	13.131713	20.974863	2.256116	77.747725	119.817335
Max	9969.5	17.169679	50.814247	2.389681	105.199425	150.754426

Table 4: Summary statistics of prediction well (Well-6)

STATISTIC	DEPTH	CAL	LLD9	RHOB	GR
Count	11793	11793	11793	11793	11793
Mean	6969.898796	12.931096	16.141283	2.175738	58.804504
STD	1702.41542	1.084508	18.881235	0.062719	22.827318
Min	4018.5	11.926034	-3.569161	2.069205	24.317694
25%	5496	12.150969	1.364376	2.129316	37.562266
50%	6970	12.338398	6.176394	2.160316	53.52101
75%	8444	13.450124	28.161993	2.218303	79.376749
Max	9918	15.768325	60.791391	2.333862	102.195299

Resistivity (LLD9)

Well-7

The Mean resistivity value in this well is 27.0358Ωm while STD is 43.2074, suggesting a wide range of resistivities. The Minimum resistivity value of 0.0696Ωm indicates highly conductive zones, likely shales or water-bearing sands with high salinity. While the maximum value 1876.614Ωm indicates very high resistivity zones. The large range between 25th percentile (1.48825) and 75th percentile (31.9575) confirms significant lithological and fluid variations.

Well-1

The Mean resistivity value in this well (13.5535Ωm) is lower than that of Well-7, suggesting potentially less resistive overall formations. This well also has a lower STD value (14.9909), indicating less variability in resistivity. The negative minimum value of -1.594149Ωm is physically impossible, suggesting a bad data point or an outlier. This was removed using the Percentile-based capping method of outlier removal. The Maximum value of 50.814247 Ωm which is significantly lower than that of Well-7, further supports the idea of fewer or less resistive zones.

Well-6

The Mean resistivity value of 16.141283 Ωm which is higher than Well-1 but lower than Well-7, suggests an intermediate resistivity profile. The STD value (18.881235) is also intermediate, implying some variability in resistivity. The negative Minimum value -3.569161 Ωm also suggests a bad data point or an outlier. The Maximum resistivity value of 60.791391 Ωm which is higher than Well-1 but still significantly lower than Well-7, suggests some resistive zones intermediate between Well-7 and Well-1

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

Bulk density (RHOB)

Reflects formation density, typically lower in porous sands and higher in shales and denser lithologies.

Well-7

The Mean RHOB value of (2.1604g/cm^3) and STD value of (0.1022) indicate a relatively consistent density range. While the Minimum (1.434g/cm^3) and Maximum (2.5701g/cm^3) suggest presence of very porous sands (low density) and denser formations/shales (high density). The range suggests a good mix of sand and shales.

Well-1

The Mean RHOB value (2.2042g/cm^3) in this well is slightly higher than Well-7, suggesting a slightly denser overall formation, possibly more shales or compacted sands while the STD value of 0.0797 is lower than Well-7, indicating less variation in density. The Minimum (2.06769g/cm^3) and Maximum (2.389681g/cm^3) show a narrower range, confirming less variability in density, potentially less extreme porosity variations compared to Well-7.

Well-6

The Mean RHOB value (2.175738g/cm^3) in this well is intermediate between Well-7 and Well-1. While the STD value (0.062719) is the lowest among the three wells, indicating the most consistent density profile, implying less variation in porosity and lithology. The Minimum (2.069205g/cm^3) and Maximum (2.333862g/cm^3) values show the narrowest range, reinforcing the idea of less variability in density.

Gamma ray (GR)

Gamma ray readings reflect the shale content in the formation. Higher values (>50API) indicate clay-rich or shale formations, while lower values (<50 API) suggest cleaner, non-radioactive formations like sandstones.

Well-7

The Mean gamma value of 56.5479API and STD value of 23.9306 indicate a mix of shales and sands. While the Minimum value (26.702 API) suggests clean sand intervals the Maximum (120.1977API) indicates radioactive shales. The 25th percentile (36.375API) and 75th percentile (81.3605API) provide a good bracket for typical sand and shale values.

Well-1

The Mean gamma value of 58.6579API is slightly higher than that of Well-7, implying a slightly higher shale content on average. However, the STD value (22.2230) is similar, indicating a similar mix of sands and shales. The Minimum (26.368276) and Maximum (105.199425) are comparable to Well-7, confirming the presence of clean sands and shales.

Well-6

The Mean gamma value (58.804504API) is the highest among the three wells, suggesting the highest overall shale content. STD (22.827318) is similar to the other wells, indicating the presence of both sands and shales. The Minimum gamma value of 24.317694API is the lowest, indicating some very clean sand intervals. While the Maximum (102.195299API) is comparable, indicating shales.

SONIC

Measures the time it takes for a sound wave to travel through a formation, inversely related to velocity. Higher transit time (lower velocity) indicates more porous or less consolidated formations (sands), while lower transit time (higher velocity) indicates denser formations (shales).

Well-7

The Mean value of 111.8373 µs/ft and STD value of 13.7102 µs/ft show a moderate range. While the Minimum sonic value of 27.8 µs/ft is unusually low, possibly indicating a very dense or fast formation, or a data outlier/tool issue. This value is quite anomalous for typical Niger Delta sediments, which usually have sonic values well above 50-60 µs/ft.

The Maximum value of 187.2 µs/ft indicates very porous or unconsolidated intervals, typical of loose sands.

