applied sciences

Article

Advancing Geotechnical Evaluation of Wellbores: A Robust and Precise Model for Predicting Uniaxial Compressive Strength (UCS) of Rocks in Oil and Gas Wells

Mohammadali Ahmadi

Department of Chemical and Petroleum Engineering, University of Calgary, 2500 University Dr. NW, Calgary, AB T2N 1N4, Canada; mohammadali.ahmadi@ucalgary.ca

Abstract: This study examines the efficacy of various machine learning models for predicting the uniaxial compressive strength (UCS) of rocks in oil and gas wells, which are essential for ensuring wellbore stability and optimizing drilling operations. The investigation encompasses Linear Regression, ensemble methods (including Random Forest, Gradient Boosting, XGBoost, and LightGBM), support vector machine-based regression (SVM-SVR), and multilayer perceptron artificial neural network (MLP-ANN) models. The results demonstrate that XGBoost and Gradient Boosting offer superior predictive accuracy for UCS in drillability, as indicated by low Mean Absolute Percentage Error (MAPE) values of 3.87% and 4.18%, respectively, and high R^2 scores (0.8542 for XGBoost). These models emerge as optimal choices for UCS prediction focused on drillability, offering increased accuracy and reliability in practical engineering scenarios. Ensemble methods and MLP-ANN emerge as frontrunners, providing valuable tools for improving wellbore stability assessments, optimizing drilling parameter selection, and facilitating informed decision-making processes in oil and gas drilling operations. Moreover, this study lays a foundation for further research in drillability-centred predictive modelling for geotechnical parameters, advancing our understanding of rock behaviour under drilling conditions.

Keywords: uniaxial compressive strength; wellbore stability; drilling; least squares support vector machine; prediction

check for updates

Citation: Ahmadi, M. Advancing Geotechnical Evaluation of Wellbores: A Robust and Precise Model for Predicting Uniaxial Compressive Strength (UCS) of Rocks in Oil and Gas Wells. *Appl. Sci.* **2024**, *14*, 10441. https://doi.org/10.3390/app142210441

Academic Editor: Stefano Invernizzi

Received: 24 August 2024
Revised: 5 November 2024
Accepted: 7 November 2024
Published: 13 November 2024

1. Introduction

Geotechnical evaluation of wellbores, which is crucial for hydraulic fracturing design, sand production management and prediction, fault stability, reactivation analysis, and wellbore stability, requires accurate and comprehensive knowledge of the mechanical properties of rocks [1].

The uniaxial compressive strength (UCS) of rocks is a fundamental measure of drillability in geotechnical engineering, particularly in the oil and gas sectors. UCS quantifies the maximum axial load a rock can endure before failure, reflecting the rock's resistance to compressive forces. In drilling applications, higher UCS values signal harder formations, which demand increased energy and precise adjustments to drilling parameters—such as bit type, weight on bit, and rotary speed—to achieve efficient penetration. Accurately estimating UCS enables engineers to anticipate the energy requirements and bit wear, aiding in optimal drilling parameter selection and cost reduction.

UCS significantly influences the rate of penetration (ROP) in drilling. Rocks with lower UCS allow for higher ROP, facilitating faster drilling and minimizing tool wear, while those with higher UCS reduce ROP, requiring adjustments to drilling methods and equipment. Therefore, reliable UCS prediction models are vital for enhancing drillability assessments, proactively managing wellbore stability and improving drilling efficiency in complex geological formations.

One commonly used method to evaluate these properties is through the measurement of UCS in the laboratory [2]. However, this method can be highly sensitive to the loading process of the core sample and is destructive in nature [3]. As an alternative, log-based methods that indirectly measure rock strength have been proposed, but their precision and accuracy have not been fully validated by reliable data [4]. Given the challenges and expenses associated with standard laboratory tests, indirect methods are more promising for practical use [2].

Sonic travel time is a physical property of rocks that is utilized for studying rock mechanics and reservoir evaluation. This property varies based on the lithology, rock textures, fluid content, and porosity of the rock [5]. In cases where limited core samples are available, sonic and neutron logs can be used to estimate the rock properties. Researchers have developed six experimental equations related to carbonate rock strength for measuring geophysical properties, which are listed in Table 1 [6–10]. Another approach for predicting rock strength is to use drilling data based on ROP models [11,12]. These models can be used with all types of bits, although tri-cone bits (TCBs) are preferred due to their wide range of use [13].

Table 1. Empirical relationships between UCS and petro-physics logs in carbonates.

Formula	Region	Remarks
$\mathrm{UCS} = 143.8 \times \exp(-6.95\varnothing)$	Middle East	$0.05 < \phi < 0.2$ and $30 < \mathrm{UCS} < 150$ MPa
$\mathrm{UCS} = 135.9 \times \exp(-4.8\varnothing)$	-	$0 < \phi < 0.2$ and $10 < \mathrm{UCS} < 300$ MPa
$\mathrm{UCS} = 276(1 - 3\varnothing)^2$	Korobcheyev deposit, Russia	-
$\mathrm{UCS} = \left(\frac{7682}{\Delta t}\right)^{1.82}/145$	-	-
$\mathrm{UCS} = 10^{\left(2.44+\frac{109.14}{\Delta t}\right)}/145$	-	-
$\mathrm{UCS} = 7600 \times \exp(-0.064\Delta t)$	Middle East	-

Koolivand-Salooki et al. [14] developed a method for determining the UCS of rock formations using Genetic Programming (GP). This approach utilized parameters such as total formation porosity, Bulk Density, and water saturation obtained from various logging techniques, including sonic, neutron, gamma ray, and electric logs. The elastic moduli were derived from compressional and shear sonic logs using mathematical correlations, and the rock UCS was estimated using empirical correlations by Wang and Plumb. The study involved analyzing approximately 5000 data points from three wells in an Iranian oil field to develop the GP model for UCS prediction. The model was fine-tuned using UCS data from core samples and validated with two separate datasets. The estimated UCS values from the GP model closely matched those obtained from analytical methods based on well-log data [14].

McElroy and colleagues [15] introduced an ANN modelling approach for predicting the UCS of oil well cement, specifically class "H". This research analyzed 195 cement samples, incorporating varying concentrations of pre-dispersed nanoparticles, including nanosilica (nano-SiO_2), nanoalumina (nano-Al_2O_3), and nanotitanium dioxide (nano-TiO_2), across different temperature conditions. The effectiveness of these nanoparticles was assessed through transmission electron microscopy (TEM) images. The ANN model included one input layer, one hidden layer, and one output layer. Its performance was superior to that of Multi-Linear Regression (MLR) and Random Forest (RF) regression algorithms in terms of statistical accuracy. Based on their findings, the developed ANN model was a highly accurate and non-destructive alternative approach to traditional UCS tests, offering cost and time-saving advantages to the petroleum industry [15].

Hiba et al. [16] carried out a study to investigate the geomechanical parameters used for field planning and development, specifically focusing on the tensile and UCS values of rock. Given the time-consuming nature of laboratory measurements, the researchers employed non-destructive techniques to expedite and enhance the reliability of predictions.

They used an ANN to predict Ts and UCS based on drilling data obtained from two Middle Eastern fields. The ANN was highly accurate in predicting both parameters during the training phase and was effective in predicting them during the testing and validation phases with an average AAPE of 0.59% [16].

Ibrahim et al. [17] investigated the use of machine learning to predict the UCS and tensile strength (T0) of carbonate rocks from well-log data. They utilized RF and decision tree (DT) algorithms on data from a Middle Eastern reservoir, identifying gamma ray, compressional time, and Bulk Density as key predictive factors. The study found both models to be highly accurate, with RF slightly outperforming DT. Specifically, RF achieved a correlation coefficient (R) of 0.97 and an absolute average percentage error (AAPE) of 0.65% for UCS prediction, and an R of 0.99 and AAPE of 0.28% for T0. These results suggest that machine learning offers a reliable and efficient method for estimating rock strength parameters, though further research is needed for other geological formations [17].

This study aims to evaluate the effectiveness of various machine learning models in predicting the UCS of rocks within the context of oil and gas wells, which is a key factor in maintaining wellbore stability and optimizing drilling operations. By comparing Linear Regression, ensemble methods (such as Random Forest, Gradient Boosting, XGBoost, and LightGBM), support vector machine-based regression (SVM-SVR), and multilayer perceptron artificial neural network (MLP-ANN) models, this research seeks to identify the most reliable and accurate approaches. This study not only highlights the importance of selecting suitable machine learning models for geotechnical applications but also advances the field by providing valuable insights into rock behaviour under drilling conditions. These insights pave the way for further research, ultimately improving the understanding of geomechanical properties and their impact on drilling operations.

2. Different ROP Models

The ROP directly affects the drilling cost per foot drilled. Tommy Warren (1981) [18] proposed a two-term ROP approach that could be employed for drilling optimization purposes, drilling conditions, formation properties, and bit type. Another model with three terms was derived in 1987 by Warren for tri-cone bits (TCBs) named the three-term approach. Hareland and Hoberock improved the approach in 1993 to include bit wear out, differential pressure impact, and hole cleaning factors [19]. There are other models defined for Polycrystalline Diamond Compact (PDC) bits. For example, Hareland and Rampersad proposed a drag bit model in 1994 [20].

2.1. Modified Warren Model

The central idea for this model comes from the scientific fact that under steady-state drilling circumstances, the removal rate is equal to the rate at which chips are formed. Therefore, ROP is regulated by the formation cutter process, the cutting removal process, and a combination of these factors mentioned above. This approach correlates ROP to rotary speed, weight on bit (WOB), bit size, and rock strength from generalized response curves and dimensional analysis, which is formulated as follows [12,19]:

$$\mathrm{ROP} = \mathrm{W_f}\left(\mathrm{f_c}(\mathrm{P_e})\left(\frac{\mathrm{aS^2D_{bit}}^3}{\mathrm{RPM}\cdot\mathrm{WOB}^2} + \frac{\mathrm{b}}{\mathrm{RPM}\cdot\mathrm{D_{bit}}}\right) + \frac{\mathrm{c}\rho\mu\mathrm{D_{bit}}}{\mathrm{F_{jm}}}\right)^{-1} \tag{1}$$

where ROP stands for the drilling rate in terms of (ft/h), $\mathrm{D_{bit}}$ denotes the bit diameter in terms of (in), S represents the confined rock compressive strength in terms of (psi), WOB represents the weight on the bit in terms of ($\mathrm{lb_f}$), RPM stands for the rotary speed in terms of (rev/min), μ stands for the plastic viscosity in terms of (cp), ρ denotes the mud density in terms of (ppg), $\mathrm{F_{jm}}$ represents the modified impact force in terms of ($\mathrm{lb_f}$), fc($\mathrm{P_e}$) denotes the chip hold-down function (dimensionless), $\mathrm{W_f}$ stands for the bit wear dimensionless factor, and a, b, and c represent the bit dimensional parameters. The first part of the equation defines the rock breaking-up rate. The second term accounts for the distribution of WOB, as WOB is increased with the number of teeth and the teeth penetrating further into the rock. As noted earlier, F_{jm} stands for the modified impact force, which is formulated as follows:

$$\mathbf{F_{jm}} = \left(\mathbf{1} - \mathbf{A_v}^{-\mathbf{0.122}}\right) \times \mathbf{F_j} \tag{2}$$

where A_v is the jet-to-fluid returning velocity ratio, and then A_v (for three jets) can be obtained using Equation (3):

$$\mathbf{A_v} = \frac{\mathbf{V_n}}{\mathbf{V_f}} = \frac{\mathbf{0.15d_b}^{\mathbf{2}}}{\mathbf{3d_n}^{\mathbf{2}}} \tag{3}$$

where d_n is the nozzle diameter (in.) and d_b is the bit diameter (in.). For a fixed value of impact force and a fixed bit size from Equation (4), the measured impact force must be independent of nozzle size as follows:

$$\mathbf{F_j} = \mathbf{0.000516\rho q v_n} \tag{4}$$

where q represents the pump flow rate in terms of (gpm) and V_n stands for the nozzle velocity in terms of (ft/s). To consider the resultant force on a formation cutting caused by the bit, which is called the chip hold-down function, Equation (5) can be used:

$$f_c(P_e) = c_c + a_c(P_e - 120)^{b_c} \tag{5}$$

where P_e stands for the effective differential pressure. a_c, b_c, and c_c are lithology and permeability-dependent parameters, which are demonstrated in Table 2 [19]. Circulation pressure at the bottom hole is defined as the summation of the annulus pressure drop and static mud column pressure. This variable can be determined through the following equation:

$$P_{ECD} = 0.052\rho TVD + \Delta P_{ann} \tag{6}$$

where TVD is true vertical depth in ft and P_{ECD} is in psi. During the drilling operation, the teeth of a bit start to wear out. Teeth area increment is due to bit wear that reduces stress on each cutter. Hareland developed the below equation to determine the bit wear:

$$\mathbf{W_f} = \mathbf{1} - \frac{\mathbf{\Delta BG}}{\mathbf{8}} \tag{7}$$

$$\mathbf{\Delta BG} = \mathbf{W_c}\sum \boldsymbol{WOB_i \cdot RPM_i \cdot A_{abr_i} \cdot S_i} \tag{8}$$

A_{abri} stands for the relative rock abrasiveness and W_c represents the wear coefficient. Therefore, an inverted ROP model can measure rock strength provided that the drilling condition is actual [12]. The apparent rock strength log (ARSL) along the wellbore can be determined through the below equation:

$$\mathbf{S} = \sqrt{\frac{\mathbf{RPM \cdot WOB^2}}{\mathbf{a \cdot f_c(P_e) \cdot ROP \cdot D_{bit}}^{\mathbf{3}}} - \frac{\mathbf{b \cdot WOB^2}}{\mathbf{a \cdot D_{bit}}^{\mathbf{4}}} - \frac{\mathbf{c \cdot \rho \cdot \mu \cdot RPM \cdot WOB^2}}{\mathbf{a \cdot f_c(P_e) \cdot F_{jm} \cdot D_{bit}}^{\mathbf{2}}}} \tag{9}$$

Table 2. Chip hold-down coefficients.

