

# Prediction of Young's Modulus for Reservoir Rocks using Machine Learning

VAN Tu Truong<sup>1</sup>, KHAC Long Nguyen<sup>1,\*</sup>, THE Vinh Nguyen<sup>1</sup>, TIEN Hung Nguyen<sup>1</sup>,  
DUC Thinh Kieu<sup>2</sup>, VAN Do Nguyen<sup>3</sup>, QUOC Hung Le<sup>4</sup>, SERGIO Vinciguerra<sup>5</sup>

<sup>1</sup> Hanoi University of Mining and Geology, Hanoi, Vietnam

<sup>2</sup> Thuy Loi University, Hanoi, Vietnam

<sup>3</sup> Vietnam Petroleum Institute-VPI, Vietnam

<sup>4</sup> Joint Venture Vietsovpetro, Vietnam

<sup>5</sup> University of Turin, Italy

\* Corresponding author: [nguyenkhaclong@hmg.edu.vn](mailto:nguyenkhaclong@hmg.edu.vn)

**Abstract:** Young's Modulus is an important mechanical property of rocks in hydraulic fracturing design, as it determines their stiffness and indicates the extent to which they will deform elastically under uniaxial compressive stress. Young's Modulus for reservoir rock can be determined via laboratory testing or using well log geophysical data. However, the laboratory testing depends on the availability of core samples, while data analysis based on well log data requires appropriate conversion formulas, which demand significant time, cost, and complexity.

The approach presented in this study is based on the fact that real time drilling data, such as weight on bit (WOB), torque on bit (TOR), standpipe pressure (SPP), rotary speed (RPM), rate of penetration (ROP), and drilling fluid flow rate (FLOWIN), are readily available in the early stages of the drilling process without incurring additional costs. Two machine learning algorithms were used to correlate drilling data with the dynamic Young's Modulus : Random Forest and Decision Tree. Two different datasets were used in this study, the first dataset was utilized to develop and train the model, while the other dataset was used to validate the developed models. The first of the two methods employed demonstrated a remarkable match between the given values and the predicted values. The correlation coefficients ranged from 0.71 to 0.95, with the average absolute percentage error being less than 5%.

Based on the obtained results, the use of drilling data combined with artificial intelligence models for predicting Young's Modulus proves to be a viable approach.

**Keywords:** Young's Modulus; Drilling data; Hydraulic Fracturing; Random Forest ; Decision Tree

## 1. Introduction

Rock Young's Modulus refers to the ability of rock to recover elastic deformation caused by external loading. Young's Modulus is one of the geomechanical characteristics that define the relationship between external forces and the resulting deformation (Fjaer et al., 2008). Young's Modulus of rock is a key factor influencing hydraulic fracturing design, drilling performance, in-situ stress calculations, and wellbore stability (Hammah et al., 2006; Nes et al., 2005).

Depending on the method of determination, there are two common types of Young's modulus. Static Young's modulus is typically determined from the stress-strain curve obtained under slow, sustained loads, while dynamic Young's modulus is derived from the velocity of elastic waves propagating through the material. The former can be determined directly through compression tests on core samples and the latter is derived from the velocities of compressional and shear waves obtained from well log geophysical data (Barree et al., 2009). The static Young's Modulus (E) represents the stiffness of rock and soil and is defined as the ratio of strain to stress according to Hooke's Law (Equation 1). Additionally, Equation (2) is used to determine the dynamic Young's Modulus using the acoustic wave method.

$$\sigma = E_{\text{sta}} \varepsilon \quad (1)$$

$$E_{\text{dyn}} = \frac{\rho V_s^2 (3V_p^2 - 4V_s^2)}{V_p^2 - V_s^2} \quad (2)$$

Whereas:  $\sigma$  is stress,  $\varepsilon$  is strain,  $\rho$  is the bulk density ( $\text{g/cm}^3$ ),  $E_{\text{sta}}$  and  $E_{\text{dyn}}$  are the static and dynamic Young's Modulus (GPa),  $V_s$  and  $V_p$  are the shear and compressional wave velocity (km/s), respectively.

Laboratory testing methods offer high accuracy but typically require the availability of cores for testing, measuring equipment such as uniaxial testing machines, which can be time-consuming and costly (Müller et al., 2019; Ifeobia & Ahmad, 2020). However, laboratory sample data are often limited in number and do not provide continuous measurements along the borehole depth due to the high costs associated with sampling and testing.

To obtain reliable Young's Modulus values from well log geophysical data, appropriate conversion formulas under reservoir conditions are required. Developing such empirical conversion formulas demands significant time, cost and complexity. Additionally, key parameters needed to calculate the Young's Modulus, such as bulk density, compressional wave velocity, and shear wave velocity, are not always available in geophysical logging measurements (Siddig et al., 2021; Darvish et al., 2015).

Currently, by applying Artificial Intelligence (AI) techniques using input parameters from drilling data obtained in the early stages of drilling well construction (Gowida & Elkatatny, 2020; Gowida et al., 2020), geomechanical parameters can be predicted more simply, quickly, and cost-effectively compared to traditional methods.

The oil produced from the Oligocene formation in the White Tiger field, Vietnam, has lower production rates compared to other formations. The reservoir rocks in this formation are primarily lenticular in shape, with low permeability and high heterogeneity. For the Oligocene oil reservoirs with low storage capacity due to various factors, it is necessary to apply geo-technical solutions to improve oil recovery, including hydraulic fracturing. To develop an effective hydraulic fracturing (HF) plan, one of the key information required is to assess the geomechanical parameters. These parameters provide the initial stress conditions of the formation before fracturing and serve as accurate inputs for predicting rock deformation during and after the HF process, as well as determining the necessary fracture initiation pressure. The geomechanical parameters for HF include Poisson's ratio, Young's modulus, compressive strength, minimum horizontal stress, and maximum horizontal stress. Here, we present a new method to forecast Young's Modulus using various AI techniques, with drilling parameters as input data for the Oligocene formation in the White Tiger field to support the HF process.

## **2. Research Methodology**

### **2.1 Methods and Data Description**

In this study, a model based on the Random Forest and Decision Tree algorithms was used to predict dynamic Young's Modulus using the real time drilling parameters such as weight on bit (WOB), torque on bit (TOR), standpipe pressure (SPP), rotary speed (RPM), rate of penetration (ROP), and drilling fluid flow rate (FLOWIN). The actual dynamic Young's Modulus values obtained from wave velocity logs were used to train and validate the model.

