Journal of Oil, Gas and Petrochemical Technology 12(2): 45-65, Summer and Autumn 2025

RESEARCH PAPER

Artificial Intelligence in Reservoir Characterization: Predicting Shale Volume with ANN, RF, and ET

Mohammad Yasin Hosseini [1], Ali Ranjbar [2*], Mohammad Mahdi Hosseini [3]

1. Faculty of Petroleum, Gas, and Petrochemical Engineering, Persian Gulf University, Bushehr, Iran

2.Corresponding Author. Department of Petroleum Engineering, Faculty of Petroleum, Gas, and Petrochemical Engineering, Persian Gulf University, Bushehr, Iran. E-mail address: Ali.ranjbar@pgu.ac.ir

ARTICLE INFO

Article History:
Received 31 May 2025
Revised 29 August 2025
Accepted 18 September 2025

Keywords:
Shale Volume Prediction
Well Log Data
Machine Learning
Artificial Neural Network
Random Forest
Extra Trees
decision-making.

ABSTRACT

Shale volumes are essential for lithology identification, reservoir evaluation, and stratigraphic correlation in the subsurface formation analysis. This study evaluates the performance of three machine learning algorithms—Random Forest (RF), Extra Trees (ET), and Artificial Neural Network (ANN)—for shale volume prediction using the conventional well log data. The models were trained and tested on a comprehensive dataset from one of the oil fields in southern Iran, incorporating parameters such as sonic travel time (DTC), bulk density (RHOZ), resistivity (RT), neutron porosity (HTNP), and caliper (HCAL). Results demonstrated that ANN achieved superior accuracy with an R^2 of 0.9779 and Root Mean Squared Error (RMSE) of 0.0130 API, outperforming both RF (R^2 = 0.9640, RMSE = 0.0166 API) and ET (R^2 = 0.9007, RMSE = 0.0275 API). While ANN excelled in capturing complex nonlinear relationships, tree-based methods offered faster training times and greater interpretability through feature importance metrics. The findings highlight ANN as the preferred choice for high-fidelity shale volume prediction, whereas RF provides a balanced solution for scenarios requiring both speed and competitive accuracy. This study underscorses the transformative potential of machine learning in petrophysical analysis, offering practical recommendations for model selection based on the project-specific needs..

How to cite this article
Hosseini M Y, Ranjbar A, Hosseini M M, Artificial Intelligence in Reservoir Characterization: Predicting Shale Volume with ANN, RF, and ET, Journal of Oil, Gas and Petrochemical Technology, 2025; 12(2): 45-65. 10.22034/jogpt.2025.527316.1141

1. INTRODUCTION

Predicting shale volume is paramount in comprehensive reservoir evaluation, particularly within the challenging domain of unconventional reservoirs. This critical parameter directly dictates the reliable estimation of hydrocarbon reserves and profoundly influences the design and efficacy of hydraulic fracturing operations. Machine learning (ML) methods offer a powerful approach to this task by leveraging extensive well log data and integrating various geological parameters, often outperforming traditional methods in complex formations where conventional techniques

*Corresponding Author Email: Ali.ranjbar@pgu.ac.ir **

struggle to capture the non-linear relationships between the log responses and shale content [1].

Shale volume, fundamentally representing the proportion of shale within a reservoir rock, is a cornerstone parameter in robust reservoir characterization workflows. Its accurate estimation is essential for several interconnected reasons. Shale content significantly impacts a reservoir petrophysical properties, notably porosity and permeability. For instance, the presence of clay minerals within shale can reduce effective porosity by occupying pore spaces and decreasing permeability through restricting the fluid flow pathways. Studies have demonstrated how an increasing proportion of clay minerals, such as chlorite and montmorillonite, can significantly reduce the reservoir quality by clogging pore throats and affecting fluid flow pathways, while also showing that porosity and permeability of shale samples can significantly decrease under pressure [2, 3]. Advanced reservoir models, crucial for simulating fluid flow and production, require precise shale volume inputs to accurately represent the heterogeneous nature of the subsurface and reliably estimate hydrocarbon content [4]. In low-permeability unconventional reservoirs, such as shale gas and tight oil plays, economic hydrocarbon production is heavily reliant on the successful creation of an extensive and conductive fracture network through hydraulic fracturing. Understanding the distribution and proportion of shale is critical because the mineral composition of shale, including its clay content and bedding structures, significantly influences hydraulic fracture propagation paths and the complexity of the generated fracture network [5, 6]. Accurate shale volume prediction allows for the optimized proppant placement, fluid selection, and overall fracture design, directly impacting stimulated reservoir volume and production rates [7]. Furthermore, distinguishing between shale and other lithologies (e.g., sandstone, limestone, dolomite) is fundamental for accurately defining reservoir boundaries and identifying potential pay zones. Shale intervals often indicate non-reservoir rock, while a reduction in shale content typically correlates with higher reservoir quality, making shale volume a key parameter in delineating productive zones from non-productive ones [8].

Traditional methods for shale volume estimation, predominantly relying on single well logs such as the gamma ray log (e.g., Larionov, Stieber, or Clavier methods), often fall short due to inherent limitations. These methods are frequently based on empirical relationships derived from specific geological settings and may not be universally applied due to the local geological variations [1]. The gamma ray log's response, while generally correlated with the clay content, can be significantly affected by factors other than the clay minerals, such as the presence of radioactive minerals like uranium, thorium, or potassium-rich minerals (e.g., potassium feldspars, micas, glauconite), or even organic matter, leading to inaccurate shale volume estimations [8, 9]. Additionally, the actual relationship between well log responses (e.g., gamma ray, resistivity, density, neutron porosity) and shale content is inherently complex and non-linear, influenced by multiple interacting geological factors [10, 11]. Traditional methods struggle to capture these intricate, multivariate dependencies because they often rely on the simplified linear or empirical models. Machine learning techniques directly address these limitations by learning complex, multivariate patterns from large and diverse datasets. Unlike empirical equations, ML algorithms can identify subtle correlations and non-linear relationships across multiple well logs and geological parameters [12]. For instance, recent studies have successfully employed machine learning models like Random Forest and Support Vector Machines to provide more accurate shale volume estimations, demonstrating their ability to handle the complexity and non-linearity of well log data compared to conventional methods [13]. Artificial Neural Networks (ANNs) have also been shown to effectively estimate shale volume from well log data, highlighting their capability to learn complex patterns and overcome limitations of the traditional approaches by utilizing multiple logs as input [14].

Machine learning approaches have become increasingly valuable for predicting shale volume in reservoir characterization, offering significant advantages over traditional methods. As a critical parameter influencing key petrophysical properties like porosity, permeability, and saturation, accurate shale volume estimation is essential for the reliable assessment of hydrocarbon reserves and reservoir producibility [1, 15]. Various machine learning algorithms have demonstrated effectiveness in this application, each with unique strengths.

Artificial Neural Networks (ANNs) excel at modeling the complex relationships between well log data and shale volume through their robust predictive capability [9, 16]. Support Vector Regression (SVR) proves particularly effective for handling high-dimensional data and capturing nonlinear relationships in shale volume estimation, with recent studies demonstrating its utility in predicting tight oil recovery [15, 17]. Among ensemble methods, Random Forest (RF) stands out by combining multiple decision trees to enhance prediction accuracy while mitigating overfitting risks, showing strong generalization for noisy datasets in the reservoir characterization [15]. For handling complex datasets with superior efficiency, Extreme Gradient Boosting (XGBoost) has shown remarkable performance, especially in predicting TOC content from wireline log data [18]. When dealing with sequential data patterns, Long Short-Term Memory (LSTM) networks, as a specialized recurrent neural network architecture, have successfully been applied to predict shale gas horizontal well productivity, demonstrating their value in temporal analysis of reservoir characteristics and outperforming traditional methods [19, 20]. These machine learning techniques collectively represent a paradigm shift in shale volume prediction, offering more sophisticated and reliable alternatives to the conventional approaches.

The use of machine learning (ML) for shale volume prediction offers several advantages. Firstly, ML models can significantly reduce costs by decreasing the reliance on expensive logging tools and wireline services [21]. Secondly, these techniques enhance accuracy, yielding precise shale volume predictions that support reliable reservoir evaluation and lithology identification [22, 23]. Finally, AI-driven models improve efficiency by competently processing large datasets and discerning complex patterns, thereby streamlining reservoir characterization workflows [24].

The effectiveness of these machine learning methods is further highlighted in various studies, which are summarized in the table below, detailing their application to well log prediction.

Table 1- Summary of machine learning methods used in well log prediction across studies.