Formation	Permeable	Impermeable
P_e	$P_h - P_p$	P_h
a_c	0.0050	0.014
b_c	0.7570	0.470
c_c	0.1030	0.569

The rock strength from Equation (15) is at the bit operation condition at the bottom of the hole. In conventional drilling operations, the hydrostatic pressure caused by the mud is higher than the pore pressure. In order to calculate the unconfined counterpart, a failure index should be defined as follows [18,21]:

$$\boldsymbol{S_0 = \frac{S}{(1 + a_s \cdot {p_e}^{b_s})}} \tag{10}$$

where a_s and b_s are failure criteria fitting constants.

2.2. Drag Bit Models

Drag bits have fixed cutter blade parts integrated into the body. PDC-bit ROP equations are employed to estimate confined rock compressive strength [22] as follows:

$$S = \frac{WOB}{N_c \cdot A_p} \tag{11}$$

$$A_p = \sin\theta \left[(\frac{d_c}{2})^2 \cos^{-1}\left(1 - \frac{2P}{d_c \cos\theta}\right) - \left(\frac{d_c P}{2\cos\theta}\right)\left(\frac{P d_c}{\cos\theta} - \frac{P^2}{\cos^2\theta}\right)^{0.5} \right] \tag{12}$$

$$R_e = \frac{D_{bit}}{2\sqrt{2}} \tag{13}$$

$$A_v = \cos\alpha \sin\theta \left[(\frac{d_c}{2})^2 \cos^{-1}\left(1 - \frac{2P}{d_c \cos\theta}\right) - \left(\frac{d_c P}{2\cos\theta}\right)\left(\frac{P d_c}{\cos\theta} - \frac{P^2}{\cos^2\theta}\right)^{0.5} \right] \tag{14}$$

$$ROP = W_f \frac{14.14 N_c (RPM) A_v}{d_{bit}} \tag{15}$$

where α is the cutter side rake angle and θ is the cutter back rake angle. Equation (15) gives the output volume from each cutter (A_v) in in^2 at a PDC bit. Via Equations (13) and (14), the penetration of the PDC cutter can be calculated. To estimate each cutter's projected contact area, the penetration of each PDC is used with Equation (12); the confined compressive strength can be calculated by Equation (11). The level of wear out of the bit is obtainable using Equations (7) and (8).

2.3. Shortcomings of Traditional ROP Models

Although ROP models like the Modified Warren and drag bit models provide useful theoretical frameworks, they lack flexibility when applied to diverse formations or varying drilling conditions. Factors such as bit wear, formation heterogeneity, and variations in fluid properties significantly affect the accuracy of these models. Traditional models also rely on empirical coefficients that are specific to particular formations or bit types, limiting their generalizability across different geological contexts. Furthermore, these models often fail to adapt to real-time data, resulting in a lag between field observations and model predictions, which can impact decision-making during drilling operations.

To address these limitations, this study leverages machine learning models to predict UCS with improved accuracy, enabling better estimates of drillability and ROP. Machine learning approaches, particularly ensemble methods, can overcome the limitations of traditional models by capturing non-linear relationships and adapting to real-time data changes. By enhancing UCS predictions, these models support more reliable drillability assessments across a wide range of geological conditions, ultimately improving drilling efficiency and cost management.

3. Theory

In this study, we employed a diverse array of machine learning models renowned for their efficacy in predictive tasks. The selected models encompassed both traditional and ensemble learning approaches, aiming to comprehensively evaluate their performance in predicting the uniaxial compressive strength (UCS) of rocks in oil and gas wells.

3.1. Linear Regression

Linear Regression, a fundamental statistical method, serves as a baseline model in our analysis. It establishes a linear relationship between the input features (X) and the target variable (y). The model's goal is to find the best-fitting linear equation, expressed as follows [23,24]:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{16}$$

where β represents the coefficients and ϵ denotes the error term. Linear regression assumes a linear relationship between the independent and dependent variables and is sensitive to outliers [25].

3.2. Random Forest

Random Forest is an ensemble learning technique renowned for its robustness and versatility in predictive modelling tasks [26]. It operates by constructing multiple decision trees during training, where each tree is trained on a bootstrap sample of the data and makes predictions independently [27]. The final prediction is then derived by aggregating the predictions of all individual trees, typically through averaging for regression tasks or voting for classification tasks. The output of a Random Forest model can be formulated as follows [26]:

$$y_{i,\ predicted} = \sum_{j=1}^{N} f_j(x_j) \tag{17}$$

where $y_{i,\ predicted}$ represents the predicted value for the i-th data point, N is the total number of trees in the forest, and $f_j(x_j)$ denotes the prediction of the j-th decision tree for the i-th data point.

Each decision tree in the Random Forest is trained using a subset of the available features, typically selected randomly at each split. This random feature selection helps to decorrelate the trees and enhance the diversity of the ensemble, thereby reducing the risk of overfitting [28].

Random Forest offers several advantages over individual decision trees. Firstly, it is capable of capturing non-linear relationships and interactions between features, making it suitable for complex datasets [29,30]. Secondly, aggregating the predictions of multiple trees tends to yield more stable and reliable predictions, which are less susceptible to noise and outliers in the data [31]. Moreover, Random Forest inherently provides a measure of feature importance, allowing for the identification of the most influential variables in the prediction process [32].

3.3. Gradient Boosting

Gradient Boosting is a powerful ensemble learning technique that constructs a predictive model in a stage-wise fashion by sequentially optimizing the residuals of the previous models [33]. It is renowned for its ability to handle complex datasets and produce high-accuracy predictions across various domains. Unlike Random Forest, which builds independent trees in parallel, Gradient Boosting builds trees sequentially, with each subsequent tree aiming to correct the errors made by the previous ones. The prediction of a Gradient Boosting model can be expressed as follows [34]:

$$y_{i,\ predicted} = \sum_{t=1}^{T} \gamma_t h_t(x_i) \tag{18}$$

where $y_{i,\ predicted}$ represents the predicted value for the i-th data point, T is the total number of trees in the ensemble, γ_t denotes the learning rate or shrinkage parameter associated with the t-th tree, and $h_t(x_i)$ represents the prediction of the t-th tree for the i-th data point. At each iteration, Gradient Boosting fits a new decision tree to the negative gradient of the loss function with respect to the current model's predictions. This process effectively minimizes the residual errors of the previous model, leading to a gradual improvement in prediction accuracy. The final prediction is obtained by summing the predictions of all individual trees, weighted by the corresponding learning rates [35].

Gradient Boosting offers several advantages over other machine learning algorithms. It is capable of capturing intricate patterns and non-linear relationships within the data, making it suitable for complex regression tasks. Moreover, by iteratively minimizing the errors of preceding models, Gradient Boosting tends to produce highly accurate predictions, often outperforming other ensemble methods. However, Gradient Boosting is more sensitive to overfitting compared to Random Forest, necessitating careful tuning of hyperparameters such as the learning rate, tree depth, and regularization parameters. Additionally, the sequential nature of Gradient Boosting makes it less scalable for large datasets and computationally intensive tasks.

3.4. XGBoost

XGBoost, short for Extreme Gradient Boosting, is an optimized implementation of Gradient Boosting, renowned for its exceptional performance and scalability in predictive modelling tasks. It builds upon the principles of Gradient Boosting by introducing several algorithmic optimizations and parallelized computing techniques, making it one of the most widely used algorithms in machine learning competitions and real-world applications. The XGBoost model's prediction can be represented by the following equation:

$$y_{i,\ predicted} = \sum_{t=1}^{T} f_t(x_i) \tag{19}$$

where $y_{i,\ predicted}$ represents the predicted value for the i-th data point, T is the total number of trees in the ensemble, and $f_t(x_i)$ represents the prediction of the t-th tree for the i-th data point. Similar to Gradient Boosting, XGBoost sequentially fits decision trees to the negative gradient of the loss function with respect to the current model's predictions. XGBoost offers several advantages over traditional Gradient Boosting methods, including enhanced performance, scalability, and robustness to overfitting.

3.5. Support Vector Machine with Support Vector Regression (SVM-SVR)

Support Vector Machine (SVM) with Support Vector Regression (SVR) is a powerful machine learning technique renowned for its effectiveness in capturing complex relationships and handling non-linear regression tasks. SVM-SVR builds upon the principles of SVM for classification tasks, extending them to the realm of regression by formulating the problem as a function approximation task. Mathematically, the prediction of an SVM-SVR model can be represented as follows [36–38]:

$$y_{i,\ predicted} = \sum_{j=1}^{N} \alpha_j\, K\left(x_i, x_j\right) + b \tag{20}$$

where $y_{i,\ predicted}$ represents the predicted value for the i-th data point, N is the total number of support vectors, α_j denotes the Lagrange multipliers associated with the j-th support vector, $K\left(x_i, x_j\right)$ represents the kernel function, which computes the similarity between the i-th and j-th data points, and b is the bias term. The objective of SVM-SVR is to find the optimal hyperplane (or decision boundary) that maximizes the margin between data points while minimizing the error between the predicted and actual values. This is achieved by solving the following optimization problem [39]:

$$min_{w,b,\xi,\xi^*} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i{}^*) \tag{21}$$

Subject to:

$$y_i - w^T \phi(x_i) - b \le \epsilon + \xi_i \tag{22a}$$

$$w^T\phi(x_i) + b - y_i \leq \epsilon + \xi_i{}^* \tag{22b}$$

$$\xi_i,\ \xi_i{}^* \geq 0 \tag{22c}$$

where w represents the weight vector, b is the bias term, ξ_i and $\xi_i{}^*$ are slack variables that allow for deviations from the margin, ϵ is the margin of tolerance, and C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the error. SVM-SVR can capture non-linear relationships and high-dimensional interactions through kernel functions, allowing for greater flexibility in modelling complex datasets. Additionally, SVM-SVR inherently performs feature selection by focusing only on the support vectors, thereby reducing the computational complexity and memory requirements, especially for large-scale datasets.