The predicted dynamic Young's modulus from the more accurate built model was converted to static conditions using an empirical correlation for static Young's modulus. Due to the limited availability of core samples, some of the converted Young's modulus values will be compared to those obtained from laboratory tests on core samples.

The data used in this study were collected from two wells, A and B, drilled through the Oligocene formation in the White Tiger field, Vietnam. The lithologies of the upper and lower Oligocene formations in the wells consist of shale, mudstone and sandstone as given in Fig 1. A full data set of drilling parameters is available for the two wells, spanning from well log data to core testing. The dataset includes Young's Modulus values calculated from geophysical logging data while drilling an "8.5-inch" borehole section through the Oligocene formation. Well A contains a total of 714 data points, with 70% used for model training and 30% for model testing. Meanwhile, 196 data points from Well B were used to validate the developed model.

The model was trained with Young's Modulus as the output parameter, the drilling parameters and actual dynamic Young's Modulus values from wells A and B used as input. The predicted Young's Modulus values for Well B were compared with values obtained from core testing.

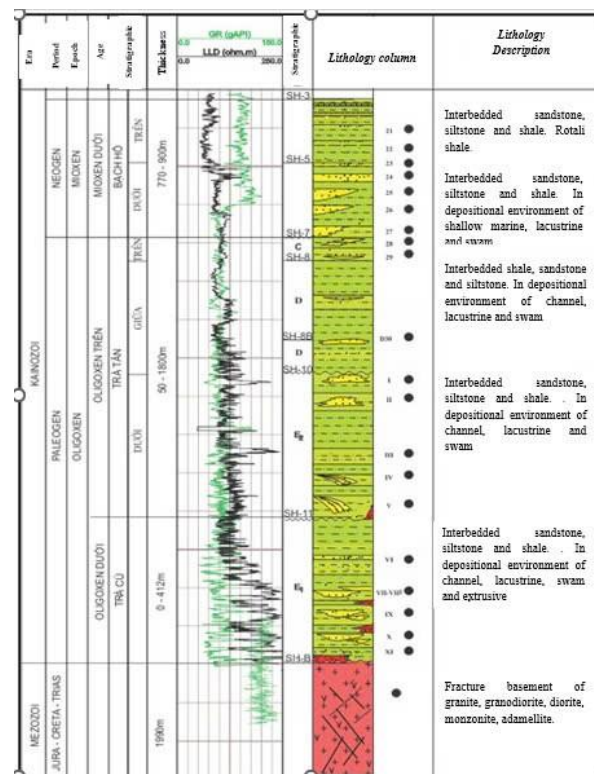


**Fig. 1.** Lithology column for Well A, the White Tiger Field

## 2.1 Data Analysis

Selecting input parameters for training the model is a crucial step in determining its reliability. In order to decide which parameters should be used as input data for training the network, we analyzed the relationship between drilling parameters and the Young's Modulus based on datasets from Wells A and B in the White Tiger field.

The selection of input parameters for training and testing is a key factor controlling the model. The correlation coefficients between the Young's Modulus and various drilling parameters and their relative importances are shown in Fig 2a. As seen in Fig 2a, all correlation coefficients between drilling parameters and the Young's Modulus are less than 0.25. Therefore, applying artificial intelligence models is expected to yield better results than linear regression methods, as AI can better match more complex relationships.

Fig 2b shows a relatively strong correlation between Young's Modulus and certain drilling parameters, such as standpipe pressure (SPP), torque on bit (TOR), weight on bit (WOB), and rate of penetration (ROP). Lower correlation coefficients for other parameters do not necessarily imply the absence of a relationship with the Young's Modulus. Instead, they indicate that a linear equation does not adequately describe the relationship between the inputs and outputs.

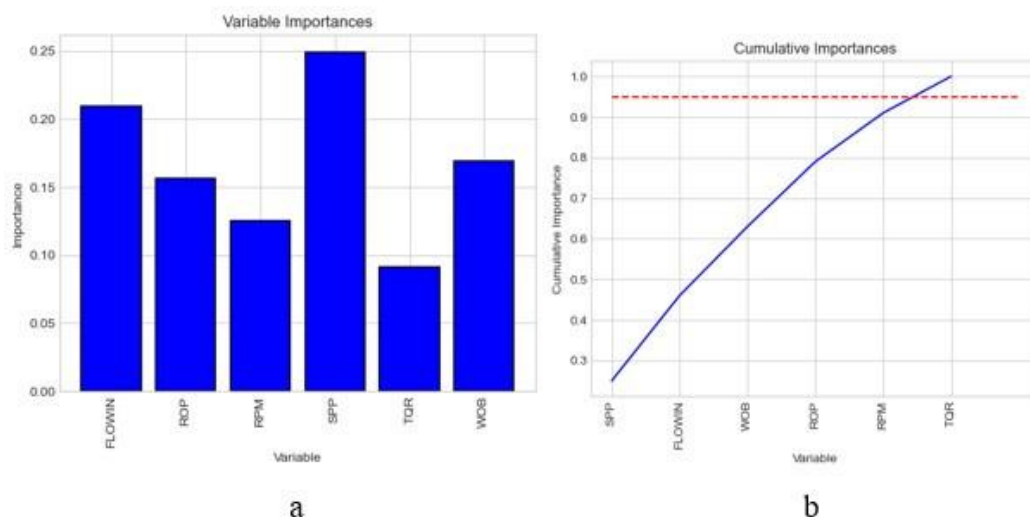


**Fig. 2.** Correlation relationship between Young's Modulus and the input parameters

Before data was input into machine learning algorithms, the datasets were denoised and outliers were identified and removed using data normalization methods (Tripathy, 2013). The data analysis involved examining the correlation between variables and normalizing the data. The statistical analysis of the dataset used to develop the prediction model is presented in Tab. 1.