Input Parameters	Output Parameters	Method	Conclusions	Authors
Gamma Ray, Resistivity, Density, Neutron Porosity, X-ray Diffraction data	Shale Volume	Support Vector Regression (SVR), Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB), Deep Neural Networks (DNN)	ML models, especially Gradient Boosting, showed significantly higher accuracy (correlation coefficient of 0.98 for GB) compared to traditional methods (0.27-0.52) in complex heterogeneous reservoirs like the Bakken Formation, which are influenced by radioactive minerals.	N. Bettir, A. Dehdouh, I. Mellal, A. Kareb and M. Rabiei[1]
Gamma Ray, Density, Neutron, Sonic, Core data for calibration	Shale Volume	XGBoost	XGBoost demonstrated outstanding performance with a low validation set Root Mean Square Error (RMSE) of 0.078, indicating its reliability for fast and consistent shale volume estimation.	M. Ali[25]

Gamma Ray, Spontaneous Potential, Resistivity, Density, Neutron	Shale Content / Shale Volume	CNN-BiGRU-VAE Neural Network (Convolutional Neural Network - Bidirectional Gated Recurrent Unit - Variational Autoencoder)	The hybrid neural network model effectively described the strong non-linear mapping relationship between well logging parameters and shale content, offering an accurate approach for prediction.	H. Zhang and W. Wu[11]
General well logs likely)	Shale Volume	Machine Learning Methods (Specifics not detailed in snippet)	ML methods are applied for shale volume estimation in various fields, suggesting their utility in complex geological settings.	P. Ebrahimi, A. Ranjbar, Y. Kazemzadeh and A. Akbari[12]
Depth, Vp, GR, Density, Magnetic susceptibility	Shale Volume	Artificial Neural Network	An ANN model was successfully developed to predict shale volume from well log data, demonstrating a robust correlation. The procedure was found to be easy, straightforward, and likely to give reasonable results.	Vu, D.H. and Nguyen, H.T. [9]
Well information, completion and hydraulic fracturing parameters, production data	Cumulative Gas Production	Deep Neural Network (DNN) (Multi-layer perceptron)	DNN model effectively predicts cumulative gas production, with improved prediction performance when using principal component analysis to extract important information and identifying key variables through variable importance. This indirectly relates to shale volume, as it impacts reservoir quality and productivity.	D. Han and S. Kwon[26]
Caliper, Depth, Gamma Ray, Resistivity, Sonic, Density, Neutron logs and core data	Shale Volume, Porosity, Permeability, Water Saturation	ANFIS, ELM, GP, AdaBoost, Decision Tree (DT)	Machine learning models, particularly AdaBoost (RMSE of 0.0152 and AARE% of 3.1610 for water saturation), demonstrated high accuracy in enhancing petrophysical evaluation and identifying productive zones. While this study focuses on multiple petrophysical parameters, shale volume is an implicit and critical component for these estimations.	B. Rezaei Mirghaed, A. Dehghan Monfared and A. Ranjbar[27]

In this study, machine learning techniques were employed to predict shale volume, aiming to significantly boost the accuracy and efficiency of reservoir characterization and lithology discrimination. This approach directly facilitates a more cost-effective and reliable evaluation of subsurface formations, which is crucial for optimizing hydrocarbon exploration and production.

The novelty of this work lies in the comparative analysis of advanced ensemble learning methods like Extra Trees and Random Forest, alongside a powerful Artificial Neural Network (ANN), for shale volume prediction. While machine learning has been applied to this problem before, our research provides a robust head-to-head comparison of these specific techniques, highlighting their individual strengths and limitations when applied to a comprehensive set of well log data. It is hypothesized that these advanced algorithms, particularly the ensemble methods, will capture complex, non-linear relationships within the data that traditional methods often miss, leading to superior prediction accuracy and better uncertainty quantification.

To achieve this, we used a rich dataset consisting of Depth, Compressional Slowness (DTC), Clay

Volume (VCLC), Formation Temperature (TEMP), Compensated Neutron Porosity (HTNP), True Resistivity (RT), Bulk Density (RHOZ), and Caliper (HCAL) logs. These parameters were meticulously selected to provide a comprehensive representation of the petrophysical and geological characteristics influencing shale content. We rigorously pre-processed this data and handled any missing values to optimize the input for our models. The performance of each machine learning model will be thoroughly evaluated using industry-standard metrics such as R-squared and Mean Squared Error, allowing for a quantitative assessment of their predictive capabilities. Ultimately, our paper will not only present the optimal machine learning approach for shale volume estimation but also provide insights into the importance of each input parameter, contributing valuable knowledge for future reservoir characterization efforts.

While machine learning has been applied to shale volume prediction in prior studies, the novelty of this work lies in presenting a direct comparative evaluation of the three widely used yet distinct approaches—ANN, Random Forest, and Extra Trees—on a real-world field dataset. By highlighting the trade-offs between accuracy, computational cost, and interpretability, this study provides practical guidelines for selecting the most suitable model in different reservoir characterization scenarios, which have not been systematically addressed in previous research.

2. Data Description

The dataset used in this study consists of petrophysical well log measurements acquired from a well in an oil field in southern Iran. The primary target variable is the shale volume, which serves as a key indicator of lithology and shale content, with values typically ranging between 0 and 1.

The input logs and measurements include key geological and petrophysical parameters collected over a range of depths from [4305] to [4554] meters. These comprise Sonic Travel Time (DTC), representing compressional wave slowness values between [48.5] and [78.6] µs/ft; Bulk Density (RHOZ), indicating formation density within [2.1] to [2.7] g/cm^3; Resistivity (RT), reflecting deep resistivity readings from [0.13] to [15002] Ω·m; Neutron Porosity (HTNP), measuring hydrogen content as a dimensionless fraction; and Caliper (HCAL), which records borehole diameter in inches. Together, these logs provide a comprehensive set of measurements essential for analyzing subsurface formations.Table 2 below displays the dataset

Table 2- Statistical profiling of well-log parameters

Parameter	Mean	Standard Error	Median	Mode	Standard Deviation	Sample Variance	Kurtosis	Skewness	Range
Depth	4429.66	1.7795	4429.66	4305.15	71.9523	5177.14	-1.2	0	249.022
DTC	62.44	0.206	64.1992	57.5263	8.3282	69.3593	-1.3126	-0.1615	30.1094
VCLC	0.8935	0.0015	0.8886	0.8175	0.062	0.0038	-0.2441	-0.1225	0.4237
TEMP	135.102	0.0596	135.102	130.935	2.4084	5.8005	-1.2	0	8.3354
HTNP	0.1023	0.0015	0.1074	0.0115	0.0621	0.0039	-1.1655	-0.0282	0.2308
RT	259.199	20.4257	57.7567	0.1291	825.915	682135	106.327	8.8034	15002.3
RHOZ	2.4623	0.0027	2.449	2.3791	0.1101	0.0121	-0.5255	0.29	0.6613
HCAL	6.1924	0.0198	5.863	5.7711	0.8019	0.643	13.0423	3.0023	6.4587
Vshale	0.3454	0.0021	0.335	0.266	0.085	0.0073	12.3135	1.404	1.00

characteristics.

This comprehensive collection of well logs provides a robust foundation for training and evaluating the machine learning models in this study.

3. Methodology

This study evaluated the performance of the three machine learning algorithms—Random Forest (RF), Extra Trees (ET), and Artificial Neural Network (ANN)—for the shale volume prediction. These methods were selected based on their well-documented success in geophysical applications [28-30].

3.1. Random Forest (RF)

Random Forest (RF) is a powerful and widely used ensemble learning method for both classification and regression tasks. It operates by constructing a multitude of decision trees at training1 time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. This approach effectively combines the "wisdom of crowds" to achieve higher accuracy and reduce overfitting compared to the single decision trees.

The core idea behind Random Forest (RF), introduced by Leo Breiman in 2001 [28], is to strategically introduce randomness at two key stages to build a robust ensemble model. Firstly, it employs Bootstrap Aggregation (Bagging): instead of training each decision tree on the entire dataset, RF creates multiple subsets of the original training data by randomly sampling with replacement [28, 31, 32]. Each of these bootstrapped samples is then used to train an independent decision tree. This "bagging" strategy plays a crucial role in reducing the variance of the model and preventing overfitting [33].

Secondly, Random Feature Subspace is introduced during the tree construction process. When building each individual decision tree within the forest, at each split node, a random subset of features is considered for splitting, rather than all available features [28, 34]. For a regression problem, it is common to consider features (where p is the total number of features). This clever step further decorrelates the trees, as they are less likely to rely on the same dominant features, leading to a more diverse and robust ensemble [32, 35].

The prediction of the ensemble is subsequently obtained by averaging the individual tree out-

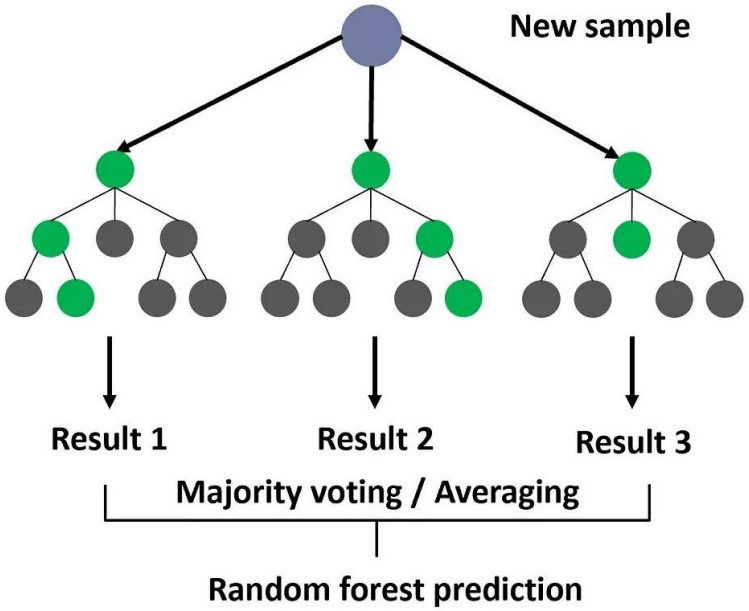

Figure 1- Schematic representation of the Random Forest algorithm.

puts. For regression, if is the prediction of the b-th tree for an input x, the final Random Forest prediction H(x) is given by:

$$H(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \quad (1)$$

where B=200 is the total number of trees [28]. This averaging process effectively smooths out the individual tree predictions, significantly reducing their variance and consequently improving the overall generalization ability of the model. Figure 1 visually depicts the complete Random Forest algorithm.