3.6. Multilayer Perceptron Artificial Neural Network (MLP-ANN) Model

The multilayer perceptron artificial neural network (MLP-ANN) model is a versatile and powerful deep learning architecture widely used for regression and classification tasks. It consists of multiple layers of interconnected neurons, where each neuron receives input from the previous layer, applies a transformation function, and passes the result to the next layer [40]. MLP-ANN is capable of learning complex patterns and non-linear relationships within the data, making it well suited for modelling intricate datasets. The predicted output value using an MLP-ANN model can be formulated as follows:

$$y_{i,\ predicted} = f_{out}\left(W^{(2)} \cdot f_{hidden}\left(W^{(1)} + b^{(1)}\right) + b^{(2)}\right) \tag{23}$$

where $y_{i,\ predicted}$ represents the predicted value for the i-th data point, x_i denotes the input features for the i-th data point, f_{hidden} and f_{out} are activation functions applied to the hidden and output layers, respectively, $W^{(1)}$ and $W^{(2)}$ are weight matrices connecting the input to the hidden layer and the hidden to the output layer, respectively, and $b^{(1)}$ and $b^{(2)}$ are bias vectors for the hidden and output layers, respectively.

MLP-ANNs are trained using an optimization algorithm such as stochastic gradient descent (SGD) or its variants, which iteratively adjusts the weights and biases of the network to minimize a loss function. The loss function measures the discrepancy between the predicted and actual values, and the optimization algorithm seeks to find the optimal set of parameters that minimizes this discrepancy. MLP-ANNs offer several advantages, including their ability to learn complex patterns and relationships in the data, their flexibility in handling various types of data, and their scalability to large datasets. However, they require careful tuning of hyperparameters such as the number of hidden layers, the number of neurons per layer, and the choice of activation functions to achieve optimal performance.

4. Methodology

The methodology employed in this study aimed to comprehensively evaluate the predictive performance of various machine learning models in forecasting the uniaxial compressive strength (UCS) of rocks encountered in oil and gas wells. The analysis centred on five key input parameters: weight on bit (WOB), Sonic Transit Time (DT), Density Tool Reading (NPHI), rate of penetration (ROP), and Bulk Density (RHOB). Each of these features plays a significant role in determining the UCS of the rock. For instance, Bulk Density provides insight into the mineral composition and density of the rock, which directly correlates with its mechanical strength. Sonic Transit Time reflects the elasticity and acoustic properties of the rock, while Neutron Porosity indicates the porosity level, which affects fluid saturation and overall rock strength.

A correlation heatmap and scatter plots were generated (Figures 1 and 2) to visually depict the relationships between UCS and the input variables, facilitating a preliminary understanding of the dataset's characteristics.

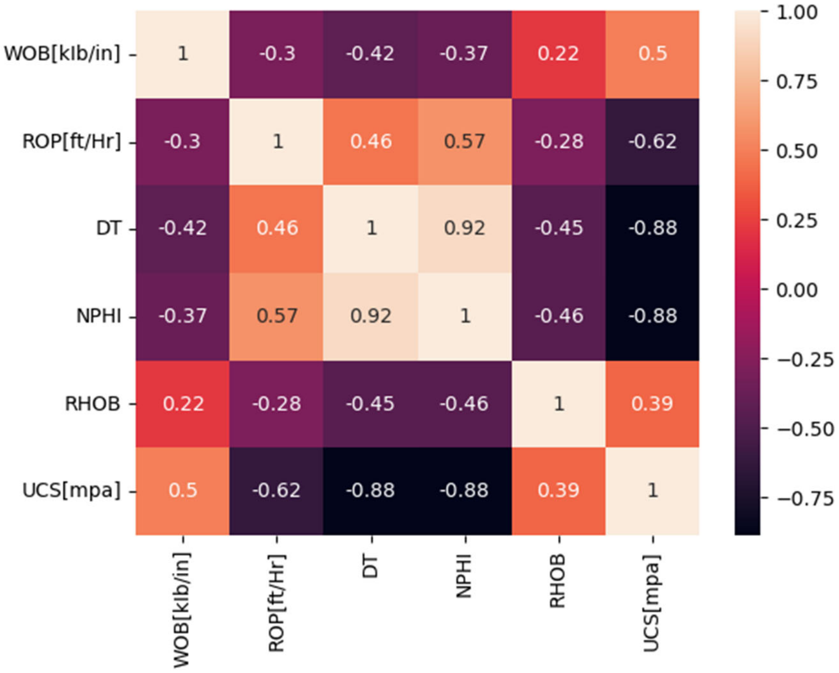


Figure 1. Correlation heatmap between UCS and the input parameters.

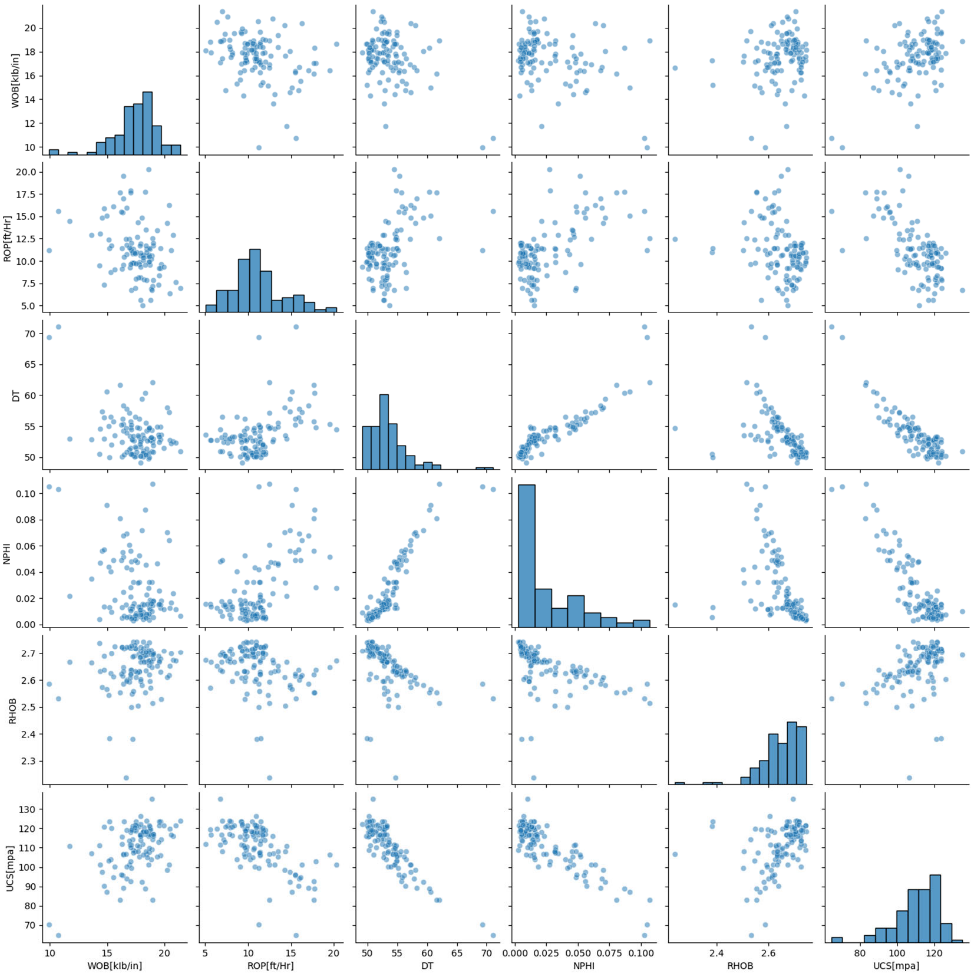


Figure 2. Scatter plots between the inputs and UCS.

The dataset, comprising 111 data points, was meticulously curated to ensure its representativeness and suitability for model training. To assess the model's performance accurately and prevent overfitting, we employed a standard 80:20 train–test split. During model development, we also conducted cross-validation with 6 folds to further validate the models' performance and ensure robustness. This approach allowed us to mitigate the risks of overfitting, ensuring that our findings are not only reliable but also generalizable to unseen datasets. While our current dataset comprises 111 data points, which may limit the generalizability of our conclusions, we acknowledge this limitation and plan to expand our dataset in future studies to further validate our results and improve model robustness.

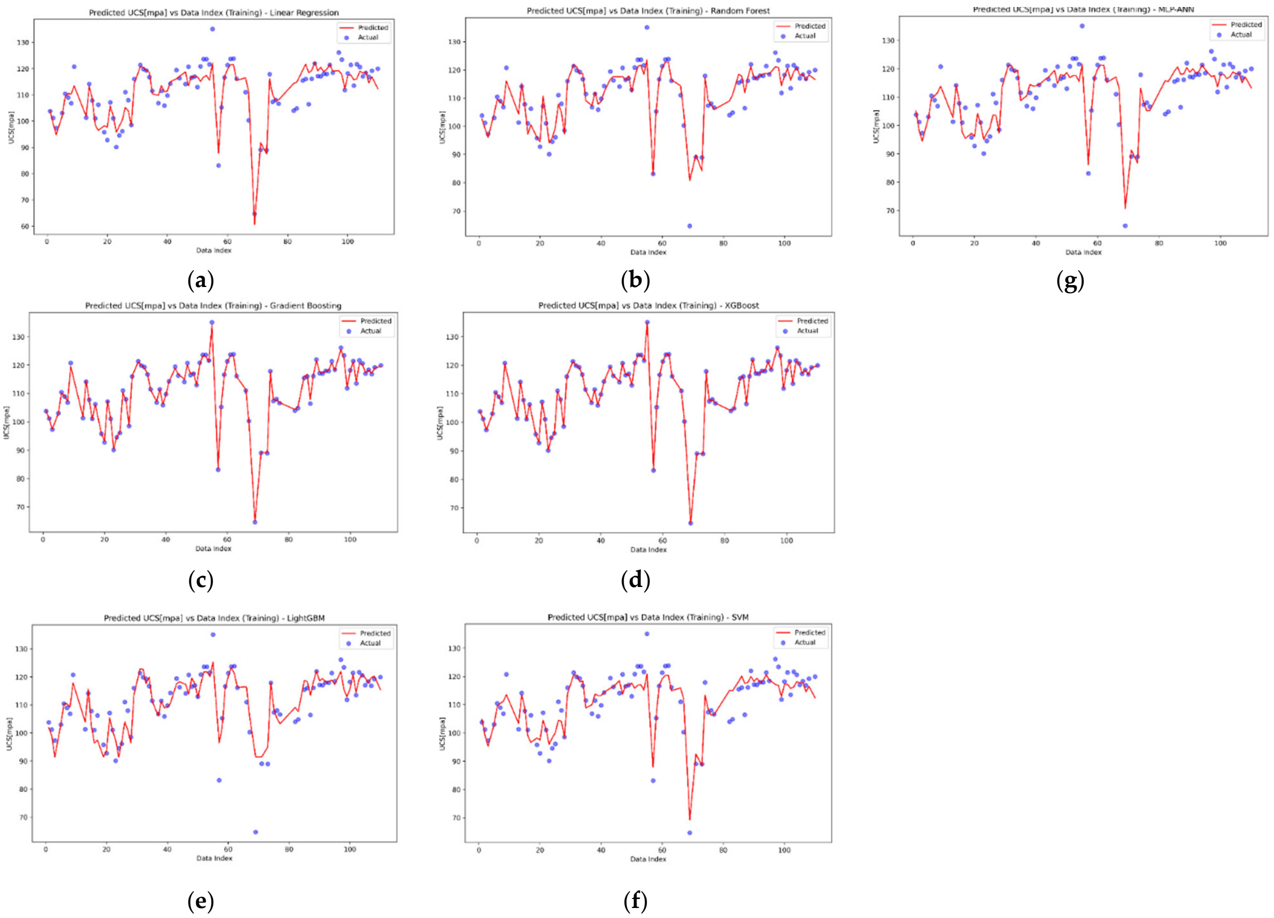
Each machine learning model, including Linear Regression, Random Forest, Gradient Boosting, XGBoost, SVM-SVR, and MLP-ANN, underwent a systematic training process. This process involved feeding the models with the training dataset and iteratively adjusting their parameters to minimize prediction errors and optimize predictive accuracy. For the Random Forest model, the best hyperparameters identified are a maximum depth of 10, a minimum number of 2 leaf samples, a minimum sample split of 5, and 100 estimators. In the case of Gradient Boosting, the optimal settings include a learning rate of 0.05, a maximum depth of 5, and 100 estimators, which facilitate effective learning without overfitting. XGBoost demonstrates improved performance with a learning rate of 0.15, a maximum depth of 5, and 100 estimators. The MLP-ANN model exhibits the best performance with a logistic activation function, a hidden layer configuration of 100 neurons, a constant learning rate, and stochastic gradient descent (SGD) as the solver. For the SVM, optimal hyperparameters comprise a regularization parameter C of 10, a gamma value set to "scale", and a linear kernel. Finally, LightGBM achieves its best predictive results with a learning rate of 0.15, a maximum depth of 5, and 100 estimators.

