Journal of Chemical and Petroleum Engineering (JChPE)

Print ISSN: 2423-673X Online ISSN: 2423-6721

A New Approach in Predicting the Higher Heating Value of Natural Gas from Ghana's Oil Fields

Eric Broni-Bediako *, **Sampson Oware**, **Solomon Asante-Okyere**

1. Petroleum and Natural Gas Engineering, School of Petroleum Studies, University of Mines and Technology, Tarkwa, Ghana. E-mail: ebroni-bediako@umat.edu.gh
2. Ghana Gas Company Limited, Ghana. E-mail: owaresampson@gmail.com
3. Petroleum and Natural Gas Engineering, School of Petroleum Studies, University of Mines and Technology, Tarkwa, Ghana. E-mail: sasante-okyere@umat.edu.gh

ARTICLE INFO

Article History:
Received: 25 July 2023
Revised: 29 August 2023
Accepted: 30 August 2023

Article type: Research

Keywords:
Adaptive Boost,
Artificial Neural Networks,
Extreme Gradient Boost,
Higher Heating Value,
Linear Regression

ABSTRACT

The heating value of natural gas is used to determine the quality of the gas sample, hence accurate prediction of heating value helps in controlling the issue of underbilling and overbilling between a gas aggregator and an off-taker. Moreover, the heating value of natural gas is not a fixed value and the accuracy of it in real-time is essential. This study was focused on the prediction of the Higher Heating Value (HHV) of natural gas based on percentage gas compositions obtained from Ghana's offshore oil fields using Artificial Neural Networks (ANN), Adaptive Boost (AdaBoost), Extreme Gradient Boost (XGBoost), Linear Regression (LR). These algorithms were modelled to determine the best predictive model using 2021 sample data on gas specifications. Eighty percent (80%) of the data was used for training and the remaining 20% was used for testing. The performance of each algorithm was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R^2 and Adjusted R^2. XGBoost performed better than all the other predictive models with an R^2 and adjusted R^2 of 91.18% and 90.93% respectively and RMSE, MAE, and MAPE of 1.7302, 0.5393 and 0.57% respectively. The incorporation of this method provides a diverse approach to the analysis of the pipeline dynamic results of the heating value of natural gas.

Introduction

Natural gas is a multi-substance fossil energy formed underneath the earth's surface [1]. Methane (CH4), the largest component of natural gas constitutes a carbon (C) atom and four hydrogen (H) molecules [2]. Natural gas is colourless, odourless, and amorphous and gives off a valuable amount of energy when it undergoes complete combustion. The combustion of fossil fuels like coal or oil emits large quantities of harmful compounds like nitrous oxide, carbon dioxide and sulphur oxide. Comparatively, during the combustion of natural gas, the emission of sulphur oxide is negligible and that of nitrous oxide and carbon dioxide is lower, which helps reduce the problem of acid rain and greenhouse effects [3]. The world's shift in energy preference from fossil fuels to natural gas is because natural gas serves as a cleaner source of energy [4]. With reference to the BP statistical review of the World Energy 2022 edition, global

* Corresponding Author: E. Broni-Bediako (E-mail address ebroni-bediako@umat.edu.gh)

Journal of Chemical and Petroleum Engineering, 2023, 57(2): 321- 341.
Publisher: University of Tehran, College of Engineering *DOI: 10.22059/jchpe.2023.362770.1444*

natural gas demand grew 5.3% in 2021, recovering above pre-pandemic 2019 levels and crossing the 4 trillion cubic meter mark for the first time. Its share in primary energy in 2021 was unchanged from the previous year at 24% [5]. This reveals the increasing demand for natural gas as a source of fuel over the years. Natural gas has evolved from being primarily used as local energy for heat and electricity to a more robust use in residential, industrial, and commercial heating globally dominating the world economic growth [6]. In the petrochemical industry, natural gas is used as natural industry fuel and feedstock for organic chemical industry processes in the production of ethylene and propylene [7]. Natural gas is also used in the fertiliser industry to produce ammonia. Gases such as hydrogen, sulphur, syngas, and carbon black can also be produced using natural gas [3, 7].