When constructing each decision tree, the algorithm needs a criterion to decide how to split the nodes effectively. For classification tasks, Gini impurity is a common measure used to evaluate the quality of a split [36]. It quantifies the disorder or impurity of a set of samples at a given node. Specifically, for a node m with a set of samples Dm and K classes, the Gini impurity is calculated as:

$$G(D_m) = 1 - \sum_{k=1}^{K} p_{mk}^2 \quad (2)$$

where p_{mk} represents the proportion of samples belonging to class k in node m. A lower Gini impurity value indicates a more homogeneous node, meaning that most samples within that node belong to the same class. The objective of the algorithm is to find splits that maximize the reduction in Gini impurity, thereby creating purer child nodes [37]. While Gini impurity is primarily employed for the classification scenarios, for regression tasks, measures like Mean Squared Error (MSE) or Mean Absolute Error (MAE) are typically utilized to determine optimal splits, with the aim of reducing the variance of the target variable within each resulting child node.

3.2. Extra Trees

Extra Trees, formally known as Extremely Randomized Trees, stand as a prominent ensemble machine learning method that extends the core principles of Random Forest by introducing an even higher degree of randomization during the tree construction phase. This additional layer of randomness is strategically designed to maximize the diversity among the individual decision trees within the ensemble, leading to robust predictive models that often boast improved generalization capabilities and notable computational efficiencies [29].

The fundamental departure of ET from RF lies in the mechanism of determining splits at each node during the recursive tree-building process. While a Random Forest carefully selects the optimal split point from a random subset of features (a process that still involves an optimization step), Extra Trees radically simplifies this by employing Random Threshold Selection without Optimization. For each feature randomly selected as a candidate at a given node, Extra Trees does not search for the best possible threshold that optimizes an impurity measure (like Gini impurity or MSE). Instead, it randomly selects a threshold within the observed range of that feature for that specific node data [29]. This means that if a feature ranges from 0 to 100, ET might simply pick 47.3 as a split point, rather than systematically evaluating all potential split points to find the one that yields the purest child nodes. This complete elimination of the optimization step at each split is the primary reason for ET's superior training speed, particularly on large datasets [38]. This Impact on Bias-Variance Trade-off is significant; by foregoing the local optimization, individual Extra Trees tend to have a slightly higher bias than individual trees in a Random Forest. They are "less perfect" fits to their specific bootstrap sample because their splits are not optimally chosen. However, this intentional "weakening" of the individual trees is precisely what leads to their profound decorrelation and increased diversity within the ensemble. When these highly varied (and therefore less correlated) trees predictions are aggregated, their individual errors and biases tend to cancel out more effectively. The net effect is a substantial reduction in the overall variance of the ensemble model. This often results in a more favorable bias-variance

trade-off for Extra Trees, enabling it to achieve comparable, or even superior, accuracy to Random Forest while being computationally more efficient during the training phase [38, 39].

The enhanced diversity derived from random thresholding makes Extra Trees exceptionally robust to noise and slight perturbations in the training data. Each tree sees not only a bootstrapped sample of data (if bagging is used, though pure ET sometimes uses the whole dataset without bagging) but also highly randomized splitting rules, ensuring that no single feature or split point dominates the tree construction process. This contributes to excellent generalization, making ET less prone to overfitting, even on complex or high-dimensional datasets [39].

Similar to Random Forest, the final prediction from an Extra Trees model is obtained by aggregating the outputs of all the individual trees. For regression tasks, this typically involves simply averaging the predictions from each tree:

$$H_{ET}(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \quad (3)$$

where H_{ET} (x) is the final prediction of the Extra Trees model for input x, h_b (x) is the prediction of the b-th individual tree, and B is the total number of trees in the ensemble. For classification problems, a majority voting scheme is usually applied. This averaging mechanism further leverages the diversity of the individual trees to produce a stable and accurate final output.

3.3. Artificial Neural Network

The core of this system is a feedforward Artificial Neural Network, a type of neural network where connections between the nodes do not form a cycle. This particular ANN is designed to model the complex, nonlinear relationships found within well log data. To combat the common problem of overfitting, where a model learns the training data too well and performs poorly on new, unseen data, dropout regularization is implemented [30, 40]. Dropout, originally proposed by Srivastava et al., [40], works by randomly "dropping out" (setting to zero) a certain percentage of neurons during training, which prevents complex co-adaptations on the training data and encourages the network to learn more robust features. A general schematic of an ANN is shown in Figure 2.

The network itself follows a sequential feedforward architecture, meaning information flows in one direction from input to output through a series of layers [41, 42]. It consists of three hidden layers, which are the computational core of the network between the input and the output layers. These hidden layers are structured as follows: the first hidden layer contains 128 neurons, the second has 64 neurons, and the third contains 32 neurons. This decreasing number of neurons in successive layers often helps the network learn increasingly abstract representations of the data.

Each of these hidden layers utilizes the Rectified Linear Unit (ReLU) activation function. ReLU, defined as f(x)=max(0,x), is a popular choice because it introduces non-linearity into the network, which is crucial for modeling complex patterns that linear models can not capture [43]. It does this by outputting the input directly if it is positive, and zero otherwise. This simplicity also helps with computational efficiency.

In contrast to the hidden layers, the output layer employs a linear activation function (f(x)=x). This choice is ideal because the network is tasked with a continuous regression task, specifically shale volume prediction [44]. A linear activation function in the output layer allows the network to output a continuous range of values, which is necessary for predicting a numerical quantity like shale volume. This specific configuration, balancing the number of layers and neurons with appropriate activation functions, was carefully selected to provide sufficient model capacity to capture underlying data patterns while maintaining computational efficiency and preventing overfitting.

information from input to output through interconnected layers of neurons.

The process of teaching this ANN to make accurate predictions involves two key components:

a loss function and an optimizer. The loss function quantifies how well the network is performing. For this ANN, the Mean Squared Error (MSE) was chosen [45]. MSE calculates the average of the squared differences between the actual (true) values and the values predicted by the network. The formula for MSE is given by:

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

where n is the number of data points,Y_i represents the actual value, and $\hat{y}_i$ represents the predicted value. The goal during training is to minimize this MSE, meaning the network is getting closer to making accurate predictions.

To minimize the loss function, an optimizer is employed. Here, the Adam optimizer was used [46]. Adam is a sophisticated and widely adopted optimization algorithm that combines the benefits of the two other popular optimizers: momentum and adaptive learning rates. Momentum helps accelerate training in the right direction by considering past gradients, while adaptive learning rates mean that the learning rate (which determines the size of the steps taken during the optimization) is adjusted individually for each parameter based on the historical gradients. This combination allows Adam to converge quickly and effectively, even with complex datasets, and helps in navigating the loss landscape efficiently to find the optimal set of weights for the network. The learning rate, denoted as η, is a crucial hyperparameter that controls how much the model weights are adjusted with respect to the loss gradient during each training iteration. The general update rule for parameters θ in an optimization context often involves a learning rate:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla J(\theta_t) \quad (5)$$

where J is the loss function. Adam refines this by incorporating adaptive learning rates and moment estimates.

3.4. Metrics and Validation Protocol

The evaluation of model performance was conducted using two primary metrics that quantify different aspects of the predictions. The first, the R^2 score, measures the proportion of variance in the observed data that is explained by the model. A higher R^2 value, approaching 1, indicates that the model accounts for most of the variability in the data, reflecting good predictive power. Con-

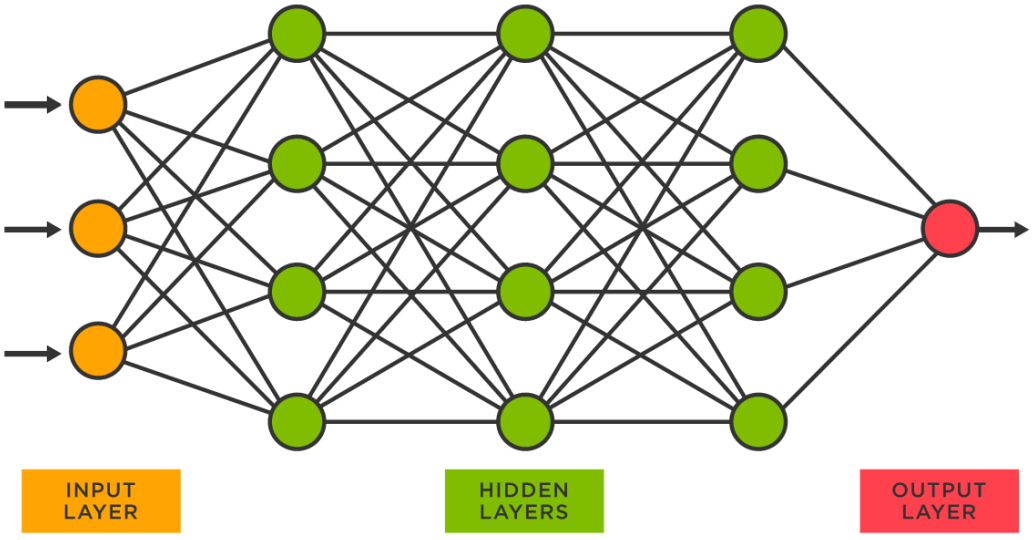

Figure 2-General schematic of a feedforward Artificial Neural Network, illustrating the flow of information from input to output through interconnected layers of neurons.

versely, lower values suggest less effective modeling. The R^2 is calculated as [47]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (6)$$

where y_i are the observed values, , $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of the observed data. This metric is widely used for regression tasks due to its interpretability.