Following model training, a rigorous evaluation was conducted using diverse statistical metrics. These metrics included Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Percentage Error (MPE), Median Absolute Error, R^2 score, Adjusted R^2 score, Mean Squared Logarithmic Error (MSLE), Mean Bias Error (MBE), and geometric and symmetric Mean Absolute Percentage Error (MAPE). By analyzing these metrics, we were able to quantify the predictive accuracy, bias, and overall performance of each model, enabling a robust comparison and selection of the most suitable model for UCS prediction in oil and gas wells.

5. Results and Discussion

The visual representations in Figures 3 and 4 offer a comprehensive overview of the comparative performance of various machine learning models in predicting UCS values for both the training and testing datasets. These figures serve as valuable tools for assessing the predictive accuracy and generalization capabilities of different models in the context of geotechnical engineering applications.

Figure 3 presents a comparative analysis of predicted and actual UCS values against the data index for a training dataset using different machine learning models. In the Linear Regression model (Figure 3a), the predicted UCS values (red line) demonstrate a general alignment with the actual values (blue dots), but significant deviations are evident, particularly in regions where the actual values exhibit sharp fluctuations. These deviations suggest that the Linear Regression model struggles to capture the non-linear patterns in the data accurately. Random Forest (see Figure 3b) and Gradient Boosting (see Figure 3c) exhibit closer alignment with the actual UCS values compared to Linear Regression. Both models reduce the magnitude of the deviations, reflecting their capability to handle non-linearities better than a simple linear model. Among these, Gradient Boosting appears to have a slight edge, maintaining a tighter fit throughout the dataset. XGBoost (see Figure 3d) further improves the accuracy, displaying a strong correlation between the predicted and actual values across the entire data index. The model's robustness in managing diverse data patterns is evident from the minimal deviations observed. This is reflected in the stability of its predictions, even when faced with variations in the underlying data characteristics, such as different geological formations or variations in drilling parameters. The minimal deviations observed in the model's predictions indicate that it can effectively generalize from the training data to unseen samples, thereby minimizing the risk of overfitting. This robustness ensures that the model remains reliable under various operational conditions, making it a valuable tool for practical applications in the field. LightGBM (see Figure 3e) demonstrates a performance comparable to XGBoost, with a similarly tight fit and minimal errors. Both boosting algorithms, XGBoost and LightGBM, seem to offer superior predictive accuracy due to their advanced ensemble learning techniques. The SVM-SVR model (see Figure 3f) shows a reasonable fit but with more pronounced deviations in certain sections of the data index. This suggests that while SVM-SVR is effective, it might not be as versatile as the boosting methods in capturing the full complexity of the UCS values. Finally, the MLP-ANN model (see Figure 3g) provides a fit comparable to the boosting models, with predictions closely following the actual values. The neural network's ability to model intricate patterns in the data contributes to its high predictive performance.



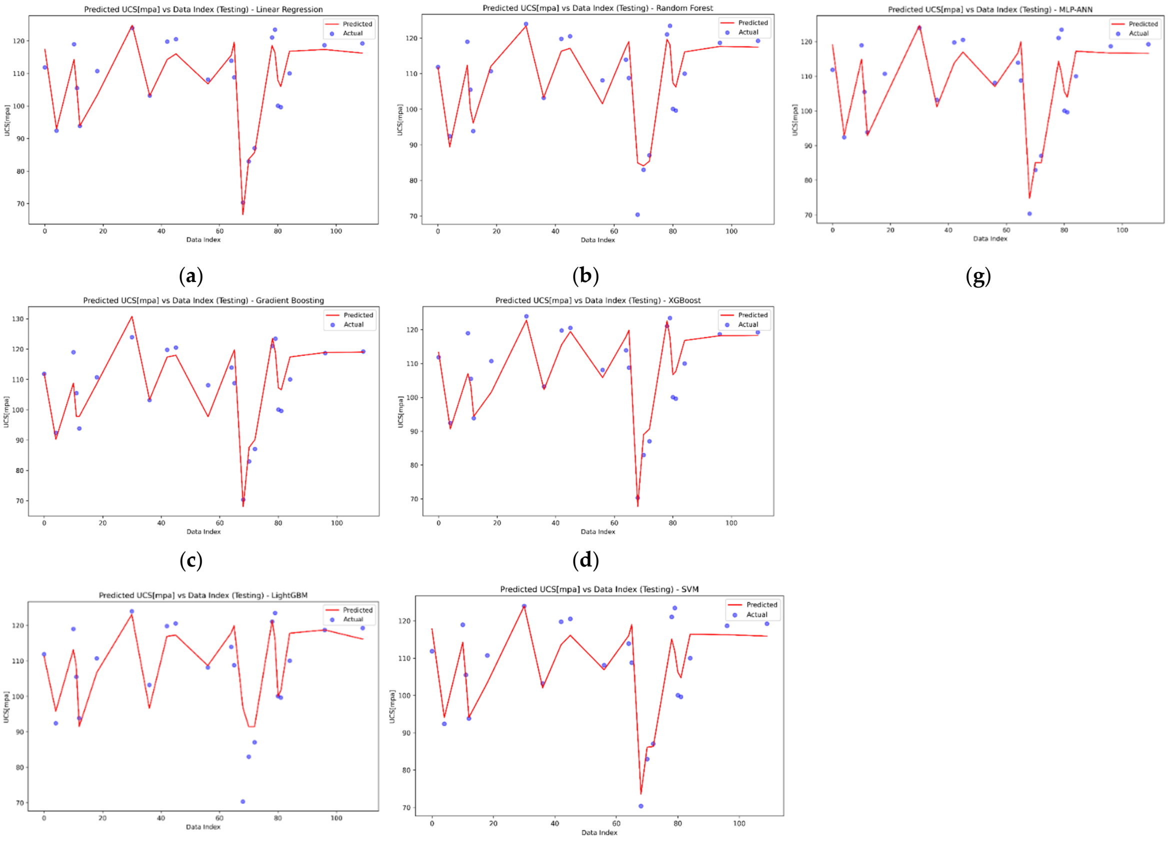
(a) (b) (g)

(c) (d)

(e) (f)

Figure 3. Comparison between the predicted values of each model and the actual UCS values versus the data index for the training dataset. (**a**) Linear Regression, (**b**) Random Forest, (**c**) Gradient Boosting, (**d**) XGBoost, (**e**) LightGBM, (**f**) SVM-SVR, and (**g**) MLP-ANN.

Figure 4 provides a comparative analysis of several predictive models in estimating UCS against actual values using a testing dataset. Linear Regression (Figure 4a) reveals substantial discrepancies between predicted and actual UCS values, which are particularly notable in several spikes and troughs where the model fails to capture the variability in the data. This reinforces the earlier observation that Linear Regression is less effective in modelling complex, non-linear relationships inherent in the dataset. As depicted in Figure 4b, Random Forest shows improved performance compared to Linear Regression, with predictions more closely aligned with actual values. However, some deviations persist, indicating that while Random Forest models handle non-linearities better, they may still miss certain data intricacies. Gradient Boosting (see Figure 4c) demonstrates a closer fit to the actual UCS values, reducing the magnitude of prediction errors compared to both Linear Regression and Random Forest. The model's ensemble learning capability enhances its predictive accuracy, although minor deviations are still present. As demonstrated in Figure 4d, XGBoost maintains a robust alignment with actual values across the testing dataset, further validating its efficacy in handling complex data patterns. The minimal deviations observed suggest that XGBoost effectively generalizes the underlying data structure. LightGBM (see Figure 4e) displays performance on par with XGBoost, with predictions closely following the actual UCS values. The model's ability to capture detailed data patterns is evident, though occasional deviations indicate slight overfitting or data-specific challenges. SVM-SVR (see Figure 4f) exhibits reasonable predictive accuracy but with noticeable deviations in several regions of the data index. This suggests that while SVM-SVR is effective in certain scenarios, it may not consistently capture the full complexity of the UCS values as effectively as ensemble methods. As depicted in Figure 4g, MLP-ANN shows a strong predictive performance, with predictions aligning closely with actual values throughout the dataset. The neural network's capability to model complex and non-linear relationships contributes to its high accuracy, although minor deviations suggest room for further optimization. The comparative analysis for both training and testing phases underscores that ensemble methods such as Gradient Boosting, XGBoost, and LightGBM, along with neural network approaches like MLP-ANN, generally outperform simpler models like Linear Regression and SVM-SVR in predicting UCS values. These advanced models demonstrate superior generalization capabilities, making them more reliable for practical applications in predicting complex, real-world phenomena.



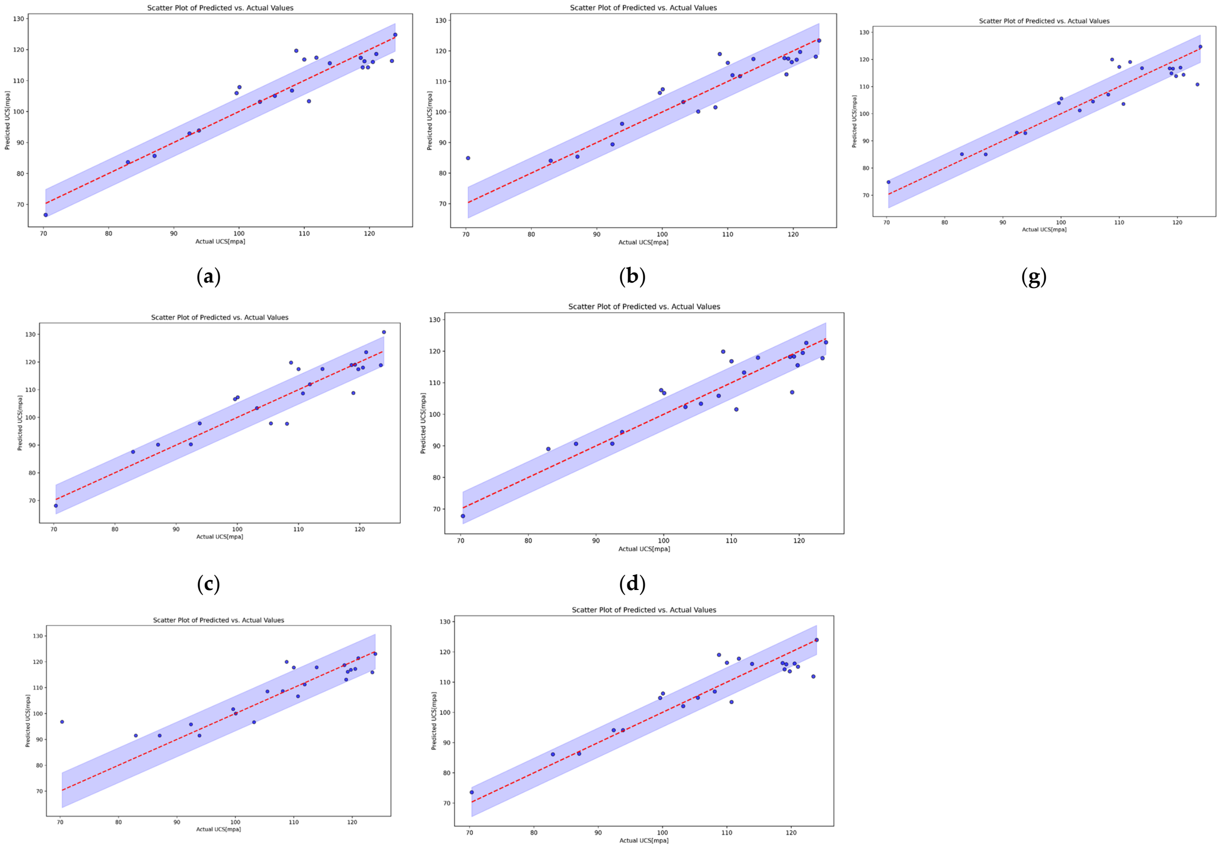
(a) (b) (g)

(c) (d)

(e) (f)

Figure 4. Comparison between the predicted values of each model and the actual UCS values versus the data index for the testing dataset. (**a**) Linear Regression, (**b**) Random Forest, (**c**) Gradient Boosting, (**d**) XGBoost, (**e**) LightGBM, (**f**) SVM-SVR, and (**g**) MLP-ANN.