The composition of a commingled gas and the heating value of natural gas are relevant in determining the quality of a natural gas substance. The heating value of natural gas is determined by standard laboratory measurement using a bomb calorimeter based on the mass rather than the volume burned [8, 9]. The heating value of natural gas is essential because it presents the content of energy of gas and the quality of gas [10, 11, 12]. The heating value of natural gas also plays an important role in the gas industry as it is used for billing purposes. Heating value determination is subject to the estimation of gas composition by Gas Chromatograph (GC) which is the most widely used instrument among instruments such as moisture analysers, gravitometers, and hydrogen sulphur monitors [13].

In Ghana, natural gas is obtained from the Jubilee, TEN and Sankofa Fields. It is predominantly used for domestic power supply for industries, transport, and cooking. This has increased natural gas consumption exponentially over the decades. Over two decades, Ghana’s natural gas consumption has increased by 52.6%. This is the result of an increase in industrial and residential demands for natural gas as their source of energy. To ensure a cleaner and more supply of energy, the country anticipates a shift from more environmentally unfriendly fuels to a relatively cheaper and cleaner natural gas-based fuel for its energy supply [14]. In Ghana, most terminal stations along the gas pipeline network where custody transfer takes place are equipped with an online GC at the end of the pipeline, close to the customer or the off-taker. The GC is incorporated with a flow computer for the estimation of the heating value and energy of the natural gas; therefore an accurate estimation of the heating value solely depends on the proper functioning of the GC. For an online GC, in unusual situations, the GC develops faults due to corona (partial discharge), thermal heating and arching. This results in a wrong GC reading, consequently resulting in a wrong diagnosis of the gas. There are also instances where the auxiliaries of the GC such as the gas carrier leaks or calibration gas get in short at the gas stations accounting for false analysis of the composition of the gas. The uncertainty of the analyses from an online GC is of utmost importance to the companies that use these figures in energy calculations, as that forms the basis for the economical transaction between the seller and buyer [15]. The traditional method for estimating heating values of natural gas as proposed in the ISO 6976:2016 must compensate for pressure and temperature at the reference point and deal with the estimation of the uncertainties associated with the heating value. The estimation of the uncertainties associated with the heating values makes it very time-consuming and laborious and does not usually promise an accurate estimation due to the large set of data to be handled, as such a more time-friendly and less tedious predicting the heating value instead of depending on previous data for billing. This study seeks to propose an alternative approach to predict the heating values of natural gas from different oil/gas fields in Ghana using machine learning models.

Machine learning techniques have been used as alternative methods in predicting the heating values of materials. Literature reveals its efficiency in making an accurate prediction. Many studies have revealed the application of machine learning in predicting Higher Heating Values

(HHV) of materials. Xing et al. [16] used ANN, Support Vector Machines (SVM), and Random Forest Regression (RFR) to predict the HHV of biomass based on their proximate and ultimate analyses. The authors used R2 to compare the accuracy of the models and the RFR algorithm performed better with R2 > 0.94. Taki and Rohani [17] used Radial Bias Function Artificial Neural Network (RBF-ANN), Multilayer Perceptron Artificial Neural Network (MLP-ANN), Support Vector Machine (SVM) and Adaptive Nero-Fuzzy Inference System (ANFIS) to predict the HHV of Municipal Waste (MW) for waste -to- energy evaluation. The authors used six different inputs which were carbon, water, hydrogen, oxygen, nitrogen, sulphur and ash. The results revealed that RBF-ANN can predict the HHV of MSW with higher accuracy than other models. Birgen et al. [18], also used ML based modelling to predict the Lower Heating Value (LHV) of municipal waste. In their work, the Gaussian Processes Regression (GPR) was used. Many other studies on using MLT in predicting Higher Heating Value (HHV) focused on different areas not directly related to predicting the HHV of natural gas, particularly from the offshore fields of Ghana. Thus, the research intends to fill this gap.

Materials and Methods

Data Acquisition

An unpublished secondary data on heating values and other related parameters by the Gas Chromatograph were obtained from Ghana's Offshore Oil Fields through Ghana National Gas Company for the prediction. Standard heating values at reference conditions of 20 °C temperature and 101.325 KPa were as well obtained. All data used were in their standard unit.