The second metric, RMSE (Root Mean Squared Error), quantifies the average magnitude of the prediction errors in the same units as the original data (API units in this case). It squares the errors

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (7)$$

to penalize larger deviations more severely, then takes the square root to bring the measure back to the original scale, making it easier to interpret. A lower RMSE indicates better model accuracy [48].

To ensure the robustness of the models, the dataset was split into training and testing sets using an 80-20 stratified partitioning scheme. The stratification process maintains the distributional characteristics of the data across both sets, preventing biases and ensuring representative sampling. For the ANN specifically, early stopping was employed during the training, monitoring the validation loss after each epoch. This method halts training if the validation loss does not improve for 20 consecutive epochs, serving as a regularization technique to prevent overfitting and ensure the model generalizes better to unseen data.

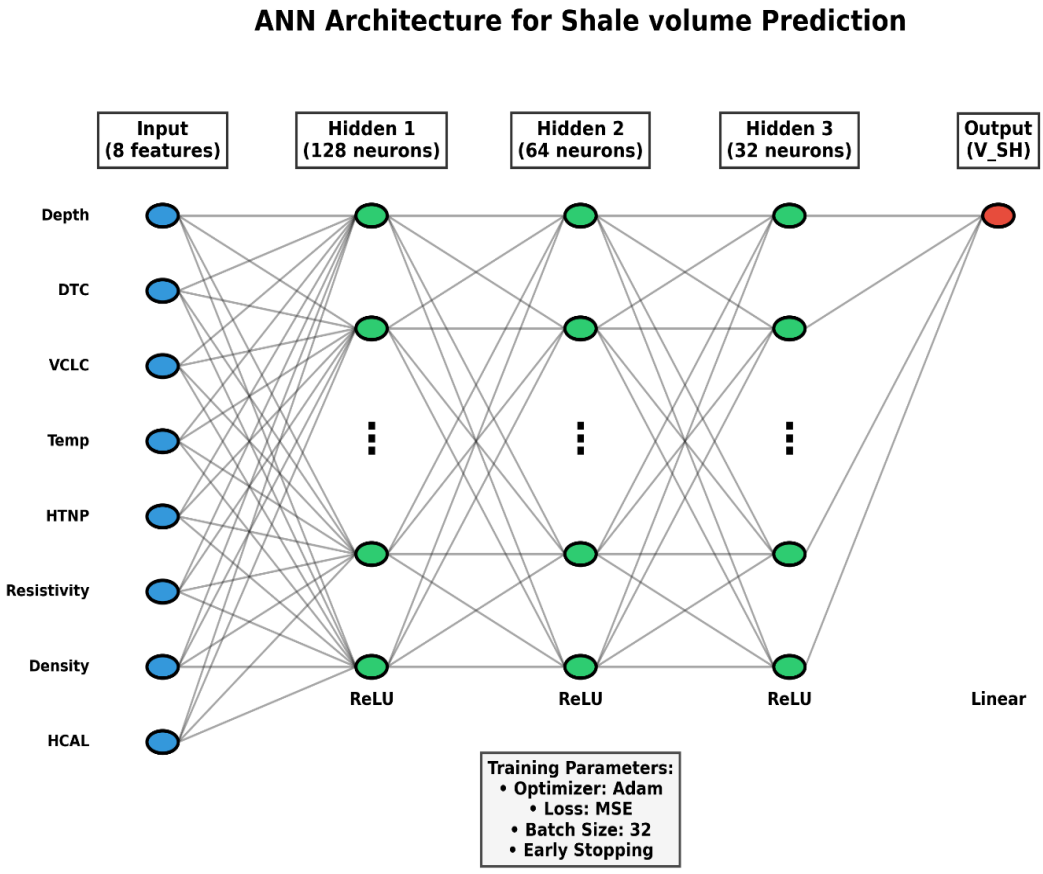

Figure 3-Visualization of the implemented feedforward neural network architecture, detailing the input

4. Results and Analysis

In this study, we employed three machine learning techniques—Random Forest (RF), Extra Trees, and Artificial Neural Networks (ANN)—to predict the shale volume. The input data used for these models included Depth, DTC, VCLC, TEMP, HTNP, RT, RHOZ, and HCAL measurements. Figure 3 illustrates the algorithms of the ANN method utilized in this study.

features, hidden layer dimensions, activation functions, and training parameters employed in the shale volume prediction model.

4.1. Preliminary statistical analysis

Based on the descriptive statistics, the characteristics of each parameter and their potential relationships with Vshale can be analyzed.

Depth has a mean of 4429.66, a standard deviation of 71.95, and a range of 249.02. Its distribution is fairly symmetric with zero skewness and a negative kurtosis (-1.2), indicating that most values are concentrated around the mean and the distribution is relatively uniform without long tails. DTC, with a mean of 62.44 and a standard deviation of 8.33, exhibits slightly negative skewness (-0.16) and negative kurtosis (-1.31), suggesting that most values cluster around the mean with few extreme deviations. VCLC, with a mean of 0.8935 and a standard deviation of 0.062, shows a tight and roughly symmetric distribution within a limited range.

TEMP has a mean of 135.10 and a standard deviation of 2.41, displaying a nearly symmetric and uniform distribution with zero skewness and negative kurtosis, indicating relatively small and predictable temperature variations. HTNP, with a mean of 0.1023 and a standard deviation of 0.0621, also exhibits limited dispersion and slight negative skewness, meaning that most values are near the mean with few outliers.

RT shows a mean of 259.2 but a much lower median of 57.76 and an extremely high standard deviation of 825.91, indicating the presence of significant outliers and a highly asymmetric distribution with positive skewness (8.8) and extreme kurtosis (106.33). This suggests that RT may not have a simple linear relationship with Vshale and might require normalization or outlier handling. RHOZ, with a mean of 2.4623 and a standard deviation of 0.1101, has a relatively symmetric and stable distribution, making it a reliable indicator for estimating Vshale.

HCAL has a mean of 6.1924 and a standard deviation of 0.802, with very high kurtosis (13.04) and positive skewness (3.0), reflecting extreme values and a non-normal distribution. Finally, Vshale itself, with a mean of 0.3454, standard deviation of 0.085, and range of 1.00, exhibits strong positive skewness (1.404) and very high kurtosis (12.31), meaning that most values are concentrated at lower shale volumes with a long tail toward higher values.

Overall, parameters such as Depth, DTC, VCLC, TEMP, and RHOZ are relatively stable with low variability and can serve as primary predictors for shale volume. In contrast, RT and HCAL are highly dispersed and skewed, requiring preprocessing or outlier management before modeling. The statistical profile of Vshale indicates a positively skewed target with a wide range, so predictive models should be capable of handling asymmetric distributions and extreme values. In summary, using stable, low-variance parameters as inputs while appropriately managing outliers in highly skewed features will likely yield accurate and reliable Vshale predictions. Generated heatmaps (Figure 4) visualize relationships between variables.

4.2. Data Preprocessing

Initially, records with missing values were removed to improve data quality and minimize the impact of incomplete data on the model[13, 49]. The dataset was then split into training and testing sets in an 80-20 ratio using stratified sampling based on Depth to ensure similar distributions in both subsets and provide a realistic model evaluation[50]. For the artificial neural network (ANN) model, features were normalized using Z-score standardization to align variable scales and facilitate faster and more stable training convergence[51].