Well-1

The Mean Sonic value (109.9341 µs/ft) in this well is comparable to Well-7. While the STD (12.7340) is slightly lower than Well 1, indicating slightly less variability in porosity/consolidation. The Minimum sonic value of 87.552851 µs/ft is more realistic than that of Well-7's minimum, suggesting no extreme dense layers or data issues. While the Maximum value of 150.754426 µs/ft which is also lower than that of Well-7 maximum, imply less extremely porous or unconsolidated zones.

In summary, the geophysical log statistics are broadly consistent with the geological characteristics of a well in the Niger Delta, indicating a promising mix of potential reservoir sands and sealing shales, characteristic of a prolific hydrocarbon province.

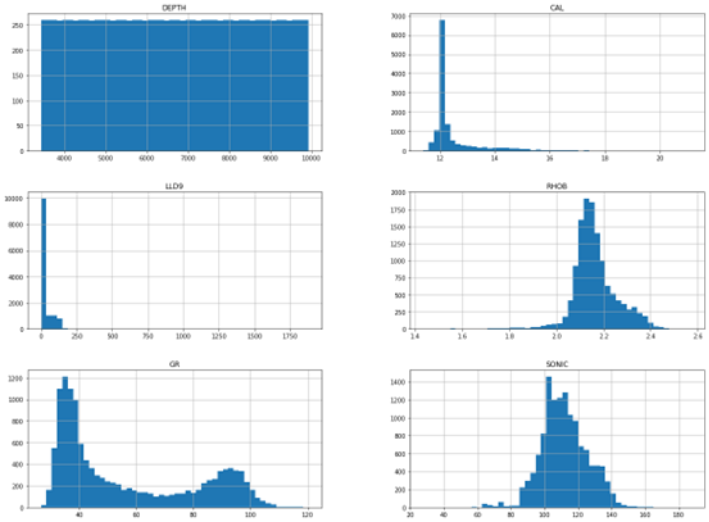

Figure 3: Histogram plot of training well (Well-7)

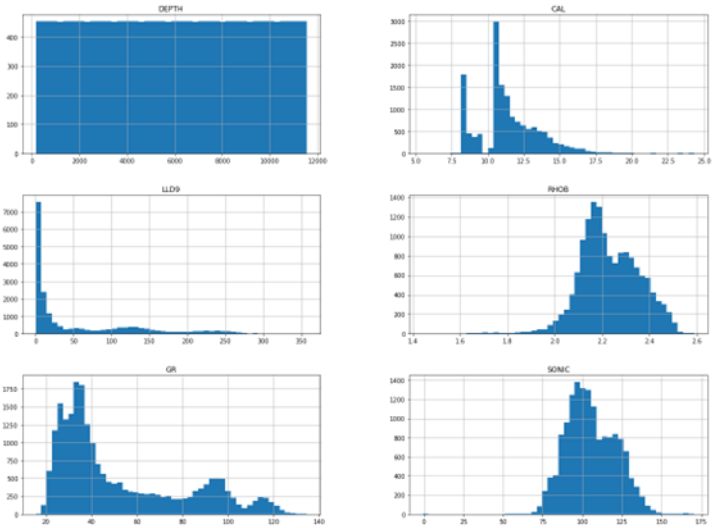

Figure 4: Histogram plot of validation well (Well-1)

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

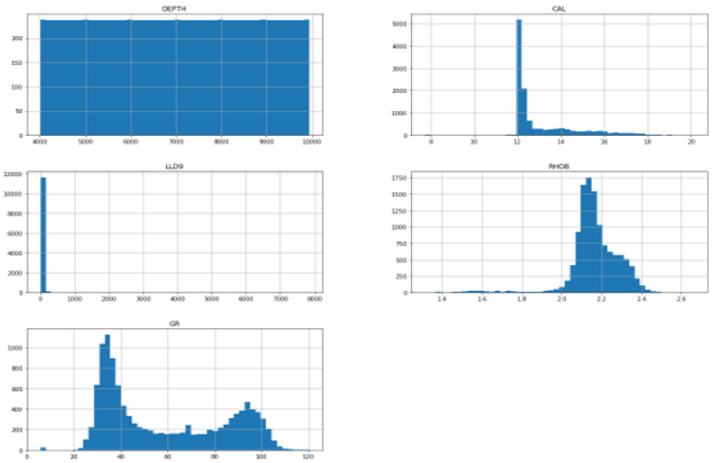

Figure 5: Histogram plot of prediction well (Well-6)

The histograms graphically reinforce the quantitative summary statistics, providing a visual confirmation of the lithological mix across the three wells in the study area. A remarkable feature of the histograms is the bimodal distribution of the gamma ray with distinct peaks. The two distinct peaks suggest the alternation between cleaner, non-radioactive formations like sandstones (lower gamma ray values) and shale-rich or clay-rich formations (higher gamma ray values). While SONIC and RHOB appear to be normally distributed with no significant outliers LLD9 shows strong peaks at low resistivity with significant outliers.

Preprocessing

The data preprocessing phase successfully addressed missing values and outliers. The Z-score method of outlier removal was effective in removing outliers in the CALIPER and SONIC logs. For the extreme values in LLD9 the IQR method gave good results while the percentile-based method showed effectiveness in removing the outliers in RHOB. This resulted in a clean dataset for model training (Figs 6a and 6b) and also an increased correlation of the different well logs feature with the target feature as revealed in the values extracted from the heatmap (table 4). Similarly, improvements were also observed in the log data after the smoothing operation (Figure 7) thus enhancing the interpretability of the well logs.