**Tab. 1.** Statistical analysis of the dataset used for the Young's Modulus prediction model

|       | FLOWIN | ROP   | RPM    | SPP    | TQR     | WOB   | YOUNG_MODULUS |
|-------|--------|-------|--------|--------|---------|-------|---------------|
| Count | 714    | 714   | 714    | 714    | 714     | 714   | 7.140000e+02  |
| Mean  | 37.78  | 17.44 | 117.47 | 195.26 | 1575.95 | 8.27  | 2.321520e+06  |
| Std   | 9.30   | 9.99  | 20.96  | 24.18  | 204.29  | 1.90  | 3.388567e+04  |
| Min   | 22.06  | 0.78  | 40.00  | 143.32 | 1037.60 | 2.33  | 2.012000e+06  |
| 25%   | 34.77  | 11.86 | 116.00 | 184.28 | 1534.15 | 7.07  | 2.313250e+06  |
| 50%   | 38.05  | 17.35 | 121.00 | 202.85 | 1593.30 | 8.41  | 2.326000e+06  |
| 75%   | 38.10  | 21.07 | 122.00 | 212.06 | 1669.88 | 9.20  | 2.341000e+06  |
| Max   | 54.64  | 45.40 | 161.00 | 224.00 | 2185.30 | 13.87 | 2.399000e+06  |

## 2.3 Data Processing

### 2.3.1 Data Normalization

During the training process, anomalous or outlier values can be considered as noise, as they may affect the reliability and reproducibility of the model. Therefore, before using the data as input for model training, the drilling parameters were adjusted and outliers were removed based on an algorithm that identifies outlier values through data normalization. The data normalization index is determined using the following formula:

$$X_{\text{standard}} = \frac{X - \text{mean}(X)}{\text{standard deviation}(X)} \quad (3)$$

Whereas:

- Mean (X) - the average value of the parameter;
- Standard deviation (X) - the standard deviation of the parameter.

The  $X_{\text{standard}}$  value was selected to identify noise in the dataset as follows (Tripathy, 2013):

- When  $X_{\text{standard}} < 2$ , the value is suitable for use;
- When  $2 < X_{\text{standard}} < 3$ , the value is questionable and it may affects the model training;
- When  $X_{\text{standard}} > 3$ , the value is considered noisy and it should be removed.

### 2.3.2 Data Standardization

After selecting the input parameters for model development, these parameters are standardized before being used for model training and testing. The input dataset consists of multiple parameters, each with different units and magnitudes. This variation can negatively impact the model performance and accuracy (<https://scikit-learn.org>). Therefore, data scaling is applied after splitting the dataset into training and testing sets to ensure that all parameters are within the same range [0,1]. This prevents scale effects with a given parameter having a significantly larger magnitude with respect to the others. The scaling is performed using the following algorithm and formula:

$$X_{\text{Normalize}} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (4)$$

Whereas:

- $X_{\text{Normalize}}$  - Normalized value;
- X – Actual value;
- $\min(X)$  – Minimum value;
- $\max(X)$  – Maximum value.

## 2.4 Model Development

To develop models that establish the relationship between Young's Modulus and drilling parameters, two machine learning algorithms were used separately: Decision Tree and Random Forest. Both of these algorithms have two capabilities—classification and regression. Within the scope of this study, only regression was applied and presented.

### 2.4.1 Decision Tree Algorithm

A Decision Tree is a type of predictive model that maps observations of an object or phenomenon to conclusions about its target value.

The training data for Decision Tree Regression is represented in the following form:  $(x, y) = (x_1, x_2, \dots, x_k, y)$

Whereas  $y$  is the target variable, which represents Young's Modulus, and  $x_1, x_2, \dots, x_k$  are the independent observed variables, representing the drilling parameters.

The process of constructing a regression tree consists of two steps (Sohil và nnk, 2022):

1. The prediction space, which consists of the set of values  $x_1, x_2, \dots, x_k$  – is divided into  $J$  distinct and non-overlapping regions:  $R_1, R_2, \dots, R_J$ .
2. For every observed variable that falls into region  $R_j$ , the prediction is assigned as the mean value of the target variable for the training observations within  $R_j$ .

To construct the optimal regions  $R_1, R_2, \dots, R_J$ , the prediction space is divided into multi-dimensional boxes in a way that minimizes the Residual Sum of Squares (RSS):

$$RSS = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \quad (5)$$

Where  $\hat{y}_{R_j}$  represents the mean value of the target variable in the  $j$ -th region.

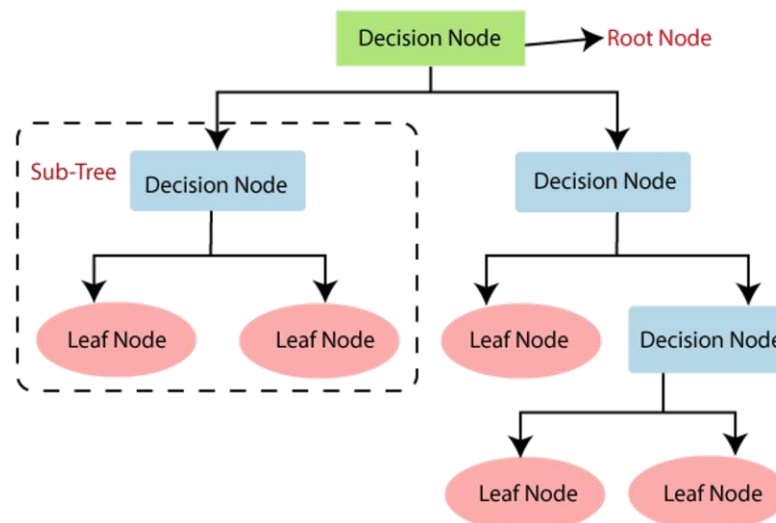


**Fig. 3.** Decision Tree Diagram (Koen, 2016)

A Decision Tree has a tree-like structure and represents a decision-making method for classifying given events. Each node in the tree indicates a class label or a specific test, which splits the data space at that given node. Each resulting subset represents a sub-space of the data corresponding to a sub-classification. This division is performed through a corresponding sub-tree. The construction of a Decision Tree can be viewed as a divide-and-conquer strategy for object classification (Pal & Jain, 2005; Jiawei et al., 2012). A Decision Tree can be described using the concepts of nodes and edges connecting the nodes.

Each node in a Decision Tree can be classified as follows:

- Leaf node (or terminal node): Also called an answer node, it represents a class of cases (records), with its label being the class name.
- Non-leaf node (or internal node): This node defines an attribute test, with the node label being the attribute name. Each branch (or edge) connecting this node to a sub-tree corresponds to one of the possible outcomes of the test. The label of the branch represents the attribute value. The topmost non-leaf node is called the

root node.