Pre-Processing of Data

The data set was explored to check if there were abnormalities within the data. The main objective of the analysis was to provide a statistical description of the data, determine outliers within the data set to check for missing values as well as to provide correlation analysis. For accuracy in results, statistical description, outliers and missing values determination, and correlation analysis were carried out before model development. Table 1 presents the statistical description of the data.

Table 1. Statistical Description of Data

	C1	C2	C3	IC4	NC4	IC5	NC5	C6+	N2	CO2	HHV
count	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021
null count	0	0	0	0	0	0	0	0	0	0	0
mean	88.4586	5.92949	3.12121	0.34455	0.61773	0.12077	0.10087	0.07177	0.42439	0.81156	1122.318
std	0.6599	0.40513	0.21891	0.02107	0.03888	0.01177	0.00997	0.01084	0.01052	0.08032	12.7767
min	86.89	0	2.29	0.28	0.47	0.04	0.03	0.01	0.4	0.43	1014.73
25%	88.05	5.85	3.01	0.33	0.6	0.11	0.1	0.07	0.42	0.8	1119.9
50%	88.47	5.98	3.1	0.34	0.61	0.12	0.1	0.07	0.42	0.82	1122.74
75%	88.65	6.15	3.28	0.35	0.63	0.13	0.11	0.08	0.43	0.85	1127.54
max	91.97	6.75	4.96	0.44	0.78	0.16	0.13	0.11	0.47	1.57	1143.27

Data Processing

Data processing is a very important part of the modelling. This is where the dataset used for the prediction of the heating value of natural gas is being processed to make important conclusions and findings. Data processing was made up of outlier determination and fixing, multicollinearity test, input and output variable selection, data splitting, data normalisation, model development and prediction, evaluation on training and testing dataset, and plotting of predicted heating value and actual heating value.

Outlier Determination and Fixing

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement, or it may indicate the experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analysis. The dataset for this study was a real-time field of daily values recorded by a Gas Chromatograph, however, there were several outliers in the dataset which could have affected the prediction if not fixed. Outliers can skew the results of the model and lead towards wrong interpretations. To identify the outliers in the dataset, visualizations with boxplots, a statistical approach using interquartile ranges, and imputation of the values of the outliers were used. Statistically, since the data obtained from the Gas Chromatograph do not follow a normal distribution, the outliers were estimated using the interquartile ranges instead of Z-scores. Eqs. 1, 2 and 3 were used to estimate the outliers from the dataset using Inter-Quartile Range (IQR) Approach.

$$Q1 - 1.5 \times IQR \tag{1}$$

$$Q3 + 1.5 \times IQR \tag{2}$$

$$IQR = Q3 - Q1 \tag{3}$$

where Q1 is the 25th percentile or lower quartile and Q3 is the 75th percentile or upper quartile and IQR is the interquartile range. Values that fell outside of the range of Eqs. 1 and 2 were considered outliers. This was done to scale the data within a specific range for accurate prediction. Python function that accepts columns from the data frame and produces the outliers using the handy pandas was built. In resolving the outliers in the dataset, an imputation approach was used whereby the mean value of each parameter was determined and used to replace the outlier.

Multicollinearity

Multicollinearity exists in a dataset when there is a high correlation between two or more independent variables in multiple linear regression. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model. A statistical technique called the Variance Inflation Factor (VIF) was used to detect and measure the amount of collinearity in a multiple regression model. A VIF of 1 will mean that the variables are not correlated; a VIF between 1 and 5 shows that variables are moderately correlated, and a VIF between 5 and 10 will mean that variables are high.

Input and Output Variable Selection

This is where the data was segregated into two. One part was the dependent variable or output variable, and the other was the independent variable or predictors. In this study, the predictors used were methane (C_1), Ethane (C_2), Propane (C_3), Isobutane (iC_4), Normal Butane ($n\text{-}C_4$), Isopentanes (iC_5), Normal Pentane (nC_5), Hexane (C_6+), Nitrogen (N_2) and Carbon dioxide (CO_2) and the independent or output variable was Heating Value (HV). In the coding, “Label” was used for the output variable which in our case is the Heating Value and “Features” was used for the predictors.