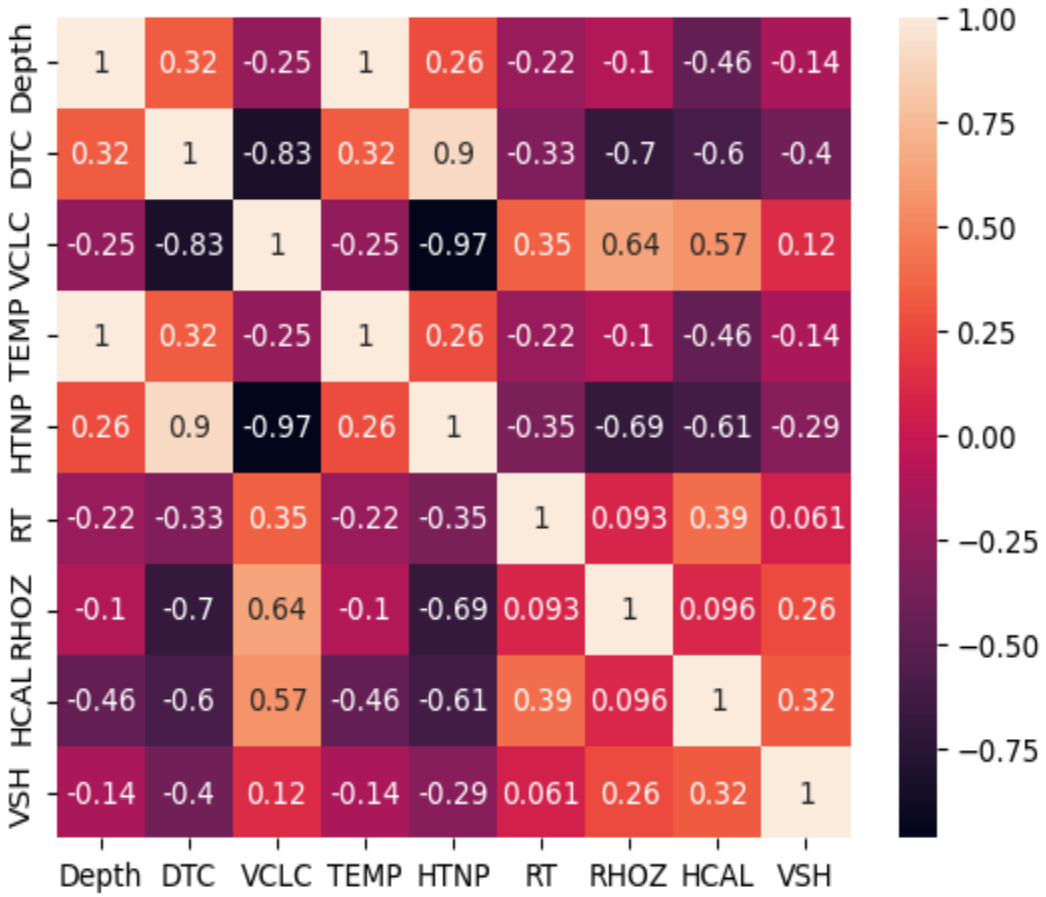

Figure 3-Heatmap showing correlations between well-log parameters: Depth, DTC, VCLC, TEMP, HTNP, RT_HRLT, RHOZ, and HCAL. Color intensity indicates relationship strength and direction, revealing subsurface property interdependencies.

4.3. Model Development

Three machine learning approaches were systematically evaluated for shale volume prediction, each offering distinct advantages and limitations in capturing the complex relationships between petrophysical measurements and shale volume response.

4.3.1. Extra tree

The Extra Trees algorithm demonstrated robust performance with an R^2 of 0.9007 and RMSE of 0.0275 API, indicating strong predictive capability. This ensemble method performance can be attributed to its unique approach of using completely randomized splits, which enhances model diversity and reduces variance.

As evident in Figure 4, the model successfully tracked shale volume trends across most depth intervals, particularly in geologically complex zones. The 4425m boundary, representing a clear lithological transition, was well-resolved, as were the characteristic high-GR shale units between 4300-4350m.

However, residual analysis revealed slightly larger prediction errors in the 4400-4425m interval, likely reflecting the difficulty of the model in capturing abrupt facies changes where the relationship between input features and shale volume response becomes non-stationary. Compared to the other methods, Extra Trees offered the fastest training time, making it particularly suitable for rapid prototyping, though with some compromise in the ultimate predictive accuracy.

Figure 4- Extra Tree model results: (a) Predicted vs. measured Vshale log with depth; (b) Resid-

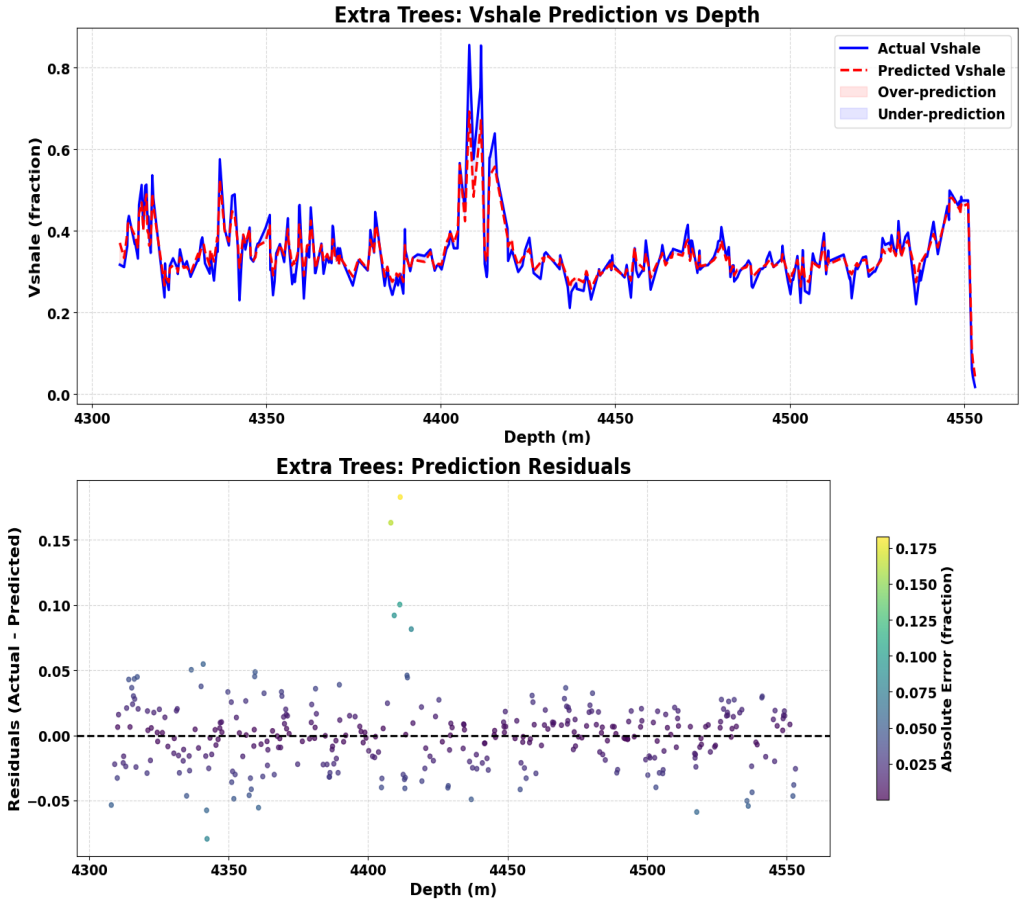

Figure 4- Extra Tree model results: (a) Predicted vs. measured Vshale log with depth; (b) Residual analysis showing systematic errors and depth zones of under/over-prediction.

ual analysis showing systematic errors and depth zones of under/over-prediction.

4.3.2. Random Forest

Building upon the Extra Trees framework, the Random Forest (RF) model achieved superior performance with an R^2 of 0.9640 and RMSE of 0.0166 API. The key distinction lies in RF's use of bootstrap aggregation and optimized split thresholds, which provided more stable predictions. This improvement was particularly noticeable at the 4420m lithological boundary, where the RF model resolved the transition with greater precision than Extra Trees.

The depth-based prediction plot (Figure 5) shows excellent agreement between predicted and actual shale volume values across most intervals, with residuals generally contained within ±3 API units. Interestingly, the model showed slightly elevated errors around 4410m, possibly indicating a localized zone where the feature shale volume relationship deviates from the overall trend. While maintaining the interpretability advantages of tree-based methods through feature importance analysis, the RF model required approximately 30% longer training time than Extra Trees, representing a reasonable tradeoff for its improved accuracy.

4.3.3. Artificial Neural Network

The ANN architecture (8-128-64-32-1) delivered exceptional performance, achieving an R^2 of 0.9793 and RMSE of 0.0126 API - the lowest among all methods tested. This deep learning approach excelled at modeling the complex, non-linear petrophysical relationships through its hierar-