A Decision Tree classifies data by starting at the root node and traversing through the tree following the branches until it reaches a leaf node, at which point the class of the given data is determined.

#### 2.4.2 Random Forest Algorithm

Random Forest is an ensemble learning algorithm proposed by Breiman in 2001 (Breiman, 2001). It constructs a large number of random decision trees using bootstrapped training samples and aggregates their predictions by averaging the results (Scornet et al., 2015). It has become a major data mining tool for both regression and classification problems. Recently, the consistency of the Random Forest algorithm was demonstrated by Scornet in 2015 (Scornet et al., 2015). Compared to other machine learning algorithms, such as neural networks, Random Forest can achieve relatively high predictive performance with only a few parameters to tune (Genuer et al., 2017).

Random Forest is a supervised classification algorithm based on Decision Trees and the Bagging and Bootstrapping methods, which have been further improved. Bootstrapping is a well-known statistical method introduced by Efron in 1979 (Efron, 1979). The Bagging method is considered an ensemble approach, where results from multiple bootstrapped samples are aggregated. The model is trained on these randomly generated samples, and the final prediction is obtained by averaging the predicted values for regression problems or selecting the class with the highest classification score.

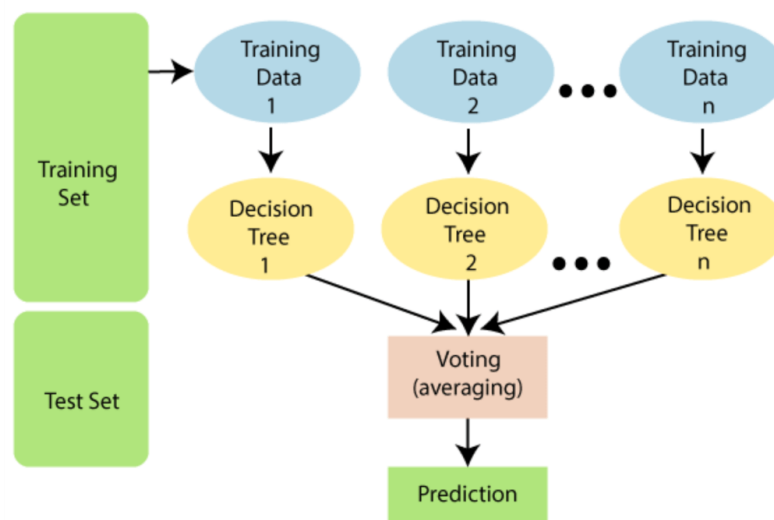


**Fig. 4.** Diagram of the Random Forest Algorithm

Random Forest Algorithm:

- Generate random datasets (bootstrap samples) from the training dataset  $D$ ;
- Use the randomly sampled subsets  $D_1, D_2, \dots, D_n$  to construct multiple decision trees  $T_1, T_2, \dots, T_n$ ;
- Combine the decision trees: For classification problems, use the majority voting strategy; For regression problems, take the average of the predicted values from all trees.

Several open-source libraries implement Decision Tree and Random Forest algorithms, one of which is scikit-learn (Pedregosa và nnk,; <https://scikit-learn.org/>). Scikit-learn is a widely used machine learning library and was chosen for this study. The parameter settings used in this research are described in the following section.

#### 2.4.3 Selection of Parameters for Random Forest and Decision Tree

The selection of parameter sets for both Decision Tree and Random Forest algorithms is described in Step 3 of the algorithm diagram in Fig 1 (Diagram of the model used to predict the Young's Modulus). The parameter sets for both algorithms are presented in Tab.s 2 and 3, respectively. (<https://scikit-learn.org/>).

**Tab. 2.** Parameter Sets Used for the Random Forest Model

| RANDOM FOREST |              |           |                   |                  |           |
|---------------|--------------|-----------|-------------------|------------------|-----------|
| n_estimators  | max_features | max_depth | min_samples_split | min_samples_leaf | bootstrap |
| 1             | auto         | None      | 1                 | 1                | True      |
| 5             | sqrt         | 1         | 2                 | 2                | False     |
| 10            |              | 5         | 3                 | 3                |           |
| 100           |              | 10        | 4                 | 4                |           |
| 200           |              | 15        | 5                 | 5                |           |
| 300           |              | 20        | 10                | 10               |           |
| 400           |              | 25        |                   |                  |           |
| 500           |              | 30        |                   |                  |           |
| 600           |              |           |                   |                  |           |
| 700           |              |           |                   |                  |           |
| 800           |              |           |                   |                  |           |
| 900           |              |           |                   |                  |           |
| 1000          |              |           |                   |                  |           |

**Tab. 3.** Parameter Sets Used for the Decision Tree Model

| DECISION TREE |           |                   |                  |
|---------------|-----------|-------------------|------------------|
| max_features  | max_depth | min_samples_split | min_samples_leaf |
| Auto          | None      | 1                 | 1                |
| Sqrt          | 1         | 2                 | 2                |
| log2          | 5         | 3                 | 3                |
| None          | 10        | 4                 | 4                |
|               | 15        | 5                 | 5                |
|               | 20        | 10                | 10               |
|               | 25        |                   |                  |
|               | 30        |                   |                  |

After performing multiple iterations of the algorithm and using the parameter sets from Tab.s 2 and 3, the optimal parameter set was selected for each model:

- For the Random Forest model: 'n\_estimators': 400; 'min\_samples\_split': 3; 'min\_samples\_leaf': 2; 'max\_features': 'sqrt'; 'max\_depth': 20; 'bootstrap': False;
- For the Decision Tree model: 'min\_samples\_split': 5; 'min\_samples\_leaf': 1; 'max\_features': 'auto'; 'max\_depth': 10.

#### 2.4.4 Model Evaluation

Random Forest and Decision Tree algorithms were used to develop the model. These algorithms utilized 70% of the dataset from Well A for model training and 30% for internal testing through multiple iterations to select the most suitable model. Once the model was finalized, data from Well B was used as

an external validation set.