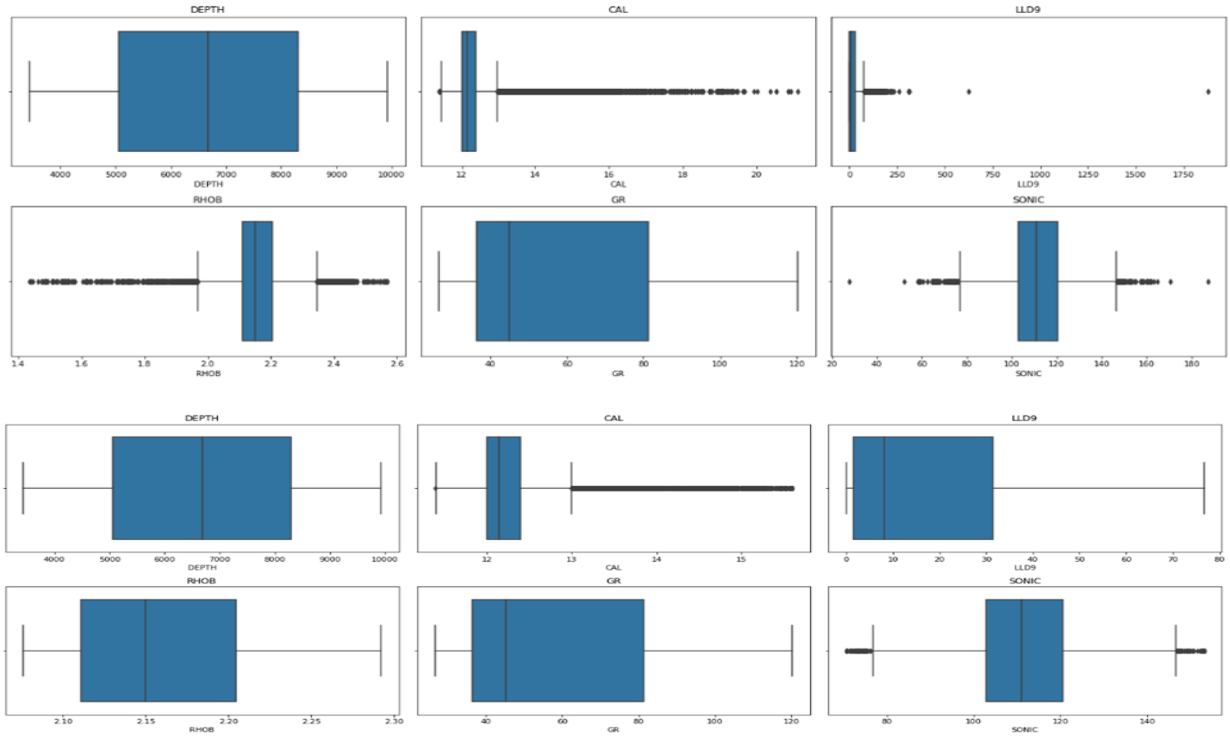

Figure 6: Box plot of training well (Well-7) (a) Before training (b) After training

Table 5: Absolute correlation of features and target (Sonic log) before and after preprocessing extracted from heatmap

Logs	**DEPTH**	**LLD9**	**RHOB**	**GR**	**CAL**
Correlation (Before preprocessing)	0.775	0.496	0.307	0.230	0.213
Correlation (After preprocessing)	0.782	0.601	0.447	0.235	0.222