Figure 5 provides a comparative analysis of the scatter plots illustrating the relationship between predicted and actual UCS values using different modelling techniques. Each subplot displays a series of points representing the actual UCS values on the x-axis and the corresponding predicted UCS values on the y-axis, along with a red dashed line indicating the ideal 1:1 prediction line and a shaded region representing prediction uncertainty. As shown in Figure 5a, the Linear Regression model shows a broad distribution of points around the ideal line, indicating a moderate fit with noticeable variance, particularly for higher UCS values. As depicted in Figure 5b, the Random Forest model demonstrates an improved alignment with the 1:1 line, suggesting better prediction accuracy and less dispersion compared to Linear Regression. As illustrated in Figure 5c,d, the Gradient Boosting and XGBoost (d) models exhibit a closer clustering of points around the ideal prediction line, signifying higher predictive precision and reduced variability. This observation highlights the effectiveness of ensemble techniques in capturing the underlying patterns in the data. Similarly, as shown in Figure 5e, LightGBM also shows a strong correlation between predicted and actual values, though with slightly more dispersion compared to Gradient Boosting and XGBoost. Furthermore, The SVM-SVR model (see Figure 5f) presents a robust performance, with most points lying near the ideal line and within the uncertainty bounds. However, there are a few outliers that deviate significantly, indicating some limitations in the model's generalization capability. Lastly, the MLP-ANN model (see Figure 5g) demonstrates a satisfactory predictive performance with a majority of points closely following the ideal line. Nonetheless, there is a slight tendency for higher variability at the extremes of the UCS range, suggesting that while MLP-ANN captures the overall trend effectively, it may struggle with extreme values.



(**a**) (**b**) (**g**)

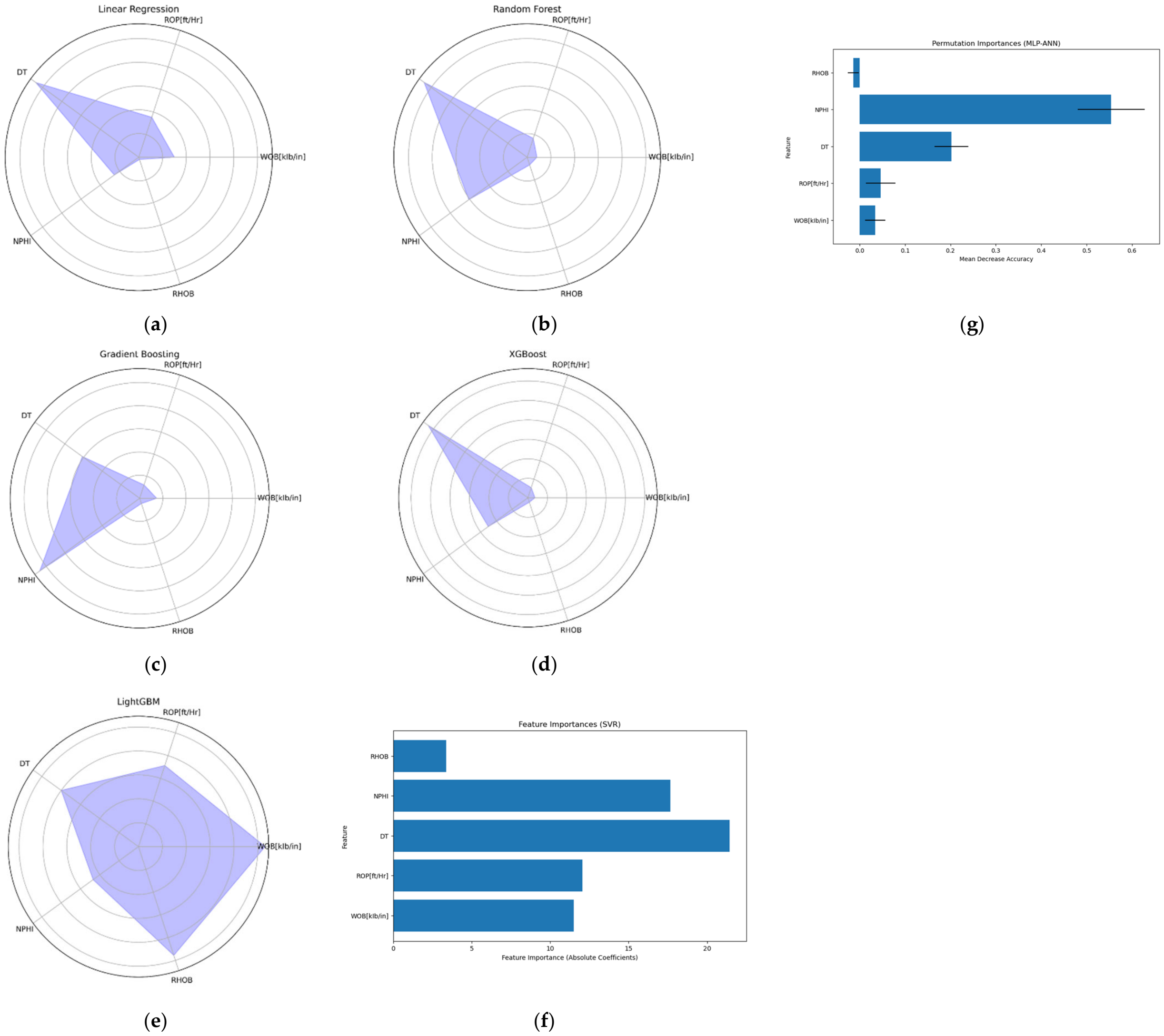
(**c**) (**d**)

(**e**) (**f**)

Figure 5. Comparison between the scatter plot of each model output versus the actual UCS values. (**a**) Linear Regression, (**b**) Random Forest, (**c**) Gradient Boosting, (**d**) XGBoost, (**e**) LightGBM, (**f**) SVM-SVR, and (**g**) MLP-ANN.

Figure 6 illustrates the relative importance of input features for various machine learning models. Understanding feature importance is essential for interpreting model behaviour and identifying key drivers of predictions. Figure 6a highlights the feature RHOB as the most influential predictor, followed by NPHI and DT in the Linear Regression model. The radar plot indicates a significant reliance on RHOB, with other features playing relatively minor roles. This suggests that in the linear model, RHOB holds a dominant explanatory power, likely due to its strong linear relationship with the target variable. Figure 6b shows a similar trend, with RHOB again being the most important feature for the Random Forest model. However, the spread of importance is slightly more balanced, with DT and NPHI also contributing significantly. This indicates that the Random Forest model captures more complex interactions among features compared to Linear Regression. Figure 6c,d both display a notable emphasis on RHOB, but with a more pronounced role for NPHI and DT for the Gradient Boosting and XGBoost models. The radar plots for these models reveal a more distributed importance among the features, suggesting that the boosting methods are effective in leveraging multiple features to enhance predictive accuracy. Figure 6e shows a more balanced distribution of feature importance, with WOB, RHOB, and NPHI all contributing significantly to the LightGBM model. This model's radar plot is more uniform compared to others, indicating that LightGBM utilizes a diverse set of features to make predictions, potentially leading to better generalization. Figure 6f provides a bar chart of feature importances based on absolute coefficients. Here, NPHI emerges as the most influential feature, followed by DT and RHOB for the SVM-SVR model. This distribution reflects the model's ability to capture complex, non-linear relationships where multiple features significantly impact the outcome. Figure 6g uses permutation importance to measure feature relevance. NPHI and DT show the highest importance, indicating their critical role in the neural network's predictions. The reliance on these features suggests that the MLP-ANN model effectively captures intricate patterns in the data.

Figure 7 illustrates the Taylor plots for seven different models. As shown in Figure 7a, in the Linear Regression model, the correlation coefficient is moderate, suggesting a reasonable but not exceptional agreement between predicted and observed UCS values. The model's standard deviation is lower than that of the observations, indicating that Linear Regression underestimates the variability in UCS values. The Random Forest model (see Figure 7b) shows a higher correlation coefficient compared to Linear Regression, indicating a stronger relationship between predictions and actual values. The standard deviation is closer to that of the observations, suggesting that Random Forest better captures the variability in UCS values. Figure 7c demonstrates an even higher correlation coefficient, nearing 1.0, for the Gradient Boosting model, which implies a very strong agreement between the model predictions and the actual UCS values. The standard deviation of the predictions aligns closely with that of the observations, indicating that Gradient Boosting effectively captures the variability in the data. As shown in Figure 7d, the XGBoost model also exhibits a high correlation coefficient, similar to Gradient Boosting, and a standard deviation that closely matches the observed values. This indicates that XGBoost is highly effective in predicting UCS values with a high degree of accuracy and reliability. As depicted in Figure 7e, LightGBM shows a strong correlation coefficient, slightly less than that of Gradient Boosting and XGBoost, but still indicative of a good predictive performance. The standard deviation is close to the observed values, although there is a slight deviation, suggesting some minor discrepancies in capturing the full range of data variability. Figure 7f presents a good correlation coefficient for the SVM-SVR model, although not as high as the ensemble methods like Gradient Boosting and XGBoost. The standard deviation is comparable to the observed values, indicating that SVM-SVR performs well in terms of capturing data variability, albeit with occasional prediction inaccuracies. Figure 7g shows a strong correlation coefficient and a standard deviation that aligns well with the observed values, indicating that MLP-ANN captures both the trend and variability in the UCS data effectively. However, similar to SVM-SVR, it might have occasional outliers or prediction errors.



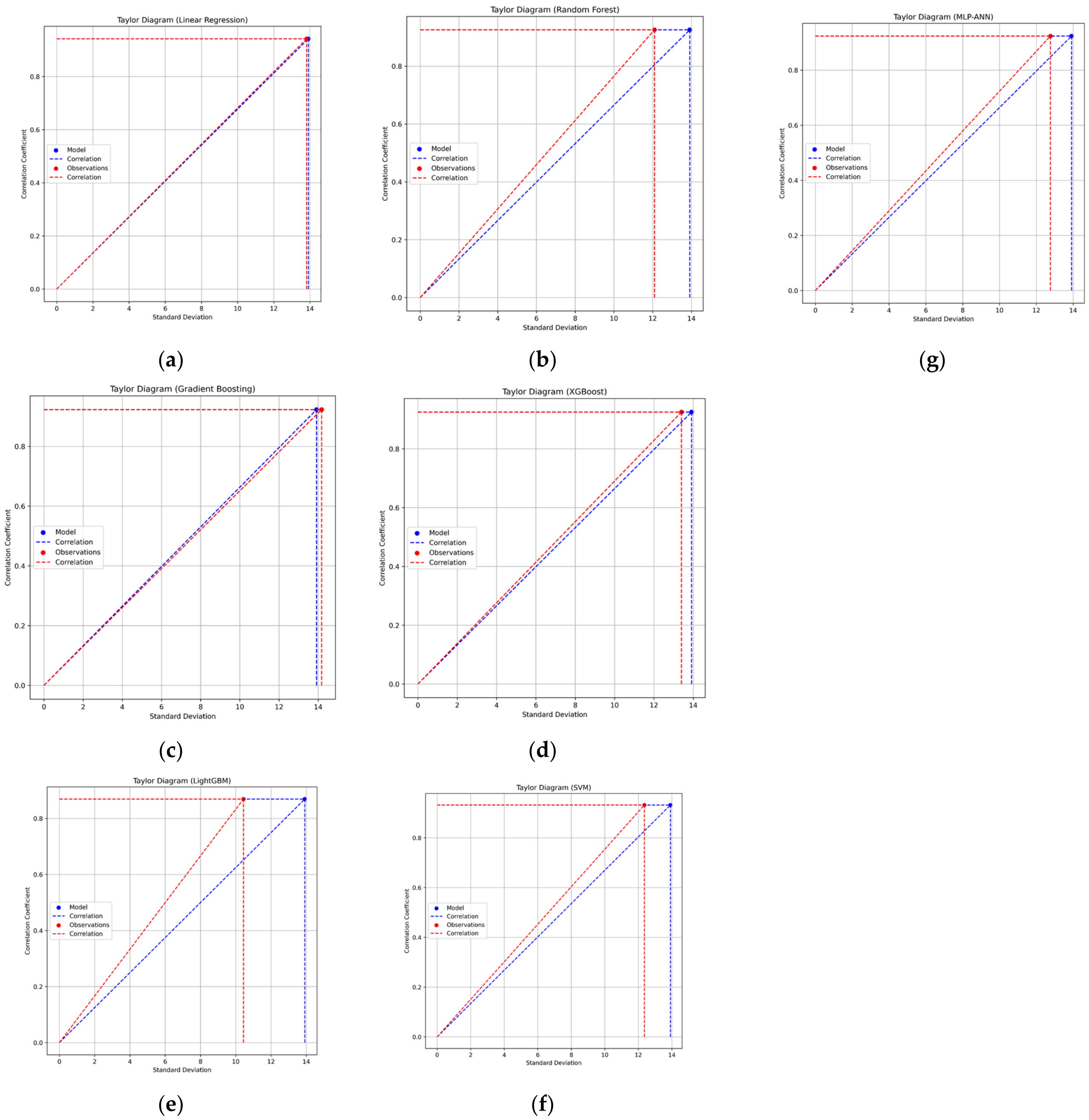
(**a**) (**b**) (**g**)

(**c**) (**d**)

(**e**) (**f**)

Figure 6. Comparison between the relative importance of the inputs for each model. (**a**) Linear Regression, (**b**) Random Forest, (**c**) Gradient Boosting, (**d**) XGBoost, (**e**) LightGBM, (**f**) SVM-SVR, and (**g**) MLP-ANN.