To evaluate all model experiments, two statistical parameters were used: the correlation coefficient (R) and the average absolute percentage error (AAPE). These metrics were calculated using Equation (6) and Equation (7):

$$R = \frac{[N \sum_{i=1}^N (x_{\text{given } i} \times x_{\text{Predicted } i})] - [\sum_{i=1}^N (x_{\text{given } i} \times x_{\text{Predicted } i})]}{\sqrt{[N \sum_{i=1}^N (x_{\text{given } i})^2 - (\sum_{i=1}^N x_{\text{given } i})^2][N \sum_{i=1}^N (x_{\text{Predicted } i})^2 - (\sum_{i=1}^N x_{\text{Predicted } i})^2]}} \quad (6)$$

$$AAPE = \frac{\sum_{i=1}^N \frac{x_{\text{given } i} - x_{\text{Predicted } i}}{x_{\text{given } i}} \times 100\%}{N} \quad (7)$$

Whereas:

- $x_{\text{given}}$  - Actual Young's Modulus from well log data;
- $x_{\text{predicted}}$  - Predicted Young's Modulus;
- $N$  - Total number of data points.
- $\max(X)$  - Maximum value.

The correlation coefficient R is interpreted as follows:

- If R is between  $\pm 0.50$  and  $\pm 1$ , it is considered a strong correlation;
- If R is between  $\pm 0.30$  and  $\pm 0.49$ , it is considered a moderate correlation;
- If R is below  $\pm 0.29$ , it is considered a weak correlation.

The average absolute percentage error (AAPE) is interpreted as follows:

- $AAPE < 5\%$ : Indicates that the prediction model has high reproducibility;
- $AAPE$  between 10% and 25%: Suggests moderate but acceptable reproducibility;
- $AAPE > 25\%$ : Indicates that the model has very low reproducibility.

### 3. Results and Discussion

#### 3.1 Training, Testing, and Validation of Young's Modulus Prediction Model Using the Random Forest Algorithm

##### 3.1.1 Training and Testing of Young's Modulus Prediction Model Using the Random Forest Algorithm

Based on drilling parameter data, including weight on bit (WOB), torque on bit (TQR), standpipe pressure (SPP), rotary speed (RPM), rate of penetration (ROP), and drilling fluid flow rate (FLOWIN), the authors focused on developing a Young's Modulus prediction model using the Random Forest algorithm. The model's output was collected, and errors were calculated during the training and testing phases. The Random Forest method demonstrated high accuracy and effectiveness in predicting Young's Modulus.

Figs 5 and 6 confirm the high accuracy of the Random Forest-based model, which achieved a mean absolute percentage error (AAPE) of 0.25% and 0.50% for training and testing phases, respectively. Meanwhile, the correlation coefficient (R) for Young's Modulus prediction model was 0.95 for the training phase and 0.71 for the testing phase.

##### 3.1.2. Validation of Young's Modulus Prediction Model Using the Random Forest Algorithm

The authors also validated the Random Forest-based model using data points from Well B, comparing the measured and the predicted values. Model evaluation metrics were calculated to demonstrate the reliability and accuracy of the developed prediction model. For the validation dataset, the model achieved a correlation coefficient (R) of 0.71 and an average absolute percentage error (AAPE) of 0.60% (Fig 7).

These results confirm the feasibility of using the empirical correlation developed in this study, based on drilling parameters, to enable real-time evaluation and prediction of Young's Modulus.