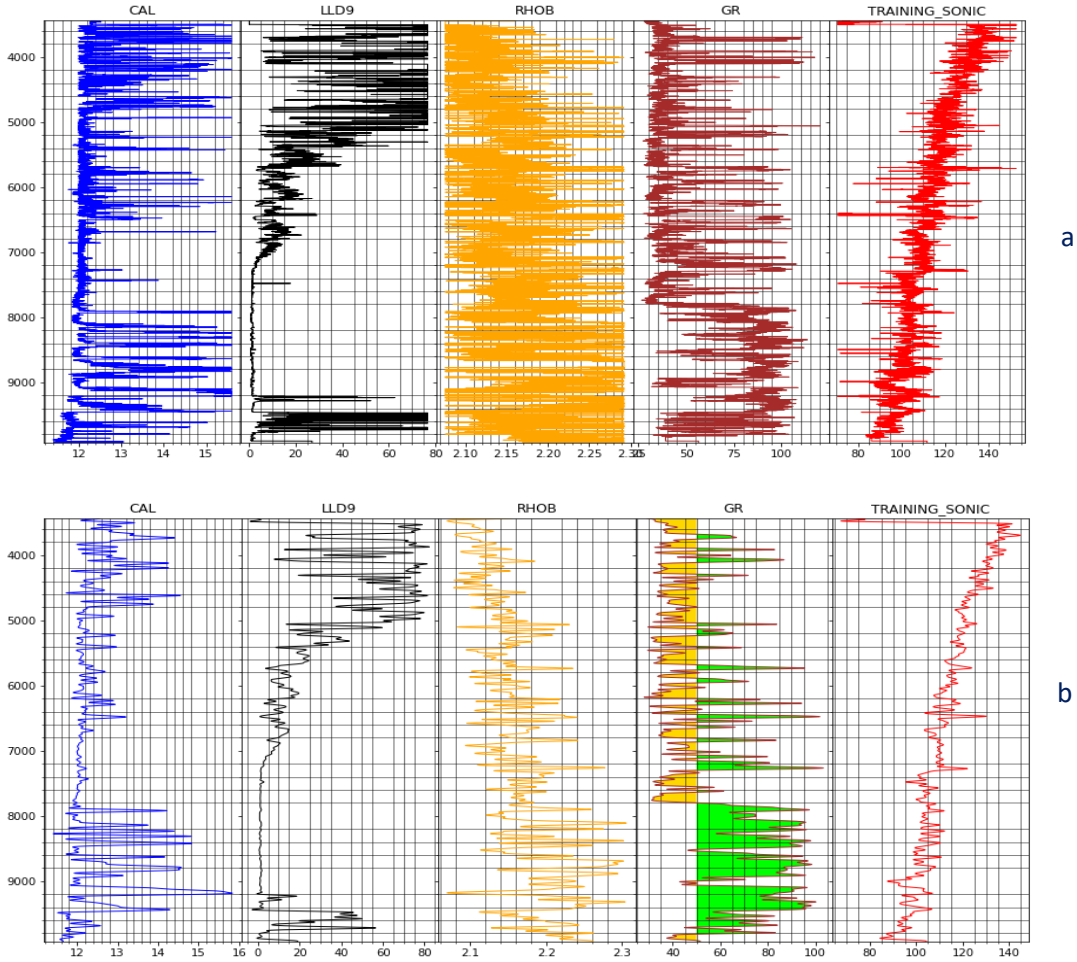

Figure 7: Result of filtering operation (a) Before filtering (b) After filtering

Model training

The five candidate models were all used during the training after the train/ test split of the training data. The hyperparameters of each model were set to the default values. The training was done by initializing and calling the model.fit() method in the scki-kit learn library.

The Random Forest model consistently demonstrated superior performance across all three-evaluation metrics (RMSE, MAE, and R^2), indicating its better capability to generalize and accurately predict sonic log values for this dataset. The performance metrics are summarized in table 6. While the scatter plots of test versus predicted sonic values are shown in Figs 8a, 8b, 8c, 8d and 8e.

Table 6: Model test performance evaluation values. Random Forest Regressor showed the lowest RMSE, MEDAE and the highest R^2, indicating a strong fit to the data

Model	**RMSE**	**MEDAE**	**R^2 Score**
Gradient Boost Regresssor	5.3604	2.2420	0.8467
Decision Tree	3.3879	1.3000	0.9388
Random Forest	2.5886	1.0380	0.9642
Linear Regression	7.9931	2.9053	0.6591
K-Nearest Neighbour	3.9097	1.6600	0.9184

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

The superior performance of the Random Forest model could be attributed to its ensemble nature, which combines the predictions of many individual decision trees. This architecture allows it to effectively capture complex, non-linear relationships between the input logs (GR, RHOB, NPHI, CALI) and the sonic log (Bi et al., [4]). While single Decision Trees can model non-linearity, they are prone to overfitting. Random Forest mitigates the problem of overfitting by averaging predictions from decorrelated trees in a process known as aggregating or "bagging", leading to better generalization (Ashby et al., [3]). KNN being non-parametric can also capture non-linearity but might be less robust with the given feature set or dimensionality. GBR also performed well, as expected from a powerful ensemble method, but Random Forest showed a slight edge in this specific evaluation.

To further understand the Random Forest model's behavior, feature importance scores were extracted. These scores indicate the relative contribution of each input feature to the prediction of the sonic log. Table 7 presents the feature importances obtained from the trained RF model. The ranking of the features or predictors shows that depth is the most important predictor. This is as expected as the speed at which acoustic waves travel through the surface is fundamentally influenced by depth as it is generally known that compaction and cementation change often in direct proportion with depth which decreases porosity and hence lowers transit time of acoustic waves between different formations in the subsurface. This is also reflected in the general trends of the sonic log signatures in the three wells. However, it is important to note that the predictors selected in this study should not necessarily be generalized to be the most important features for predicting sonic log data using machine learning algorithms since access to more log measurements or features could provide more predictors.

Table 7: Feature importances from the trained model

Feature (Input Log)	DEPTH	CAL	GR	LLD9	RHOB
Importance Score	0.9376	0.0248	0.0217	0.0080	0.0079

Compared to other studies, the accuracy achieved is competitive. For instance, some studies report correlation coefficients (related to R^2) for sonic log prediction using ensemble models in the range of 0.89 to 0.896 and RMSEs between 5.85 and 6.03 μs/ft (Saleh, [13]). Other works using drilling parameters and GR with XGBoost reported average absolute percentage errors of less than 10% (Alfaraj et al. [2]). The performance of the RF model in this study falls within these favorable ranges.

Model validation

The models were further subjected to validation by using each of the models to predict sonic from a known well (Well-1). Cross plots between the actual sonic from this well and the predicted sonic by the models are as shown in figs 9a-9e.

The core result of the comparative model validation evaluation is presented in Table 8. With an RMSE of 6.5588, an MAE of 3.6356, and an R^2 of 0.7695, the Random Forest regressor continues its superior performance over the others.