The analysis of residual distributions for various machine learning models is critical for understanding their predictive performance. Residuals, which are the differences between observed and predicted values, should ideally exhibit a random pattern centred around zero; any discernible trends or patterns may suggest that the model is not effectively capturing underlying relationships within the data. This analysis is instrumental in identifying biases in the model, revealing whether it tends to consistently overestimate or underestimate predictions. Moreover, it helps to detect non-linearities that may require additional features or interaction terms for better representation. Assessing the residuals also aids in evaluating the homogeneity of variance; the presence of heteroscedasticity can compromise the reliability of the model's predictions. Figure 8 presents the residual distributions of seven different models. As shown in Figure 8a, Linear Regression demonstrates a wide spread of residuals, with several outliers on both ends. The distribution appears slightly skewed to the left, indicating that the model tends to underpredict in some instances. The presence of multiple residual peaks suggests that the model might not fully capture the underlying data patterns, leading to heterogeneous residuals. As depicted in Figure 8b, the Random Forest technique shows a more centred residual distribution, although it still exhibits some degree of skewness to the left. The residuals are more tightly clustered around the mean compared to Linear Regression, suggesting better overall predictive accuracy. However, the model still struggles with extreme values, as evidenced by the residuals extending far from the mean.



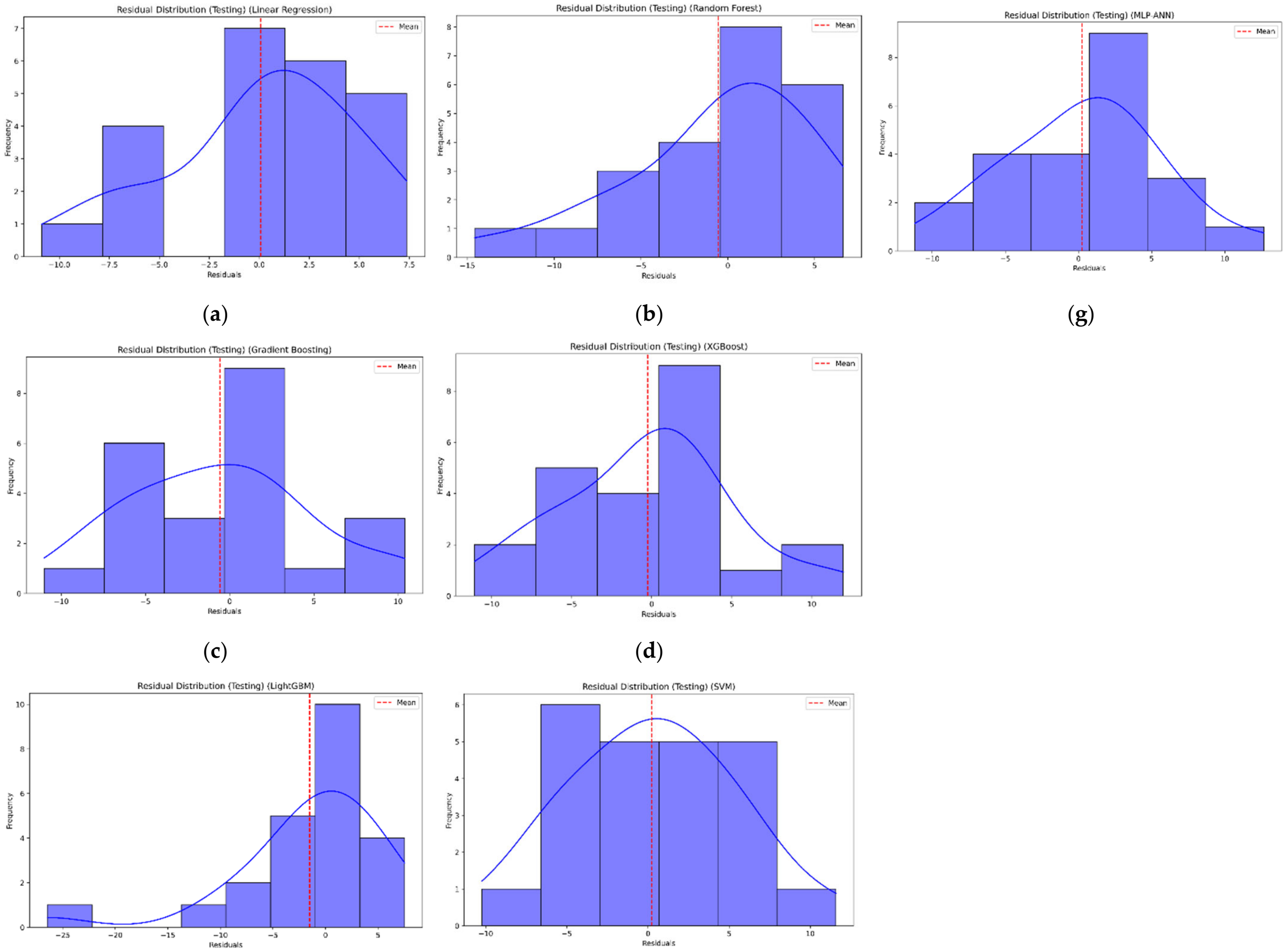
(**a**) (**b**) (**g**)

(**c**) (**d**)

(**e**) (**f**)

Figure 7. Comparison between the Taylor plot for each model. (**a**) Linear Regression, (**b**) Random Forest, (**c**) Gradient Boosting, (**d**) XGBoost, (**e**) LightGBM, (**f**) SVM-SVR, and (**g**) MLP-ANN.

Figure 8c,d present similar residual distributions for the Gradient Boosting and XGBoost models, respectively, with both showing a noticeable concentration around the mean and a reduction in extreme residuals. This indicates that these models are effective in minimizing prediction errors and handling variance within the data. Both distributions, however, show slight left skewness, suggesting occasional underpredictions. As demonstrated in Figure 8e, the LightGBM model displays a distinctive pattern with a significant peak around a small positive residual value. This indicates a slight bias in the model's predictions, consistently overestimating to a small extent. Despite this, the distribution is relatively narrow, suggesting high accuracy in most predictions. Figure 8f,g show residuals for the SVM-SVR and MLP-ANN models, respectively, with wider spreads compared to the boosting methods but narrower than Linear Regression and Random Forest. The SVM-SVR residuals are fairly symmetrical around the mean, implying balanced prediction errors, while MLP-ANN shows a right-skewed distribution, indicating a tendency towards overprediction.



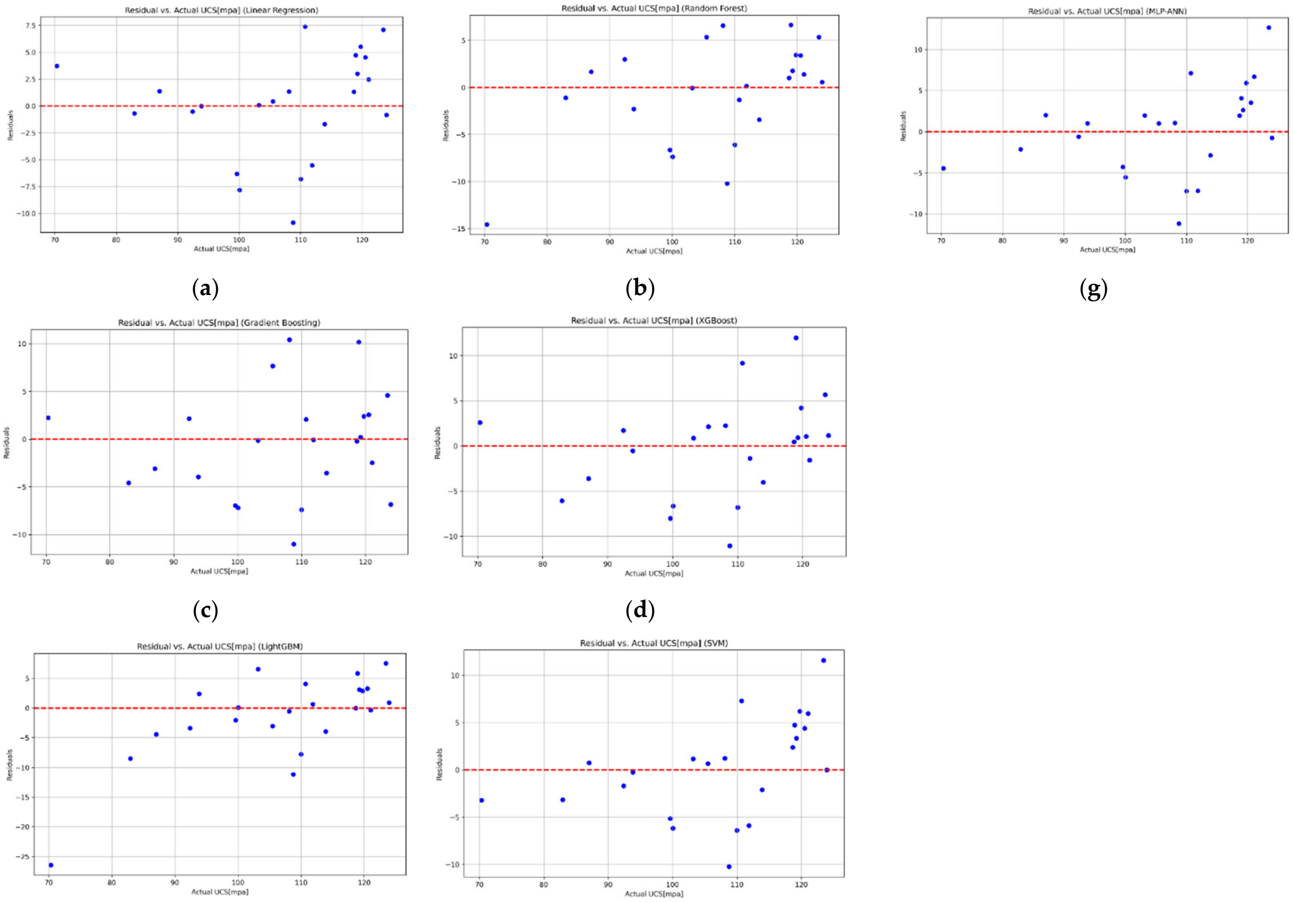
(a) (b) (g)

(c) (d)

(e) (f)

Figure 8. Comparison between the residual distribution of each model. (**a**) Linear Regression, (**b**) Random Forest, (**c**) Gradient Boosting, (**d**) XGBoost, (**e**) LightGBM, (**f**) SVM-SVR, and (**g**) MLP-ANN.