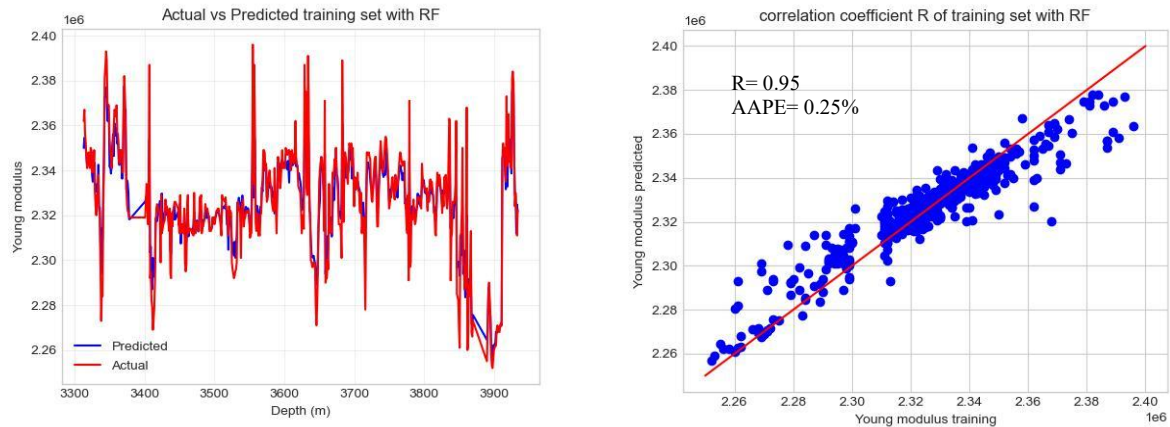


**Fig. 5.** Measured Young's Modulus vs. Predicted Young's Modulus by Random Forest for Training Phase

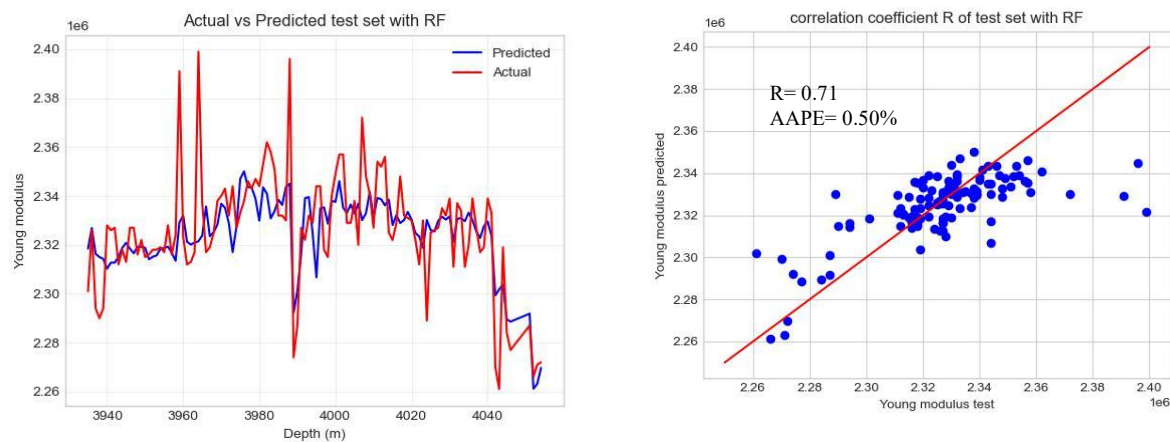


**Fig. 6.** Measured Young's Modulus vs. Predicted Young's Modulus by Random Forest for Testing Phase

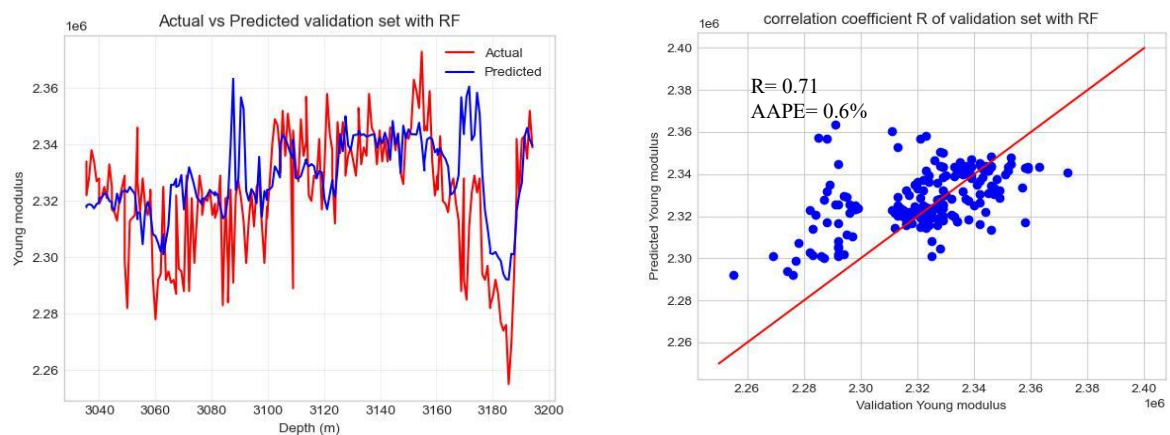


**Fig. 7.** Measured Young's Modulus vs. Predicted Young's Modulus by Random Forest for Validation Phase

The prediction model for the training, testing, and validation datasets achieved a high correlation coefficient and low error because Random Forest generates an ensemble of unpruned Decision Trees. Each tree is built on a bootstrap sample (random sampling with replacement), and at each node the best split is determined from a randomly selected subset of attributes. The generalization error of the forest depends on the accuracy of each individual tree and the degree of dependence between the trees. The Random Forest algorithm achieves high accuracy compared to other current supervised learning algorithms by maintaining low bias (the error unrelated to training data) and using randomness to reduce the correlation between trees in the forest.

### 3.2. Training, Testing, and Validation of Young's Modulus Prediction Model Using the Decision Tree Algorithm

#### 3.2.1. Training and Testing of Young's Modulus Prediction Model Using the Decision Tree Algorithm

The Decision Tree-based model was trained to predict Young's Modulus based on drilling parameters. The developed correlation allows for real-time prediction of Young's Modulus in cases where well log and geophysical data are unavailable. The results indicate that the Dynamic Young's Modulus predicted by the Decision Tree model had an error of 0.01% during the training phase and 0.68% during the testing phase.

The correlation coefficient (R) for the training dataset was 0.99, while for the testing dataset, it was 0.54. This indicates overfitting, as the model accurately predicts the training data but fails to generalize the results (Figs 8 and 9). This issue arises because the model memorizes the training data too well, making it overly dependent on it. As a result, it fails to generalize patterns for new, unseen data — a common challenge associated with Decision Tree algorithms.