Table 8: Model validation error metrics using the validation well. The Random Forest Regressor model maintained its high-performance during validation, with only minor deviations in the blind well

ML MODEL	**RMSE**	**MEDAE**	**R^2 SCORE**
Random Forest	6.5588	3.6356	0.7695 (76.95%)
Linear Regression	7.0074	3.6384	0.6972 (69.72%)
Gradient Boosting	7.2509	4.0958	0.6757 (67.57%)
KNN	7.8721	4.4687	0.6679 (66.79%)
Decision Tree	7.9359	3.8115	0.6116 (61.16%)

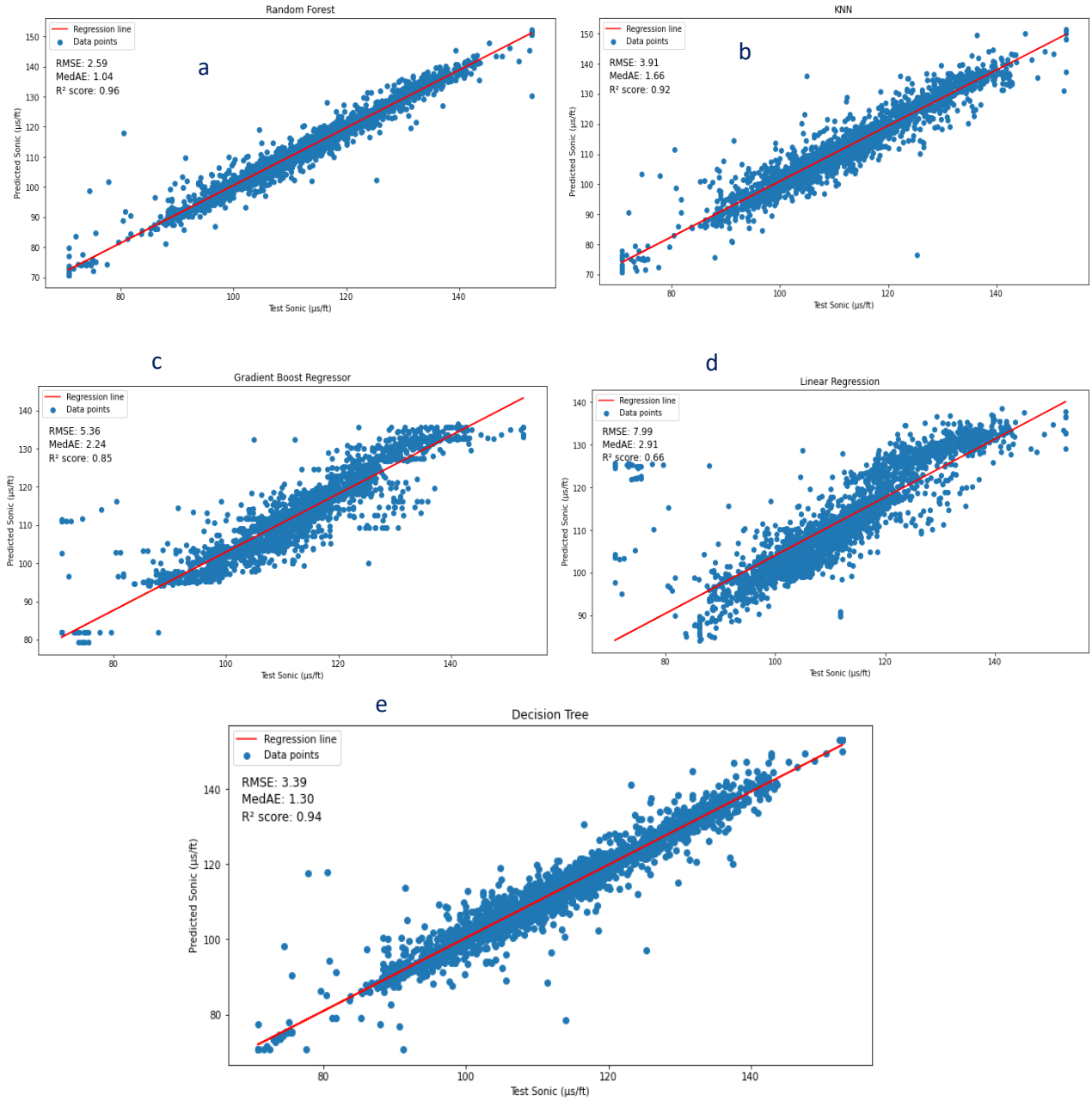

Figure 8: Scatter plots of test vs predicted sonic logs (a) Random Forest (b) KNN (c) Gradient boost (d)Linear regression (e) Decision tree