Figure 9 illustrates the residuals of predicted versus actual UCS values for seven different models. Residuals are the differences between observed values and the values predicted by the models, and analyzing these residuals helps evaluate model performance by identifying any patterns or biases in the predictions. As demonstrated in Figure 9a, in the Linear Regression model, the residuals exhibit a noticeable spread around the zero line, with a tendency to increase as the actual UCS values increase. This pattern suggests that the model may be underpredicting for higher UCS values and overpredicting for lower UCS values, indicating a potential linear bias in the predictions. Figure 9b shows residuals that are more tightly clustered around the zero line compared to Linear Regression, although there are still some noticeable outliers. The residuals do not display a clear pattern, indicating that Random Forest provides a more balanced prediction across the range of UCS values but still has room for improvement in reducing prediction errors. Figure 9c presents residuals that are fairly well distributed around the zero line, with fewer outliers than both Linear Regression and Random Forest. This indicates that Gradient Boosting has a strong predictive capability and effectively minimizes bias, providing accurate predictions across the range of UCS values. As illustrated in Figure 9d, XGBoost also exhibits a well-distributed pattern of residuals around the zero line, similar to Gradient Boosting. The absence of a clear trend or bias in the residuals further confirms XGBoost's robustness and accuracy in predicting UCS values. As depicted in Figure 9e, the LightGBM model displays residuals that are somewhat more scattered, with a few noticeable outliers, particularly at the higher UCS values. While LightGBM generally performs well, these outliers suggest occasional overprediction or underprediction, indicating variability in the model's accuracy. Figure 9f shows a relatively balanced distribution of residuals around the zero line, although there are several instances of significant positive and negative residuals. This suggests that while SVM-SVR can predict UCS values with reasonable accuracy, it may struggle with certain data points, leading to occasional large errors. Figure 9g reveals residuals that are spread more widely around the zero line, with several outliers, especially at the lower end of the UCS range. This dispersion indicates that MLP-ANN has difficulty maintaining consistent prediction accuracy across the range of UCS values, resulting in higher variability in its predictions. The residual analysis in Figure 9 highlights that ensemble methods such as Gradient Boosting and XGBoost provide the most accurate and unbiased predictions, with residuals closely clustered around the zero line and minimal outliers. Random Forest and LightGBM also perform well but exhibit slightly more variability. Linear Regression and MLP-ANN show higher dispersion and noticeable patterns in residuals, indicating potential biases and less reliable predictions. SVM-SVR offers reasonable accuracy but with occasional large residuals.



(**a**) (**b**) (**g**)

(**c**) (**d**)

(**e**) (**f**)

Figure 9. Comparison between the residuals of each model versus the actual UCS values. (**a**) Linear Regression, (**b**) Random Forest, (**c**) Gradient Boosting, (**d**) XGBoost, (**e**) LightGBM, (**f**) SVM-SVR, and (**g**) MLP-ANN.

The comparative analysis reveals that ensemble methods, particularly Gradient Boosting and XGBoost, deliver superior predictive accuracy and reliability, minimizing residuals and reducing extreme prediction errors for UCS prediction. Random Forest and LightGBM also perform well, albeit with slightly more variance. Linear Regression and MLP-ANN show moderate predictive capabilities with higher variability and wider residual distributions, indicating less precise predictions, and SVM-SVR, while generally accurate, shows better error handling than Linear Regression but not as refined as the boosting methods and is prone to occasional significant errors.

Figure 10 depicts the relationship between actual UCS data and the predictions made by the Golubev and Rabinovitch [9] model. The red triangles represent the data points, which generally align with the blue linear fit line, indicating a strong positive correlation. The R^2 value of 0.7761 suggests a moderate fit, implying that the model explains approximately 77.61% of the variance in the actual UCS data. This degree of correlation indicates that the Golubev and Rabinovitch [9] model is capable of predicting UCS values with average accuracy. However, the spread of data points around the fit line also suggests the presence of some deviations and potential outliers, which may be due to various factors such as heterogeneity or model limitations.

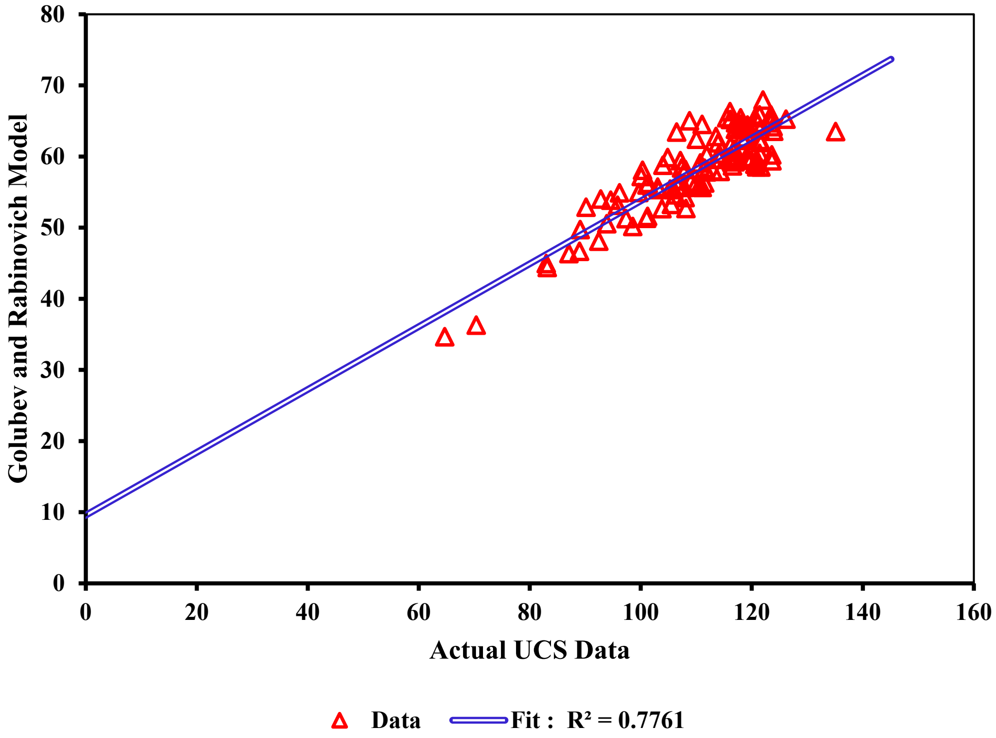


Figure 10. Scatter plot of estimated and measured UCS via the Golubev and Rabinovich model [9].

Figure 11 illustrates the comparison between real UCS data and the estimates produced by the Rzhevsky and Novick [41] model. The coefficient of determination, R^2, is 0.7656, signifying that the model accounts for approximately 76.56% of the variability in the UCS data. This R^2 value suggests that the Rzhevsky and Novick [41] model is a moderate predictor of UCS. Nonetheless, the dispersion of data points around the regression line points to some discrepancies and outliers, potentially arising from inherent model limitations.

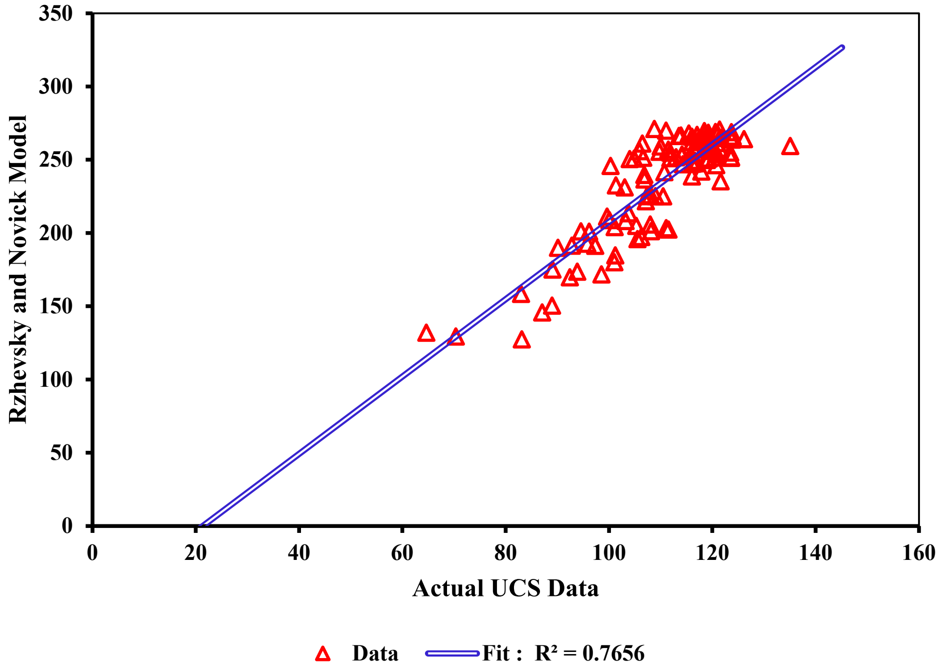


Figure 11. Scatter plot of estimated and measured UCS via the Rzhevsky and Novick model [41].

Figure 12 presents a comparative analysis of the measured UCS data against predictions made by Nabaei et al.'s [10] model. The R^2 value of 0.7674 suggests a moderately positive correlation between the model's predictions and the actual UCS values. However, some scatter around the line indicates that while the model captures the general trend of the data, there are discrepancies and potential outliers that could be attributed to variances in measurement conditions or inherent limitations of the model. The alignment of a majority of the data points along the line of best fit implies that Nabaei et al.'s [10] model can moderately estimate UCS values within a specific range, although the spread of the data suggests that further refinement of the model could enhance its predictive accuracy.

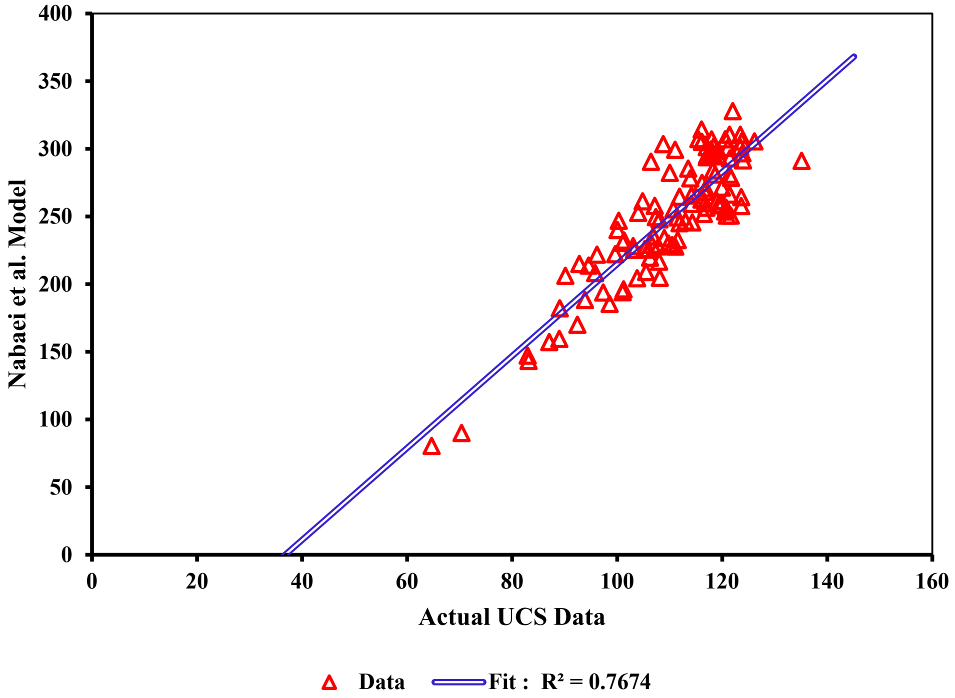


Figure 12. Scatter plot of estimated and measured UCS via Nabaei et al.'s [10] model.

Table 3 compares the outputs of different machine learning models in terms of statistical indicators such R^2, RMSE, MAE, MAE, and MPE. Linear Regression exhibited the lowest RMSE of 4.72 and MSE of 22.27, indicating its effectiveness in minimizing prediction errors. Additionally, it showed the lowest MAPE (3.38) and MPE (−0.0002), implying minimal percentage error. XGBoost and MLP-ANN also demonstrated competitive performance in terms of MAPE and MPE. In terms of MAE, XGBoost achieved the lowest value (2.578), followed closely by Linear Regression (2.993) and SVM-SVR (3.337). Linear Regression exhibited the highest R^2 value of 0.8849, indicating its ability to explain approximately 88.49% of the variance in the UCS values. SVM-SVR also demonstrated a high R^2 (0.8660), followed by XGBoost (0.8542). Adjusted R^2 values were consistent with R^2 values, with Linear Regression exhibiting the highest adjusted R^2 of 0.8511. Linear Regression and SVM-SVR exhibited the lowest MSLE values, indicating their effectiveness in minimizing logarithmic prediction errors. However, LightGBM showed the highest MSLE, suggesting higher variability in the accuracy of predictions. Regarding MBE, Linear Regression had a positive bias (0.0742), indicating slight overestimation, while other models exhibited varying degrees of bias. Overall, Linear Regression emerged as the top-performing model across multiple evaluation metrics, showcasing its robustness and effectiveness in predicting UCS values in oil and gas wells. However, XGBoost and SVM-SVR also demonstrated competitive performance, highlighting the importance of considering multiple models in predictive modelling tasks.