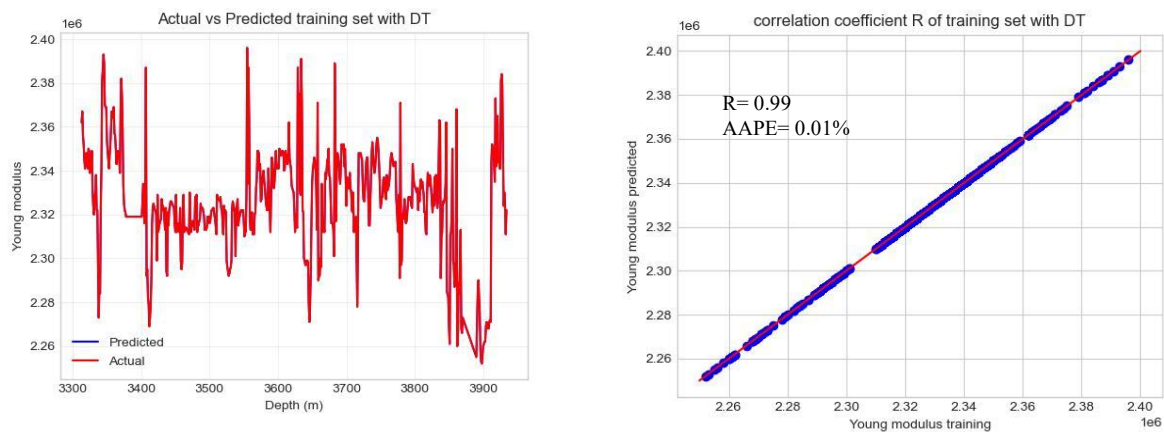


**Fig. 8.** Measured Young's Modulus vs. Predicted Young's Modulus by Decision Tree for Training Phase

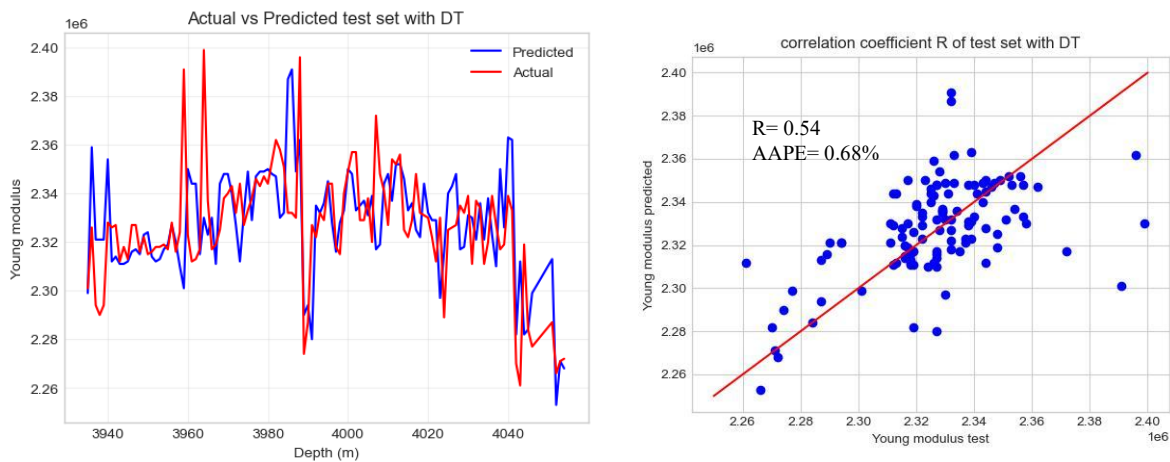


**Fig. 9.** Measured Young's Modulus vs. Predicted Young's Modulus by Decision Tree for Testing Phase

#### 3.2.2. Validation of Young's Modulus Prediction Model Using the Decision Tree Algorithm

The Decision Tree-based model predicted Young's Modulus for the validation dataset with an average absolute percentage error (AAPE) of 0.63% and a correlation coefficient (R) of 0.45. It can be observed that the validation dataset had only a moderate correlation coefficient, which is due to one of the limitations of Decision Tree algorithms. Decision Trees often struggle with time-dependent data, making it difficult to construct an optimal tree. Additionally, they are prone to overfitting, meaning they can create overly complex trees that fit the training data too well.

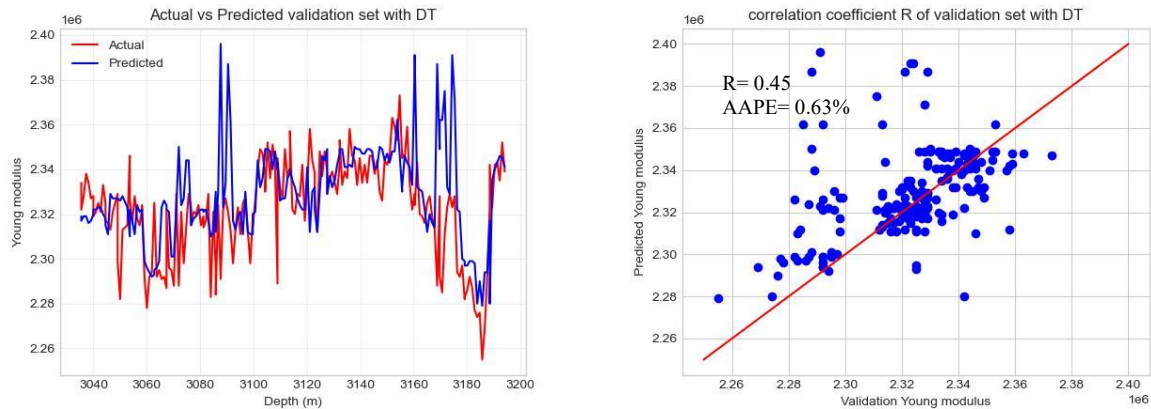


**Fig. 10.** Measured Young's Modulus vs. Predicted Young's Modulus by Decision Tree for Validation Phase

### 3.3. Compare the predicted Young's modulus values with the core sample test results

As discussed in Sections 3.1 and 3.2, it is evident that the Young's modulus values predicted by both the Random Forest and Decision Tree models have an Average Absolute Percentage Error (AAPE) below 5%. However, the correlation coefficient (R) during the testing and validation phases of the Random Forest model is higher than that of the Decision Tree model. This indicates that the Random Forest model provides better predictions of dynamic Young's modulus based on input data of drilling parameters.

With only six core samples collected from the Oligocene formation of Well B, the authors conducted a comparison between the static Young's modulus obtained from laboratory tests and the dynamic Young's modulus predicted by the Random Forest model for Well B, after converting it to static elastic modulus values to evaluate the model's accuracy.

Currently, most formulas for determining Young's modulus are developed based on core sample analysis and geophysical logging data. However, studies on the conversion relationship between static and dynamic Young's modulus specifically for the Bach Ho (White Tiger) field remain limited. In the study by P.V. Hieu (2021), an empirical correlation for static Young's modulus was proposed for the White Tiger field by the Vietnam–Russia Joint Venture (Vietsovetro). This correlation was developed based on dynamic and static measurements of core samples conducted under simulated reservoir temperature and pressure conditions (P.V. Hieu, 2021):

$$E_{sta} = 0.9018 * E_{dyn} - 1.5766 \quad (8)$$

The dynamic Young's modulus values from Well B, predicted using the Random Forest model, were converted to static conditions using Equation (8). The comparison results between the Young's modulus predicted by the Random Forest model converted by Equation (8) ( $E_{RF}$ ) and those obtained from core sample laboratory tests ( $E_{core}$ ) are shown in Fig 11.

The comparison results show a relatively good agreement between the static Young's modulus values obtained from core sample experiments and the values converted from the predicted dynamic Young's modulus using the empirical formula developed from research on the Bach Ho field. This confirms the accuracy of the Random Forest model in predicting Young's modulus for the Oligocene formation of the Bach Ho field.

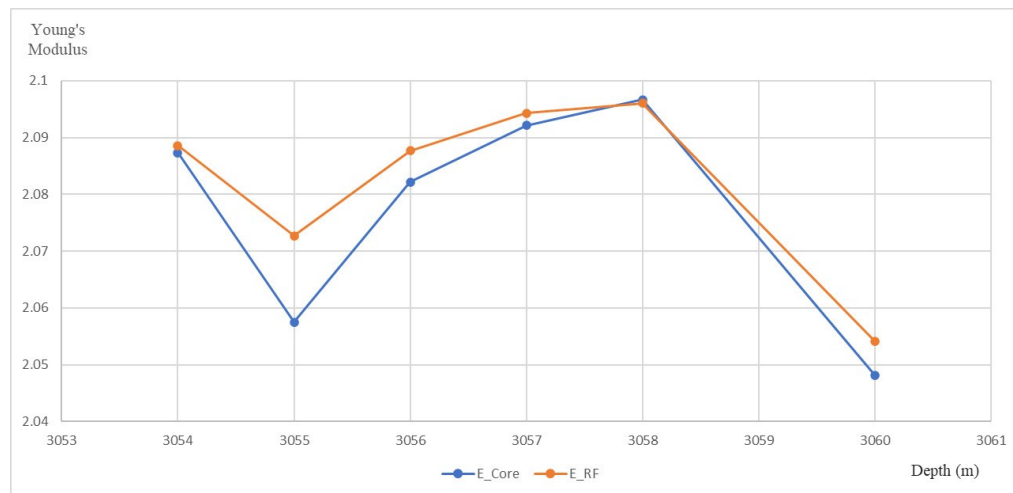


**Fig. 11.** Comparison of Young's modulus values predicted by the Random Forest model (after conversion) and values obtained from core sample laboratory tests