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

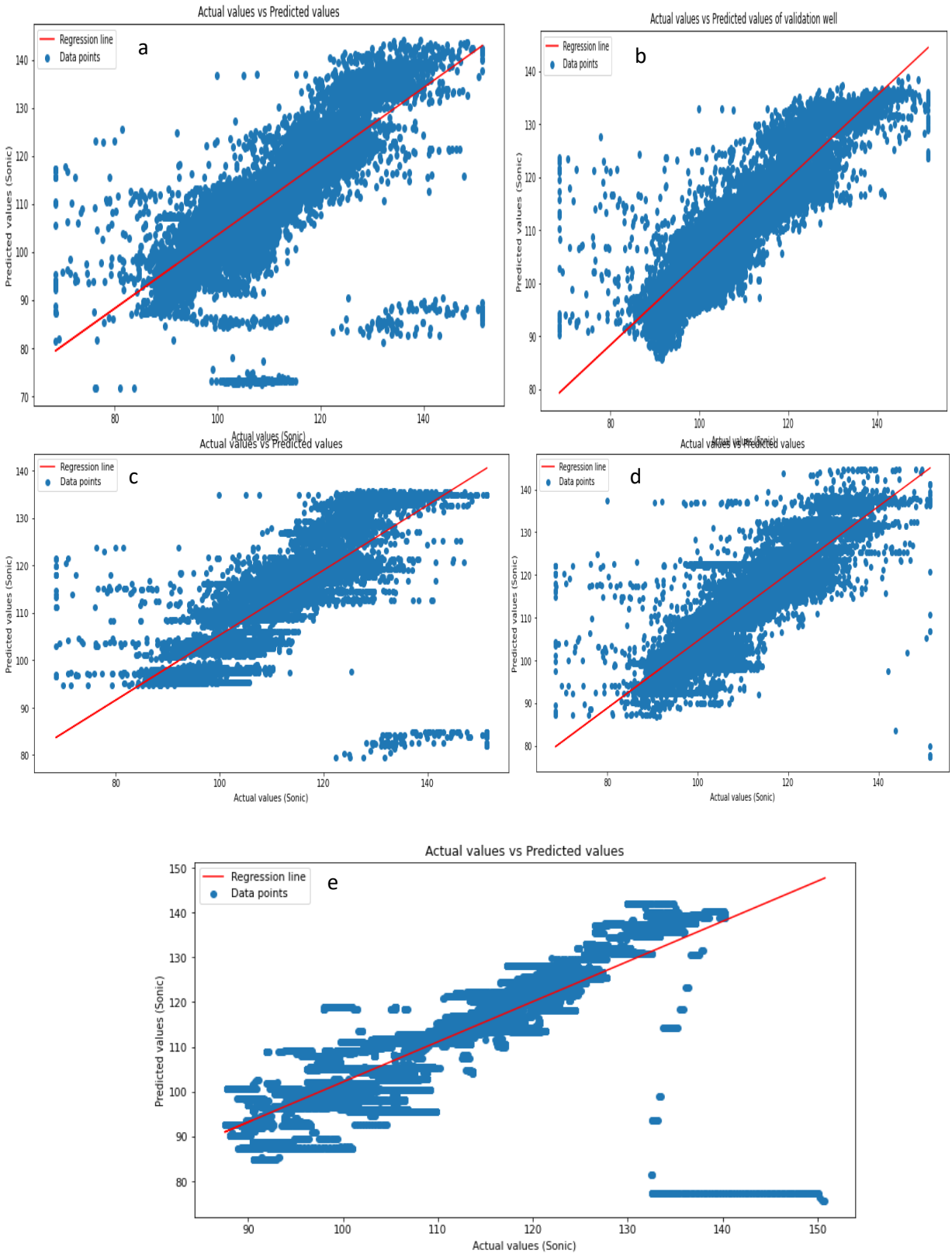

Figure 9: Scatter plots of test vs predicted sonic logs after validation (a) Random Forest (b) Linear Regression (c) Gradient Boost Regressor (d) K-Nearest Neighbour (e) Decision Tree

Model prediction

Following its superior performance on the test and validation sets, the trained Random Forest model was applied to Well-3, the designated prediction well which originally lacked a sonic log. The model

used the available GR, RHOB, NPHI, and CALI logs from Well-3 as input to generate a continuous predicted sonic log (PREDICTED_SONIC) over the logged interval. Figure 10 shows the input log alongside the newly predicted sonic log.

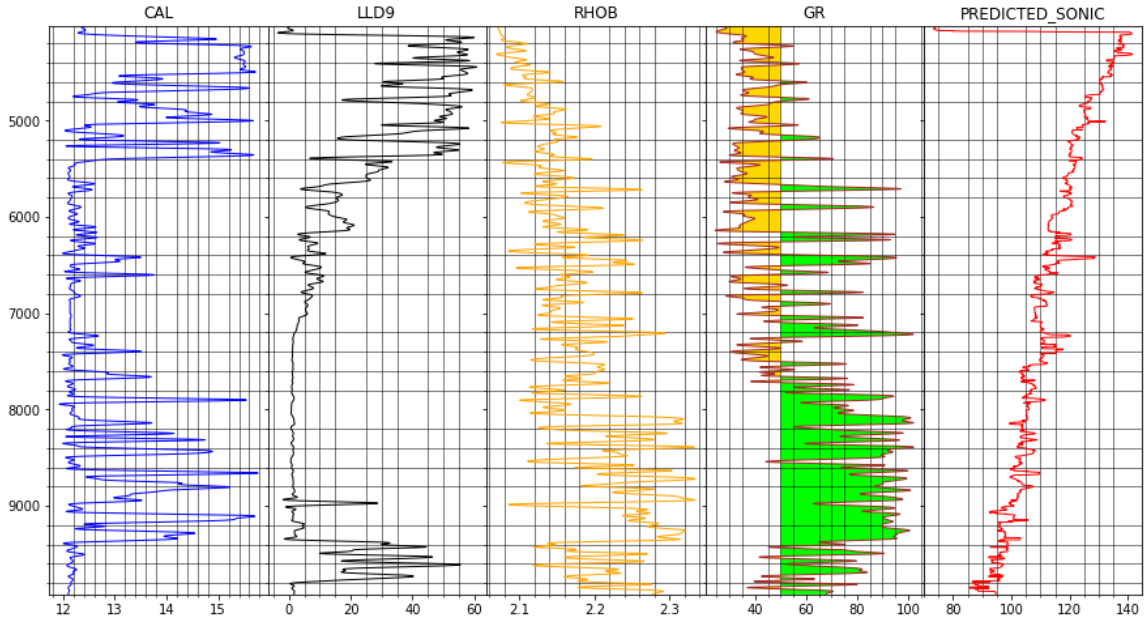

Figure 10: The predicted sonic of well 6 done using the Random Forest Regressor

Qualitative Well Correlation

To assess the geological significance and practical utility of the predicted sonic log beyond statistical metrics, a qualitative well correlation was performed. This involved visually comparing the predicted sonic log from Well-6 with the actual sonic logs from the nearby training and validation wells (Well-7 and Well-1). The wells are situated approximately 1 km apart, a distance over which significant geological markers and formation trends are often correlatable in deltaic settings like the Niger Delta (Ajaegwu [1]).