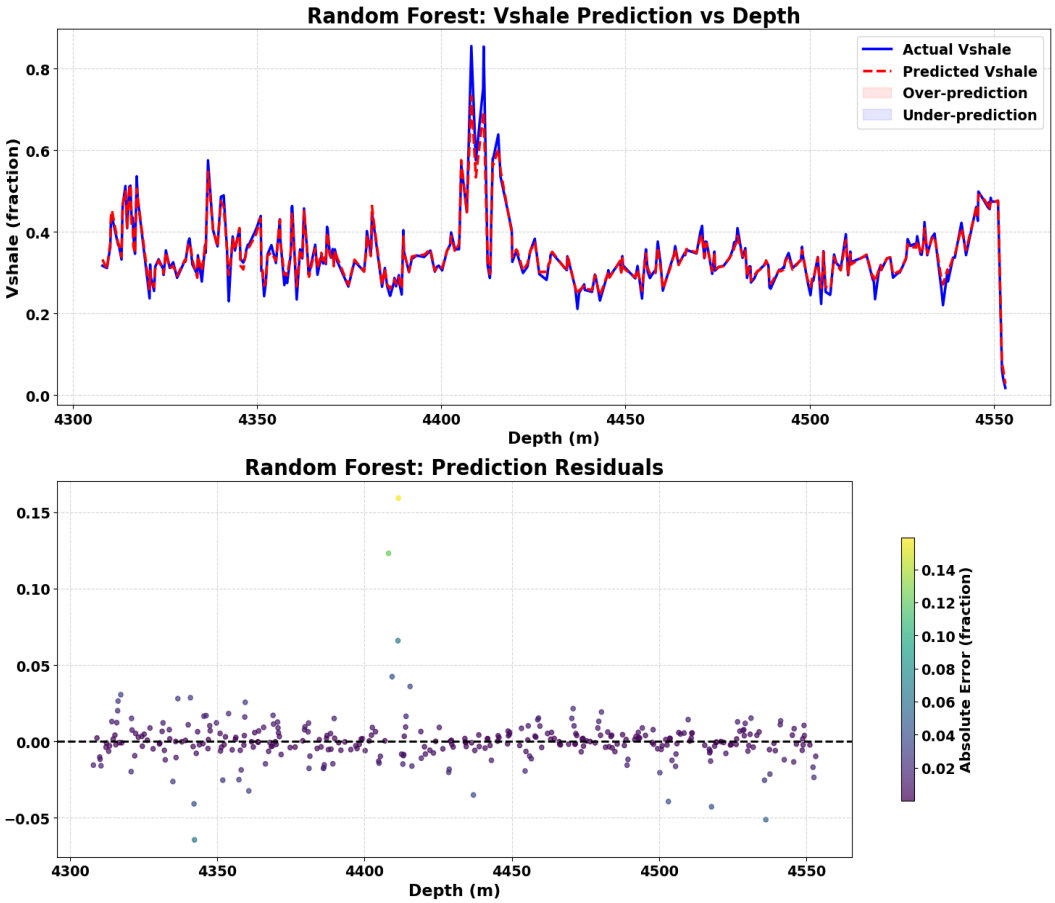

Figure 5-Random Forest model results: (a) Predicted vs. measured shale volume with depth; (b) Residual analysis showing systematic errors and depth zones of under/over-prediction.

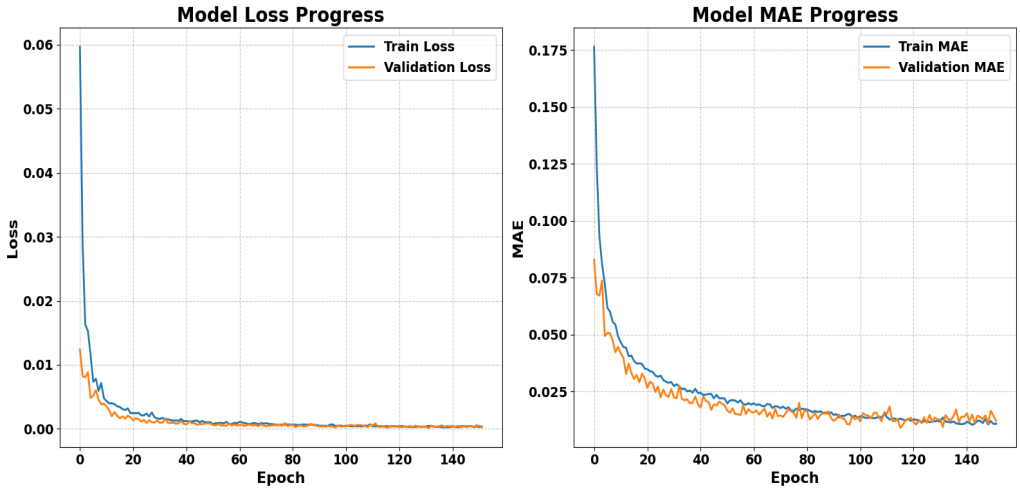

Figure 6- Training curves: (a) MSE and (b) MAE over 125 epochs. Parallel training/validation trends with dropout (30%/20%) shows effective regularization.

chical feature learning capability.

The training process (Figure 6) showed efficient convergence, reaching stability after 152 epochs with early stopping, while dropout regularization (30% in the first hidden layer, 20% in the second) effectively prevented overfitting. The final MAE of 0.0121 API reflects remarkable precision, with errors nearly an order of magnitude smaller than the tree-based methods in some intervals.

Depth-wise analysis (Figure 7) revealed consistently accurate predictions across the entire logged section, with only minor deviations (MAE = 0.008437 ± 0.007868 API) in the 4390-4395m interval. The maximum error of 0.064 API at 4381m represents an isolated case that can be potentially addressed through targeted training data augmentation. While computationally more intensive (requiring 5× more training time than RF) and dependent on careful feature scaling, the ANN performance advantages make it the preferred choice when prediction accuracy is paramount.

4.4. Hyperparameter Optimization

Hyperparameters play a critical role in determining the performance and generalization ability of machine learning models. Unlike model parameters that are learned directly from the data, hyperparameters must be carefully tuned to achieve optimal predictive accuracy and avoid overfitting or underfitting. To ensure reproducibility and robustness, a systematic hyperparameter optimiza-

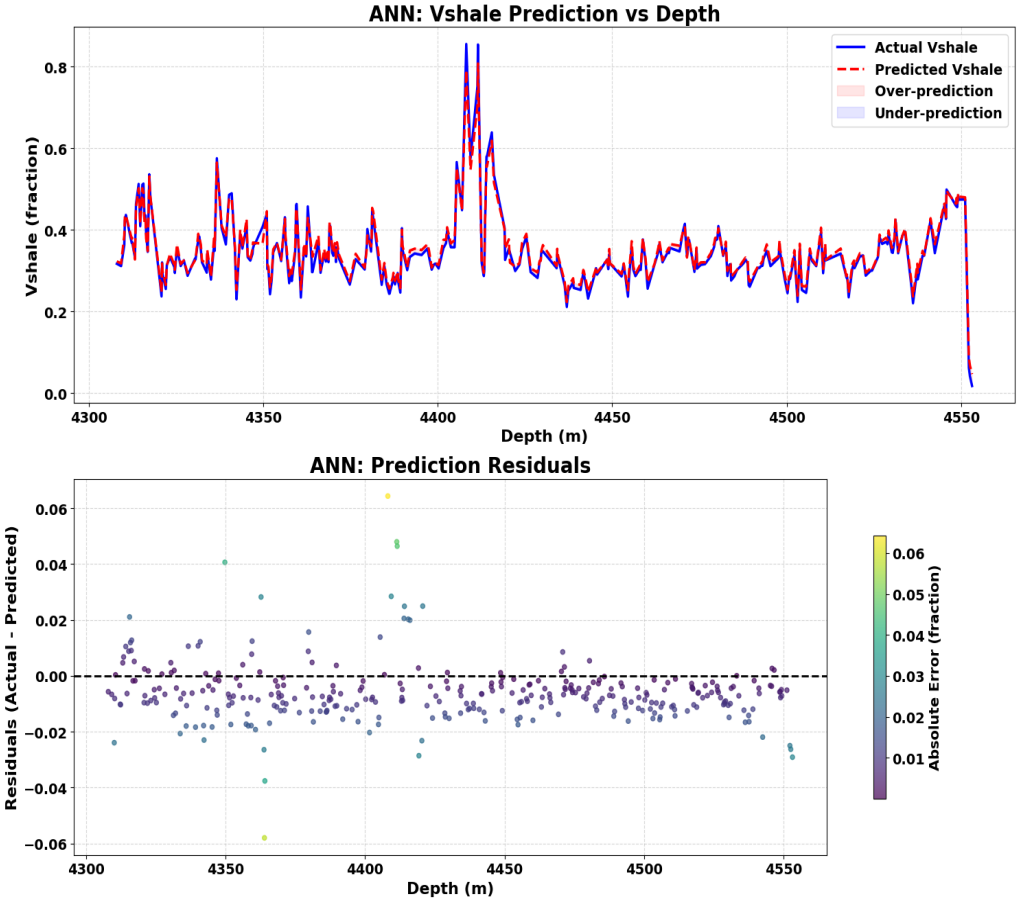

Figure 7- Artificial Neural Network model results: (a) Predicted vs. measured shale volume with depth; (b) Residual analysis showing systematic errors and depth zones of under/over-prediction.

tion process was carried out for each algorithm used in this study. The tuning procedure involved defining search ranges for key hyperparameters, applying cross-validation or validation-based strategies, and selecting the configurations that achieved the best balance between the accuracy and computational efficiency[15]. A detailed summary of the tuned hyperparameters, search strategies, and final selected values for Random Forest (RF), Extra Trees (ET), and Artificial Neural Network (ANN) is provided in Table 3.

The selected hyperparameters (Table 4) were those that provided the best trade-off between predictive accuracy and computational efficiency, as determined by cross-validation and validation set performance.

Table 3 – Summary of hyperparameter optimization for RF, ET, and ANN models

Algorithm	Hyperparameters Tuned	Search Range / Strategy	Final Selected Values	Optimization Method
Random Forest (RF)	n_estimators, max_depth, min_samples_split, max_features	n_estimators: [100–500], step 50; max_depth: [10–30]; min_samples_split: [2–10]; max_features: {sqrt, log2}	n_estimators = 300, max_depth = 20, min_samples_split = 2, max_features = sqrt	Grid Search + 5-fold CV
Extra Trees (ET)	n_estimators, max_depth, max_features	n_estimators: [100–500], step 50; max_depth: {None, 10, 20}; max_features: {sqrt, log2}	n_estimators = 400, max_depth = None, max_features = sqrt	Grid Search + 5-fold CV
Artificial Neural Network (ANN)	Architecture (layers/neurons), Dropout rates, Learning rate, Batch size, Optimizer	Layers: 2–4; Neurons per layer: [32–256]; Dropout: [0.1–0.4]; Learning rate: {0.001, 0.005}; Batch size: {16, 32, 64}; Optimizers: {Adam, RMSprop}	Architecture: (8-128-64-32-1), Dropout: 0.3 / 0.2, Learning rate = 0.001, Batch size = 32, Optimizer = Adam	Randomized Search + Validation Set Performance

4.5. Comparative Analysis of Machine Learning Model Performance for Shale Volume Prediction

The three approaches form a clear performance hierarchy in terms of prediction accuracy, with ANN outperforming RF, which in turn surpasses Extra Trees. This progression mirrors the increasing model complexity and computational requirements of each method. Table 4 and Figure 8 illustrate the detailed model performance comparisons and visualizations, respectively.