Table 3. Comparison between various statistical performance indicators for developed models.

Indicator	Linear Regression	Random Forest	Gradient Boosting	XGBoost	LightGBM	SVM-SVR	MLP-ANN
RMSE	4.72	5.35	5.54	5.31	7.24	5.09	5.33
MSE	22.27	28.61	30.71	28.23	52.35	25.94	28.43
MAPE	3.38	4.07	4.18	3.87	5.04	3.76	3.93
MPE	−0.0002	−0.9450	−0.6376	−0.4052	−2.1977	−0.0894	−0.0527
MAE	2.993	3.404	3.563	2.578	3.277	3.337	3.531
Geometric MAPE	3.36	3.97	4.17	3.85	4.74	3.75	3.92
Symmetric MAPE	3.364	3.971	4.170	3.855	4.742	3.755	3.925
R^2	0.8849	0.8522	0.8414	0.8542	0.7296	0.8660	0.8531
Adjusted R^2	0.8511	0.8087	0.7947	0.8113	0.6500	0.8266	0.8100
MSLE	0.0018	0.0030	0.0026	0.0024	0.0062	0.0020	0.0022
MBE	0.0742	−0.5570	−0.5725	−0.2447	−1.4988	0.2286	0.2345

6. Conclusions

In this study, we explored the efficacy of various machine learning models in predicting the UCS of rocks in oil and gas wells. Through rigorous experimentation and analysis, we evaluated the performance of five distinct models: Linear Regression, Random Forest, Gradient Boosting, XGBoost, SVM-SVR, and MLP-ANN. Our investigation aimed to identify the most accurate and reliable model for UCS prediction, which is crucial for optimizing drilling operations and ensuring wellbore stability in the petroleum industry.

While RHOB consistently appears as a significant feature across most models, the importance of other features such as NPHI and DT varies depending on the model used. Ensemble methods like Gradient Boosting, XGBoost, and LightGBM demonstrate a more balanced utilization of features, enhancing their predictive performance. In contrast, Linear Regression relies heavily on RHOB, reflecting its simplicity and limitations in capturing complex relationships. SVM-SVR and MLP-ANN highlight the importance of NPHI and DT, indicating their effectiveness in modelling non-linear interactions.

Our findings underscore the superiority of ensemble methods, particularly Gradient Boosting and XGBoost, in accurately predicting UCS values. These models demonstrate robustness, reliability, and superior generalization capabilities, making them ideal choices for practical applications in geotechnical engineering.

Additionally, our results highlight the effectiveness of MLP-ANN in capturing the complex, non-linear relationships inherent in the UCS dataset. While MLP-ANN exhibits strong predictive performance, it occasionally struggles with extreme values, indicating opportunities for further optimization.

Furthermore, Random Forest and LightGBM also exhibit commendable performance, albeit with slightly more variability compared to ensemble methods. These models provide viable alternatives, especially in scenarios where computational efficiency is a concern.

On the other hand, Linear Regression and SVM-SVR models, while providing moderate predictive capabilities, fall short of capturing the full complexity of the UCS dataset. These simpler models are outperformed by ensemble methods and MLP-ANN in terms of predictive accuracy and reliability.

Our study underscores the importance of employing advanced machine learning techniques for UCS prediction in oil and gas wells. By leveraging these methodologies, the petroleum industry can benefit from enhanced decision-making processes that lead to improved drilling efficiency and safety. The adoption of superior predictive models not only optimizes operational parameters but also contributes to sustainable practices by minimizing the risks associated with drilling operations. Ultimately, our research paves the way for further exploration and application of machine learning in geotechnical contexts, highlighting the significant potential for ongoing improvements in the field.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Peška, P.; Zoback, M.D. Compressive and tensile failure of inclined well bores and determination of in situ stress and rock strength. *J. Geophys. Res. Solid Earth* **1995**, *100*, 12791–12811. [CrossRef]
2. Yurdakul, M.; Ceylan, H.; Akdas, H. A predictive model for uniaxial compressive strength of carbonate rocks from Schmidt hardness. In Proceedings of the 45th U.S. Rock Mechanics/Geomechanics Symposium, San Francisco, CA, USA, 26–29 June 2011.
3. Raaen, A.; Hovem, K.; Joranson, H.; Fjaer, E. FORMEL: A step forward in strength logging. In Proceedings of the SPE Annual Technical Conference and Exhibition, Denver, CO, USA, 6–9 October 1996.
4. Nabaei, M.; Shahbazi, K.; Shadravan, A. Uncertainty analysis in unconfined rock compressive strength prediction. In Proceedings of the SPE Deep Gas Conference and Exhibition, Manama, Bahrain, 24–26 January 2010.
5. Petunin, V.V.; Yin, X.; Tutuncu, A.N. Porosity and permeability changes in sandstones and carbonates under stress and their correlation to rock texture. In Proceedings of the SPE Canada Unconventional Resources Conference, Calgary, AB, Canada, 15–17 November 2011.
6. Chang, C.; Zoback, M.D.; Khaksar, A. Empirical relations between rock strength and physical properties in sedimentary rocks. *J. Pet. Sci. Eng.* **2006**, *51*, 223–237. [CrossRef]
7. Militzer, H.; Stoll, R. *Einige Beiträge der Geophysik zur Primärdatenerfassung im Bergbau*; Neue Bergbautechnik: Leipzig, Germany, 1973.
8. Rzhevskiĭ, V.; Novik, G. *The Physics of Rocks*; Mir Publichers: Moscow, Russia, 1971.
9. Golubev, A.; Rabinovich, G. Resultaty primeneia appartury akusticeskogo karotasa dlja predeleina proconstych svoistv gornych porod na mestorosdeniaach tverdych isjopaemych. *Prikl. Geofiz. Mosk.* **1976**, *73*, 109–116.
10. Nabaei, M.; Shahbazi, K. A new approach for predrilling the unconfined rock compressive strength prediction. *Pet. Sci. Technol.* **2012**, *30*, 350–359. [CrossRef]
11. Rampersad, P.; Hareland, G.; Boonyapaluk, P. Drilling optimization using drilling data and available technologyIn Proceedings of the SPE Latin America and Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina, 27–29 April 1994.
12. Warren, T. Penetration-rate performance of roller-cone bits. *SPE Drill. Eng.* **1987**, *2*, 9–18. [CrossRef]
13. Wu, A.; Hareland, G.; Lei, L.; Lin, Y.; Yang, Y. Modeling and prediction of cone rotary speed of roller cone bits. In Proceedings of the SPE Canada Unconventional Resources Conference, Calgary, AB, Canada, 5–7 November 2013.
14. Koolivand-Salooki, M.; Esfandyari, M.; Rabbani, E.; Koulivand, M.; Azarmehr, A. Application of genetic programing technique for predicting uniaxial compressive strength using reservoir formation properties. *J. Pet. Sci. Eng.* **2017**, *159*, 35–48. [CrossRef]
15. McElroy, P.D.; Bibang, H.; Emadi, H.; Kocoglu, Y.; Hussain, A.; Watson, M.C. Artificial neural network (ANN) approach to predict unconfined compressive strength (UCS) of oil and gas well cement reinforced with nanoparticles. *J. Nat. Gas Sci. Eng.* **2021**, *88*, 103816. [CrossRef]

16. Hiba, M.; Ibrahim, A.F.; Elkatatny, S. Real-time prediction of tensile and uniaxial compressive strength from artificial intelligence-based correlations. *Arab. J. Geosci.* **2022**, *15*, 1546. [CrossRef]
17. Ibrahim, A.F.; Hiba, M.; Elkatatny, S.; Ali, A. Estimation of tensile and uniaxial compressive strength of carbonate rocks from well-logging data: Artificial intelligence approach. *J. Pet. Explor. Prod. Technol.* **2024**, *14*, 317–329. [CrossRef]
18. Warren, T.M. Drilling model for soft-formation bits. *J. Pet. Technol.* **1981**, *33*, 963–970. [CrossRef]
19. Hareland, G.; Hoberock, L. Use of drilling parameters to predict in-situ stress bounds. In Proceedings of the SPE/IADC Drilling Conference and Exhibition, Amsterdam, The Netherlands, 22–25 February 1993.
20. Hareland, G.; Rampersad, P. Drag-bit model including wear. In Proceedings of the SPE Latin America and Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina, 27–29 April 1994.
21. Winters, W.; Warren, T.; Onyia, E. Roller bit model with rock ductility and cone offset. In Proceedings of the SPE Annual Technical Conference and Exhibition, Dallas, TX, USA, 27–30 September 1987.
22. Hareland, G.; Nygaard, R. Calculating unconfined rock strength from drilling data. In Proceedings of the 1st Canada-US Rock Mechanics Symposium, Vancouver, BC, Canada, 27–31 May 2007.
23. Hope, T.M. Linear regression. In *Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 67–81.
24. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
25. Filzmoser, P.; Nordhausen, K. Robust linear regression for high-dimensional data: An overview. *Wiley Interdiscip. Rev. Comput. Stat.* **2021**, *13*, e1524. [CrossRef]
26. Genuer, R.; Poggi, J.-M.; Genuer, R.; Poggi, J.-M. *Random Forests*; Springer: Berlin/Heidelberg, Germany, 2020.
27. Babar, B.; Luppino, L.T.; Boström, T.; Anfinsen, S.N. Random forest regression for improved mapping of solar irradiance at high latitudes. *Sol. Energy* **2020**, *198*, 81–92. [CrossRef]
28. Fratello, M.; Tagliaferri, R. Decision trees and random forests. Encyclopedia of Bioinformatics and Computational Biology. *ABC Bioinform.* **2019**, *1*, 374–383. [CrossRef]
29. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
30. Aria, M.; Cuccurullo, C.; Gnasso, A. A comparison among interpretative proposals for Random Forests. *Mach. Learn. Appl.* **2021**, *6*, 100094. [CrossRef]
31. Xue, L.; Liu, Y.; Xiong, Y.; Liu, Y.; Cui, X.; Lei, G. A data-driven shale gas production forecasting method based on the multi-objective random forest regression. *J. Pet. Sci. Eng.* **2021**, *196*, 107801. [CrossRef]
32. Antoniadis, A.; Lambert-Lacroix, S.; Poggi, J.-M. Random forests for global sensitivity analysis: A selective review. *Reliab. Eng. Syst. Saf.* **2021**, *206*, 107312. [CrossRef]
33. González, S.; García, S.; Del Ser, J.; Rokach, L.; Herrera, F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Inf. Fusion* **2020**, *64*, 205–237. [CrossRef]
34. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
35. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [CrossRef]
36. Suykens, J.A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
37. Suykens, J.A.; Vandewalle, J. Chaos control using least-squares support vector machines. *Int. J. Circuit Theory Appl.* **1999**, *27*, 605–615. [CrossRef]
38. Ahmadi, M.A.; Ebadi, M.; Hosseini, S.M. Prediction breakthrough time of water coning in the fractured reservoirs by implementing low parameter support vector machine approach. *Fuel* **2014**, *117*, 579–589. [CrossRef]
39. Ahmadi, M.A.; Ebadi, M. Evolving smart approach for determination dew point pressure through condensate gas reservoirs. *Fuel* **2014**, *117*, 1074–1084. [CrossRef]
40. Ahmadi, M.A.; Ebadi, M.; Shokrollahi, A.; Majidi, S.M.J. Evolving artificial neural network and imperialist competitive algorithm for prediction oil flow rate of the reservoir. *Appl. Soft Comput.* **2013**, *13*, 1085–1098. [CrossRef]
41. Rzhevskii, V.V.; Novik, G.Y. *Fundamentals of the Physics of Rocks*; Nedra: Moscow, Russia, 1967.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.