Based on visual inspection, the logs exhibited similar trends and pattern with a general decrease in sonic values from sand to shale corresponding to expected lithological changes (higher transit time indicates more porous or less consolidated formations (sands), while lower transit time indicates denser formations (shale). Also as identified on the GR section, the logs showed similarity in facies variations at almost equal depth intervals across the three wells from sand at the top (≈ 3000-5500ft) to shale-sand intercalations at the middle (≈ 5500-7750ft) to predominantly shale facie at the bottom (≈ 7750-9920ft). This visual consistency suggests that the Random Forest model successfully learned geologically meaningful relationships from the input data, capturing underlying formation properties that exhibit lateral continuity. This step is crucial because a high R^2 score alone does not guarantee a geologically sensible prediction; the qualitative correlation acts as an essential geoscience-based validation (Hesthammer and Fossen, [8]).

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

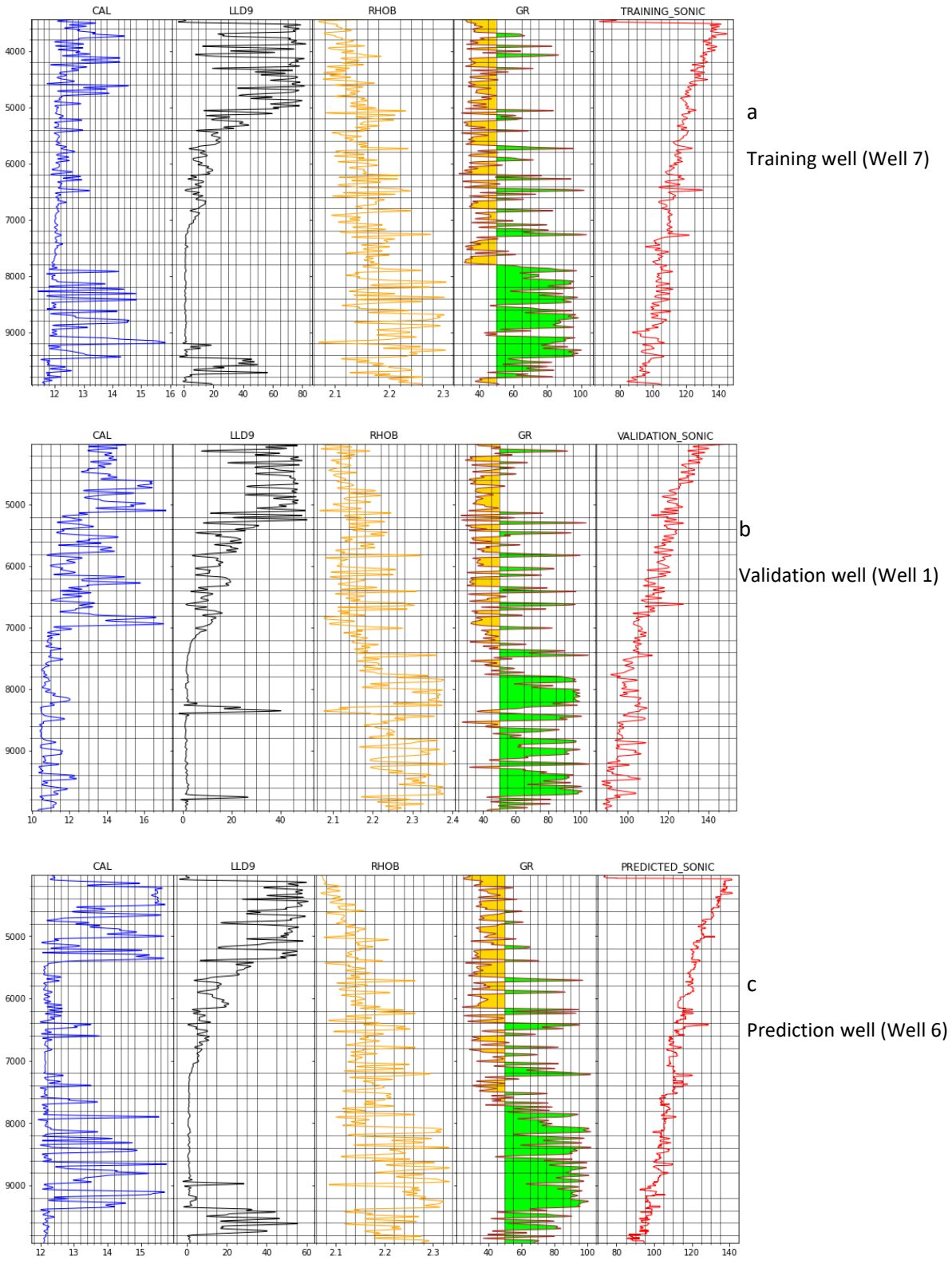

Figure 11: A supplementary QC of the predictive model using qualitative well correlation analysis among the three wells showing satisfactory results

CONCLUSIONS

This research successfully developed and evaluated a machine learning framework for predicting compressional sonic logs (SONIC) in the Zircon field, Niger Delta, Nigeria.

A comprehensive data preprocessing workflow, which included missing value imputation and outlier treatment using Z-score, IQR, and quantile-based approaches, was established to prepare high-quality well log data (Gamma Ray, Density, Resistivity, Caliper) from three wells.