Table 4- Comparative Performance of Extra Trees, Random Forest, and ANN Models

Metrics		Extra Tree	Random Forest	ANN
R^2	Train	0.94	0.99	0.98
	Test	0.90	0.96	0.98
RMSE	Train	0.020	0.009	0.012
	Test	0.027	0.017	0.013
Training Time		Lowest	Moderate	Highest
Interpretability		High	High	Low
Key Strength		Fast prediction	Balanced accuracy	Best accuracy

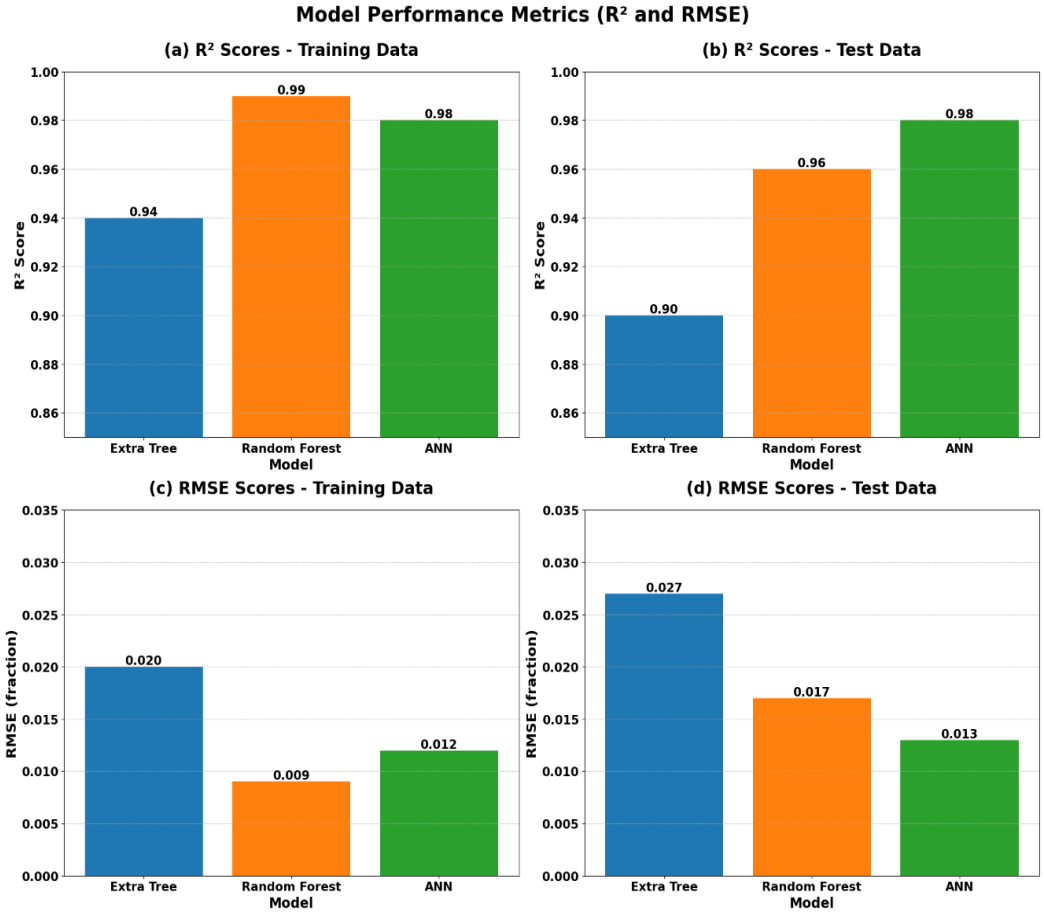

Figure 8-Comparison of Model Performance Based on R^2 and RMSE for training and testing data

5. Conclusion

This study thoroughly evaluated three machine learning models—Extra Trees, Random Forest, and Artificial Neural Networks —for predicting shale volume using the conventional well logs. The models showed a clear progression in performance, with the ANN delivering outstanding accuracy, achieving an R^2 of 0.9779 and an RMSE of 0.0130 API. The RF model followed with an R^2 of 0.9640, and Extra Trees provided a respectable R^2 of 0.9007.

The ANN model consistently outperformed the tree-based methods. Its ability to effectively capture complex, nonlinear relationships between the input well logs and the shale volume response was a key factor. The ANN's hierarchical feature extraction process allowed for precise predictions, even in areas with challenging lithological transitions. Notably, its prediction errors were 60–77% lower than those from the tree-based approaches.

While the tree-based methods (Random Forest and Extra Trees) had slightly lower accuracy, they offered distinct advantages. These models provided immediate interpretability through their native feature importance metrics, making it easy to understand which well log parameters were most influential in their predictions. They also boasted significantly faster training times, being five times quicker than the ANN. Furthermore, tree-based models performed robustly without requiring the often-complex step of feature scaling.

From a geological insight perspective, all models encountered minor challenges when predicting at abrupt lithology boundaries. This suggests that future research can benefit from incorporating spatial context or sequence modeling techniques.

The ANN's consistent accuracy across various shale zones particularly highlights its strong potential for the detailed characterization of unconventional reservoirs.

Based on these findings, we offer several practical recommendations. For time-sensitive applications where a moderate level of accuracy is sufficient, the Random Forest model presents the best balance of speed and performance. However, when maximum accuracy is paramount—such as in critical reservoir modeling or geosteering operations—the ANN is the preferred choice, even with its higher computational cost. Regarding interpretability, tree-based methods provide immediate insights into feature rankings. While interpreting an ANN decisions might require advanced techniques like SHAP analysis, this is often a worthwhile trade-off for the superior predictive gains it offers.

Ultimately, this work demonstrates the transformative potential of machine learning in petrophysical analysis. The ANN stands out as a powerful tool for high-fidelity shale volume prediction, provided that users are prepared to accommodate its computational demands and complex interpretability characteristics. The final selection of a method should always align with specific project needs, balancing accuracy, speed, and explainability.

Although the algorithms themselves are well-established, the strength of this study lies in its systematic comparison and practical recommendations for the model selection in real-world reservoir settings. The direct evaluation of ANN, RF, and ET on actual well log data from southern Iran provides unique insights into their relative strengths and limitations. This applied novelty ensures that the findings are not only academically relevant but also directly useful for petroleum engineers and geoscientists involved in reservoir evaluation and development.

References

[1] Bettir, N., et al. Improved Shale Volume Prediction Using Machine Learning Algorithms in Complex Reservoirs. in ARMA US Rock Mechanics/ Geomechanics Symposium. 2024. ARMA.

[2] Ma, Y., et al., Influence of Rock Fabric on Physical Properties of Shale Oil Reservoir Under Effective Pressure Conditions. Lithosphere, 2024. 2024(2): p. lithosphere_2023_338.

[3] Mabiala Mbouaki, A.P., et al., Petrophysical Evaluation of a Shaly Sandstone Reservoir and the Effect of Clay Minerals on Reservoir Quality: A Case Study from the Barremian Mengo Sandstone, Kouilou Basin, Republic of Congo. ACS omega, 2025. 10(10): p. 10081-10106.

[4] Ganguli, S.S. and V.P. Dimri, Reservoir Characterization, Modeling and Quantitative Interpretation: Recent Workflows to Emerging Technologies. Vol. 6. 2023: Elsevier.

[5] Bin, W., et al., Experimental study on hydraulic fracture propagation behavior in heterogeneous shale formations. Frontiers in Energy Research, 2024. 11: p. 1309591.

[6] Gong, X., X. Ma, and Y. Liu, Analysis of geological factors affecting propagation behavior of fracture during hydraulic fracturing shale formation. Geomechanics and Geophysics for Geo-Energy and Geo-Resources, 2024. 10(1): p. 102.

[7] Belyadi, H., E. Fathi, and F. Belyadi, Hydraulic fracturing in unconventional reservoirs: theories, operations, and economic analysis. 2019: Gulf Professional Publishing.

[8] Asquith, G.B. and C.R. Gibosn, Basic well log analysis for geologists. 1982: American Association of Petroleum Geologists.

[9] Vu, D.H. and H.T. Nguyen, Estimation of shale volume from well logging data using Artificial Neural Network. Tạp chí Khoa học kỹ thuật Mỏ-Địa chất, 2021: p. 46-52.

[10] Szabó, N.P., Shale volume estimation based on the factor analysis of well-logging data. Acta Geophysica, 2011. 59: p. 935-953.

[11] Zhang, H. and W. Wu, Shale content prediction of well logs based on CNN-BiGRU-VAE neural network. Journal of Earth System Science,

2023. 132(3): p. 139.

[12] Ebrahimi, P., et al., Shale volume estimation using machine learning methods from the southwestern fields of Iran. Results in Engineering, 2025. 25: p. 104506.