2.3.4 Data Splitting

The data was split into training and testing. It is very important to train the dataset very well to make an accurate prediction. 80% of the dataset was used for training and the remaining 20% was used for testing. The total dataset used for the work was 2021, out of which 1617 representing 80% were used for training and 404 representing 20% were used for testing. The various algorithms were trained on the training dataset and evaluated on the testing dataset to check accuracy.

2.3.5 Data Normalisation

Data Normalisation was done to scale the dataset between 0 and 1 for easy prediction. This scaling was done so that there would not be many outliers in the dataset. This was done to ensure fairness in the dataset for better prediction. In this study, one out of the many methods was used in normalising the dataset in (0, 1) intervals. This was the MinMax Scaler function in the learns pre-processing library. The MinMax Scaler works very well with data having outliers and preserves the relationship among the original data values. Other methods such as the StandardScaler do not work very well with outliers since outliers influence them when calculating the empirical mean and standard deviation which narrows the range of values. Eq. 4 shows the formula for the MinMax Scaler.

$$Y = \frac{Yactual - Ymin}{Ymax - Ymin} \tag{4}$$

where:

Y= Normalized value for the parameter (C_1, C_2, C_3…), Yactual = Value for individual parameter, Ymin = Minimum Value for the parameter, Ymax = Maximum Value for the parameter.

Model Development

Different algorithms were used in the prediction of the heating value of natural gas using Jupyter Notebook with Python language. A total of 2021 datasets were collected from all three oil fields in Ghana (commingled) daily over a period of five (5) years, out of which 1617 were used to train the various models and 404 were used to test and evaluate the model. The Algorithms used in this study were Linear Regression (LR), Artificial Neural Networks (ANN), Adaptive Boosting (AdaBoost) and Extreme Gradient Boosting (XGBoost).

Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical modelling technique used to establish a linear relationship between a dependent variable and two or more independent variables. The technique involves fitting a line that best captures the relationships between the independent variables and the dependent variable. Fig. 1 shows the linear regression model diagram for this work. An MLR model is given by Eq. 5:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdot\cdot\cdot + \beta_k X_k + \varepsilon \tag{5}$$

where:

The slope of y depends on the y-intercept, that is, when all variables x_1 to x_k are zero, y will be β_0. The regression coefficients β_1 and β_2 represent the change in y because of one-unit changes in x_1 and x_2, β_0 refers to the slope coefficient of all independent variables and ε term describes the random error (residual) in the model. The model coefficients (β_0, β_1, $\beta_k X_k$) are estimated by minimising the sum of squared errors of the regression model using Eq. 6

$$min(\sum_{i=1}^{k}[\mathrm{Y} - (\beta 0 + \beta 1\mathrm{X}1 + \beta 2\mathrm{X}2 + \cdots + \beta k\mathrm{X}k)\ ]2) \quad (6)$$

The resulting model can be used to predict the values of the dependent variable given some independent variables. The MLR model was set up for training by importing the LR model from sklearn.linear package. The training was done using the training dataset (consisting of 10 features with 20 records representing 80% of the dataset) and training labels (HHV). The training was completed in 58.7 milliseconds.