#### 4. Conclusion

Typically, Young's Modulus is determined using two traditional methods: the acoustic wave method and core sample compression testing in laboratories. However, these data are not always available. The prediction of Young's Modulus using drilling parameters has been explored in this study through the application of artificial intelligence. Based on the presented results, the authors have drawn the following conclusions:

- Compared to other methods used to determine Young's Modulus, applying artificial intelligence (AI) to predict Young's Modulus from drilling data establishes a stronger correlation between Young's Modulus and related parameters. This approach enhances prediction reproducibility, supporting hydraulic fracturing operations in production zones and various oil and gas industry applications. Additionally, it helps reduce the costs and time while addressing the lack of measurement data, auxiliary measurement devices, and supplementary core sample test results. Therefore, predicting Young's Modulus from drilling data offers significant practical benefits.
- Two machine learning methods were studied, and both produced reliable results. However, the Random Forest model provided better predictions of Young's Modulus. The optimization of different parameters used in the algorithm was presented, and the best-performing results were reported.
- The correlation coefficient between actual and predicted values exceeded 0.7, and the average absolute percentage error remained below 5% for the Random Forest model. The best model was selected for predicting Young's Modulus across different datasets.
- The agreement between the predicted dynamic Young's modulus converted with the empirical formula developed for the reservoir in the Bach Ho field and the values obtained from laboratory testing demonstrates the validity of the Random Forest model in predicting Young's modulus.

#### References

- Barree, R.D.; Gilbert, J.V.; Conway, M. , 2009. Stress and rock property profiling for unconventional reservoir stimulation. In: SPE Hydraul. Fract. Technol. Conf. <https://doi.org/10.2118/118703-MS>
- Breiman L., 2001. Random Forests. Machine Learning Journal Paper, 45(1), 5–32.
- Darvish H., Nouri-Taleghani M., Shokrollahi A. et al., 2015. Geo-mechanical modeling and selection of suitable layer for hydraulic fracturing operation in an oil reservoir (south west of Iran). Journal of African Earth Sciences, 111, 409–420.
- Efron B., 1979. Bootstrap Methods: Another look at the Jackknife. The Annals of statistics, 7(1), 26.
- Fjaer, E.; Holt, R.M.; Raaen, A.M.; Horsrud, P., 2008. Petroleum related rock mechanics, Vol. 53. Elsevier Science.
- Genuer R., Poggi J.-M., Tuleau-Malot C. et al., 2017. Random Forests for Big Data. Big Data Research, 9, 28–46.
- Gowida, A., and Elkatatny, S., 2020, "Prediction of Sonic Wave Transit Times From Drilling Parameters While Horizontal Drilling in Carbonate Rocks Using Neural Networks," Petrophysics, 61(5), pp.

482–494.

- Gowida, A., Elkatatny, S., Al-afnan, S., and Abdulraheem, A., 2020, “New Computational Artificial Intelligence Models for Generating Synthetic Formation Bulk Density Logs While Drilling,” *Sustainability*, 12(2), p. 686.
- Hammah, R.; Curran, J.; Yacoub, T., 2006. The influence of Young’s modulus on stress modelling results. In: *Golden Rocks 2006*, 41st U.S. Symp. Rock Mech.
- <https://scikit-learn.org/>.
- Iferobia C.C. và Ahmad M., 2020. A review on the experimental techniques and applications in the geomechanical evaluation of shale gas reservoirs. *Journal of Natural Gas Science and Engineering*, 74, 103090.
- Jiawei H., Micheline K., và Jian P., 2012. *Data Mining: Concepts and Techniques*, Morgan Kaufmann.
- Koen B., 2016. *Classifying Landsat Terrain Images via Random Forests*, Bachelor thesis Computer Science, Radboud University, Netherlands.
- Müller A., Wapler M.C., và Wallrabe U., 2019. A quick and accurate method to determine the Poisson’s ratio and the coefficient of thermal expansion of PDMS. *Soft Matter*, 15(4), 779–784.
- P.V.Hieu, 2021, PhD Thesis. РАЗРАБОТКА ТЕХНОЛОГИЧЕСКИХ РЕШЕНИЙ ПО БУРЕНИЮ СКВАЖИН В НЕУСТОЙЧИВЫХ ОТЛОЖЕНИЯХ РИФТОВОЙ ЗОНЫ МОРСКИХ МЕСТОРОЖДЕНИЙ. Gubkin Russian State University of Oil and Gas
- Siddig O., Gamal H., Elkatatny S. et al., 2021. Real-time prediction of Poisson’s ratio from drilling parameters using machine learning tools. *Sci Rep*, 11(1), 12611.
- Nes, O.-M.; Fjær, E.; Tronvoll, J.; Kristiansen, T.G.; Horsrud, P., 2005. Drilling time reduction through an integrated rock mechanics analysis. In: *SPE/IADC Drill. Conf.* .<https://doi.org/10.2118/92531-MS>
- Pal N.R. và Jain L., 2005. *Advanced Techniques in Knowledge Discovery and Data Mining*, Springer London, London.
- Pedregosa F., Varoquaux G., Gramfort A. et al., 2011. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*. *Journal of Machine Learning Research*.
- Sohil F., Sohali M.U., và Shabbir J., 2022. *An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani*, New York, Springer Science and Business Media, 2013, \$41.98, eISBN: 978-1-4614-7137-7. *Statistical Theory and Related Fields*, 6(1), 87–87.
- Scornet E., Biau G., Vert J.-P., 2015. Consistency of random forests. *Ann Statist*, 43(4).
- Tripathy S.S., 2013. Comparison of Statistical Methods for Outlier Detection in Proficiency Testing Data on Analysis of Lead in Aqueous Solution. *AJTAS*, 2(6), 233.