[13] Dehghani, M., S. Jahani, and A. Ranjbar, Comparing the performance of machine learning methods in estimating the shear wave transit time in one of the reservoirs in southwest of Iran. Scientific Reports, 2024. 14(1): p. 4744.

[14] Mondal, D., V. Srivardhan, and B. Singh. A Wavelet and Neural Network based approach towards determination of Shale Volume using well logs of Indian Coalfields. in 79th EAGE Conference and Exhibition 2017. 2017. EAGE Publications BV.

[15] Mohammadinia, F., et al., Shale volume estimation using ANN, SVR, and RF algorithms compared with conventional methods. Journal of African Earth Sciences, 2023. 205: p. 104991.

[16] Ardebili, P.N., G. Jozanikohan, and A. Moradzadeh, Estimation of porosity and volume of shale using artificial intelligence, case study of Kashafrud Gas Reservoir, NE Iran. Journal of Petroleum Exploration and Production Technology, 2024. 14(2): p. 477-494.

[17] Huang, S., et al., Support vector regression based on the particle swarm optimization algorithm for tight oil recovery prediction. ACS omega, 2021. 6(47): p. 32142-32150.

[18] Meng, Y., et al., Prediction of Total Organic Carbon Content in Shale Based on PCA-PSO-XGBoost. Applied Sciences, 2025. 15(7): p. 3447.

[19] Wang, T., et al., Productivity prediction of fractured horizontal well in shale gas reservoirs with machine learning algorithms. Applied Sciences, 2021. 11(24): p. 12064.

[20] Yang, R., et al., Long short-term memory suggests a model for predicting shale gas production. Applied Energy, 2022. 322: p. 119415.

[21] Garcia-Cifuentes, K., et al., Identification of Extended Emission Gamma-Ray Burst Candidates Using Machine Learning. The Astrophysical Journal, 2023. 951(1): p. 4.

[22] Gao, P., et al., Influence of dispersion and stabilization of active metals on Ni-Cu/AC catalyst on gas phase carbonylation of ethanol. Fuel, 2021. 292: p. 120308.

[23] Ye, L., et al., Application of machine learning in cosmic ray particle identification. ACTA PHYSICA SINICA, 2023. 72(14).

[24] Kuran, F., G. Tanırcan, and E. Pashaei, Developing machine learning-based ground motion models to predict peak ground velocity in Turkiye. Journal of Seismology, 2024. 28(5): p. 1183-1204.

[25] Ali, M., Machine learning based shale volume prediction from the Norwegian North Sea. 2021, uis.

[26] Han, D. and S. Kwon, Application of machine learning method of data-driven deep learning model to predict well production rate in the shale gas reservoirs. Energies, 2021. 14(12): p. 3629.

[27] Rezaei Mirghaed, B., A. Dehghan Monfared, and A. Ranjbar, Enhanced petrophysical evaluation through machine learning and well logging data in an Iranian oil field. Scientific Reports, 2024. 14(1): p. 28941.

[28] Breiman, L., Random forests. Machine learning, 2001. 45: p. 5-32.

[29] Geurts, P., D. Ernst, and L. Wehenkel, Extremely randomized trees. Machine learning, 2006. 63: p. 3-42.

[30] Goodfellow, I., et al., Deep learning. Vol. 1. 2016: MIT press Cambridge.

[31] Friedman, J., T. Hastie, and R. Tibshirani, Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 2000. 28(2): p. 337-407.

[32] Hastie, T., et al., Random forests. The elements of statistical learning: Data mining, inference, and prediction, 2009: p. 587-604.

[33] Liaw, A. and M. Wiener, Classification and regression by randomForest. R news, 2002. 2(3): p. 18-22.

[34] Ho, T.K., The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 1998. 20(8): p. 832-844.

[35] Pal, M., Random forest classifier for re-

mote sensing classification. International journal of remote sensing, 2005. 26(1): p. 217-222.

[36] Quinlan, J.R., Induction of decision trees. Machine learning, 1986. 1: p. 81-106.

[37] Loh, W.Y., Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2011. 1(1): p. 14-23.

[38] Kononenko, I., Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 2001. 23(1): p. 89-109.

[39] Pedregosa, F., et al., Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011. 12: p. 2825-2830.

[40] Srivastava, N., et al., Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014. 15(1): p. 1929-1958.

[41] Goodfellow, I., Deep learning. 2016, MIT press.

[42] Goodfellow, I.J., et al., An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211, 2013.

[43] Nair, V. and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. in Proceedings of the 27th international conference on machine learning (ICML-10). 2010.

[44] Géron, A., Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts. Aurélien Géron-Google Kitaplar, yy https://books. google. com. tr/books, 2019.

[45] Duda, R.O. and P.E. Hart, Pattern classification. 2006: John Wiley & Sons.

[46] Kingma, D.P., Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[47] Chicco, D., M.J. Warrens, and G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj computer science, 2021. 7: p. e623.

[48] Willmott, C.J. and K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 2005. 30(1): p. 79-82.

[49] Ebrahimi, A., et al., Estimation of shear wave velocity in an Iranian oil reservoir using machine learning methods. Journal of Petroleum Science and Engineering, 2022. 209: p. 109841.

[50] Ebrahimi, P., et al., Young's Modulus Estimation Using Machine Learning Methods and Daily Drilling Reports. Journal of Oil, Gas and Petrochemical Technology, 2023. 10(1): p. 1-24.

[51] Akbari, A., et al., Enhanced water saturation estimation in hydrocarbon reservoirs using machine learning. Scientific Reports, 2025. 15(1): p. 29846.

نشریه فناوری نفت، گاز و پتروشیمی، ۱۲(۲): ۴۵-۶۵، تابستان و زمستان ۱۴۰۴

مقاله تحقیقاتی

هوش مصنوعی در توصیف مخزن: پیش‌بینی حجم شیل با استفاده از شبکه‌های عصبی مصنوعی، جنگل تصادفی، و درختان فوق‌العاده

محمدیاسین حسینی [۱]، علی رنجبر [۲*]، محمد مهدی حسینی [۳]

۱.گروه مهندسی نفت، دانشکده مهندسی نفت، گاز و پتروشیمی، دانشگاه خلیج فارس، بوشهر

۲.*نویسنده مسئول ali.ranjbar@pgu.ac.ir

مشخصات مقاله

تاریخچه مقاله:

دریافت ۱۰ خرداد ۱۴۰۴

دریافت پس از اصلاح ۷ شهریور ۱۴۰۴

پذیرش نهایی ۲۷ شهریور ۱۴۰۴

کلمات کلیدی:

پیش بینی حجم شیل

داده‌های چاه‌نگاری

یادگیری ماشین

شبکه عصبی مصنوعی

جنگل تصادفی

درختان فوق‌العاده

*عهده‌دار مکاتبات؛

رایانامه: Ali.ranjbar@pgu.ac.ir

تلفن:

چکیده

حجم شیل برای شناسایی لیتولوژی (سنگ‌شناسی)، ارزیابی مخزن و همبستگی چینه‌شناسی در تحلیل سازندهای زیرسطحی ضروری است. این مطالعه عملکرد سه الگوریتم یادگیری ماشین—جنگل تصادفی (FR)، درختان فوق‌العاده (TE)، و شبکه عصبی مصنوعی (NNA)—را برای پیش‌بینی حجم شیل با استفاده از داده‌های معمول چاه‌نگاری ارزیابی می‌کند. مدل‌ها بر روی یک مجموعه داده جامع از یکی از میادین نفتی در جنوب ایران آموزش و آزمایش شدند، که شامل پارامترهایی مانند زمان انتقال امواج صوتی (CTD)، چگالی توده (ZOHR)، مقاومت (TR)، تخلخل نوترونی (PNTH)، و کالیپر (LACH) است. نتایج نشان داد که NNA با ضریب تبیین (R^2) ۹۷۷۹,۰ و خطای میانگین مربعات ریشه (ESMR) IPA 0310,0 به دقت بالاتری دست یافت که بهتر از FR (R^2 = 0.0469, ESMR = IPA 6610.0) و TE (R^2 = 0.7009, ESMR = IPA 5720.0) بود. در حالی که NNA در درک روابط غیرخطی پیچیده عالی عمل کرد، روش‌های مبتنی بر درخت زمان آموزش سریع‌تری را ارائه می‌دهند و از طریق معیارهای اهمیت ویژگی، قابلیت تفسیر بیشتری دارند. این یافته‌ها NNA را به عنوان گزینه ارجح برای پیش‌بینی حجم شیل با دقت بالا برجسته می‌کنند، در حالی که FR یک راه‌حل متعادل برای سناریوهایی فراهم می‌کند که هم به سرعت و هم به دقت رقابتی نیاز دارند. این مطالعه بر پتانسیل تحول‌آفرین یادگیری ماشین در تحلیل پتروفیزیکی تأکید کرده و توصیه‌های کاربردی برای انتخاب مدل بر اساس نیازهای خاص پروژه ارائه می‌دهد.

نحوه استناد به این مقاله:

Hosseini M Y, Ranjbar A, Hosseini M M, Artificial Intelligence in Reservoir Characterization: Predicting Shale Volume with ANN, RF, and ET, Journal of Oil, Gas and Petrochemical Technology, 2025; 12(2): 45-65. 10.22034/jogpt.2025.527316.1141.