Among five machine learning algorithms tested (Random Forest, Linear Regression, Decision Tree, Gradient Boost Regressor, and K-Nearest Neighbors), the Random Forest (RF) model demonstrated superior performance on the test data and an unseen validation well. It achieved an RMSE of 2.5886µs/ft, an MAE of 1.0380µs/ft, and an R-Squared value of 0.9642, and 6.5588µs/ft, 3.6356µs/ft and 0.7695 for RMSE, MEDAE and R-Squared values in the validation phase indicating a strong predictive capability.

The optimized RF model, when applied to a well lacking an original sonic log, generated a predicted sonic curve that showed remarkable similarity in log trends and patterns when qualitatively correlated with actual sonic logs from two contiguous wells approximately 1 km away. This confirmed the geological plausibility and practical utility of the predicted log.

Recommendations

The most critical recommendation is to expand the training dataset by including data from a significantly larger and more diverse set of wells from the Zircon field and, if possible, analogous fields within the Niger Delta and using multiple wells for both training and validation. Also there is a need to investigate the potential of more advanced ML algorithms, particularly deep learning models such as Long Short-Term Memory (LSTM) [14] networks or Convolutional Neural Networks (CNNs) (Saleh et al., [13]). These models are designed to capture sequential dependencies or spatial patterns in data, which could be advantageous for well log data if a sufficiently large dataset becomes available. Also, a cross-validation (k-fold) should be implemented to maximize the use of the limited data, this will help provide a more robust assessment of model stability before validation. Finally, future studies should explore the integration of complementary data sources, such as seismic attributes (which provide spatial context between wells), core data (for direct calibration of log responses to rock properties), or real-time drilling parameters (Alfaraj et al. [2]), which have shown promise in predicting logs.

References

1. Ajaegwu, N. E., Odoh, B. I., Akpunonu, E. O., Obiadi, I. I., & Anakwuba, E. K. (2012). Late Miocene to early Pliocene palynostratigraphy and palaeoenvironments of ANE-1 well, eastern Niger Delta, Nigeria. *Journal of Mining and Geology, 48*(1), 31–43.
2. Alfaraj, R. T., AlTammar, M. J., & Hamid, O. (2023). Application of machine learning for real-time prediction of sonic well logs using surface drilling parameters and gamma ray. *The Aramco Journal of Technology*, *Spring 2023*.
3. Ashby, M., Berestovsky, N., & Tobar, I. (2019). *Petrophysics-driven well log quality control using machine learning*. Anadarko Petroleum Corporation, Advanced Analytics and Emerging Technologies.
4. Bi, Jiaming & Li, Enhui & Luo, Yongxing. (2023). Petroleum Price Prediction Based on the Linear Regression and Random Forest. Applied and Computational Engineering. 8. 310-314. 10.54254/2755-2721/8/20230170.
5. Codd, E. F., 1970, A Relational Model of Data for Large Shared Data Banks: *Communications of the ACM*, **13** (6), 377-387.
6. Eke, P.O., Okeke, F.N. and Ezema, P.O. ‘’Improving the Geological Under-standing of the Niger Delta Basin of Nigeria Using Airborne Gravity Data,’’*International Journal of Geography and Geology*, vol. 5no. 5,pp. 97-103, 2016.
7. Doust, H., & Omatsola, E. (2002). *Niger Delta*. In D. E. Ajakaiye & A. W. Bally (Eds.), *Course manual and atlas of structural styles, Niger Delta*.
8. Hesthammer, Jonny & Fossen, Haakon. (2000). Uncertainties associated with fault sealing analysis. Petroleum Geoscience. 6. 10.1144/petgeo.6.1.37.
9. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. doi: 10.1109/MCSE.2007.55

Doi: https://doi.org/10.21608/jaiep.2025.430788.1027
Received: September 4, 2025; Revised: September 30, 2025; Accepted: October 17, 2025

10. Kouadio KL, Liu J, Liu R, Wang Y, Liu W (2024) K-Means Featurizer: A booster for intricate datasets. Earth Sci Inform 17:1203–1228. https:// doi. org/ 10. 1007/ s12145- 024-01236-3

11. Liu M, Nivlet P, Smith R, Ben Hasan N, Grana D. (2022) Recurrent neural network for seismic reservoir characterization. Adv Subsurf Data Anal 95–116. https:// doi. org/ 10. 1016/ b978-0-12- 822295- 9. 00010-8
12. McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.
13. Saleh K, Mabrouk WM, Metwally A. Machine learning model optimization for compressional sonic log prediction using well logs in Shahd SE field, Western Desert, Egypt. Sci Rep. 2025 Apr 29;15(1):14957.doi: 10.1038/s41598-025-97938-9.
14. Short, K.C. and Stauble, A. J. ''Outline of geology of Niger Delta,''*American Association of Petroleum Geologist Bulletin*, vol. 51,pp. .761-799, 1967.
15. Tuttle, M.L.W., Charpentier, R.R. and Brownfield, M.E. 1999. The Niger Delta Petroleum
16. Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.
17. Weber, K. J, Mandi J, Pilaar, W.F, Lehner, E., Precious, R.G (1978). The role of faults in hydrocarbon migration and trapping in Nigeria growth fault structures. 10th Annual Offshore Technology Conference Proceedings, 4: 2643-2653.
18. Weber, K. J. (1987). Hydrocarbon distribution patterns in Nigerian growth fault structures controlled by structural style and stratigraphy. Journal of Petroleum Science and Engineering, 1(2), 91-104. https://doi.org/10.1016/0920-4105(87)90001-5