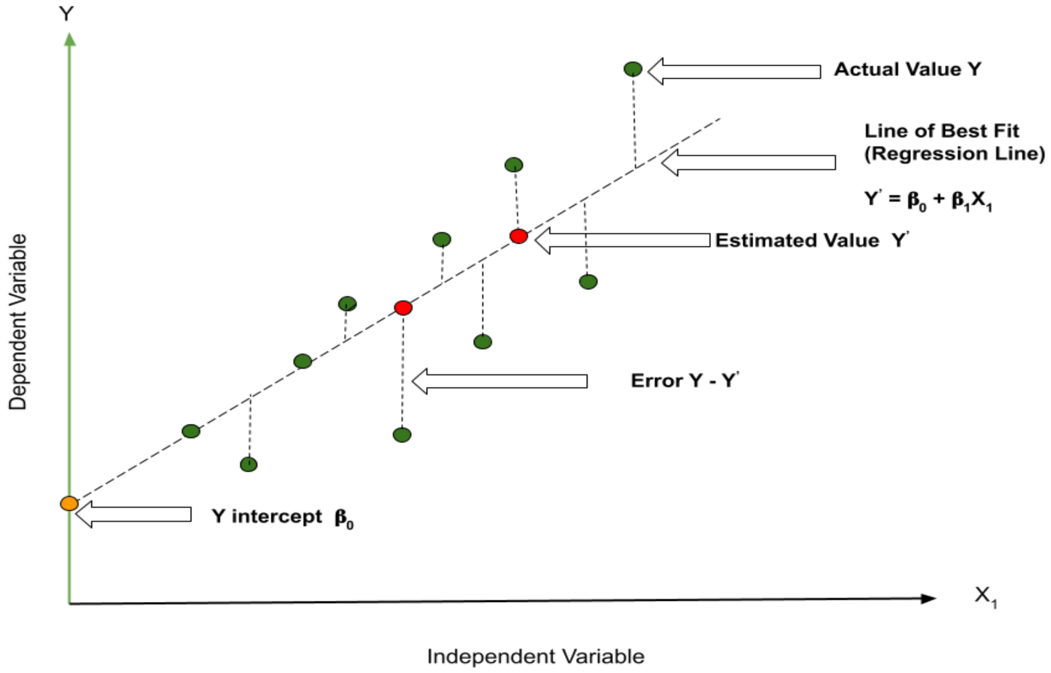


Fig. 1. Diagram of a Linear Regression Model

Artificial Neural Networks (ANN)

Artificial Neural Network model followed a sequential model architecture as in Fig. 2 with an input, hidden and output layer. An artificial neural network (ANN) can be represented mathematically using Eq. 7 for a single neuron or node:

$$y = f(\sum_{i=1}^{n} w_i \,.\, x_i + b) \quad (7)$$

where:

y represents the output of the neuron, f is the activation function applied to the weighted sum of inputs, $\sum_{i=1}^{n} w_i \,.\, x_i + b$ is the weighted sum of the input values $(x_1, x_2, x_3 \ldots x_i)$ multiplied by their corresponding weights $(w_1, w_2, w_3, \ldots w_i)$, b is the bias term added to the weighted sum and n is the number of input connections to the neuron. The activation function introduces non-linearity to the neuron's output, allowing the neural network to learn complex patterns in data. Commonly used activation functions include sigmoid (σ), hyperbolic tangent (tanh), and rectified linear unit (ReLU).

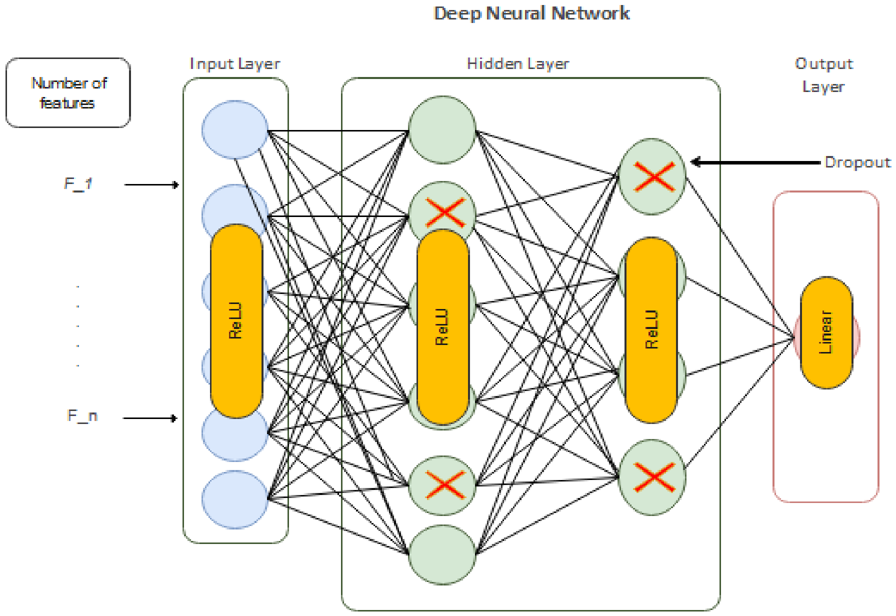
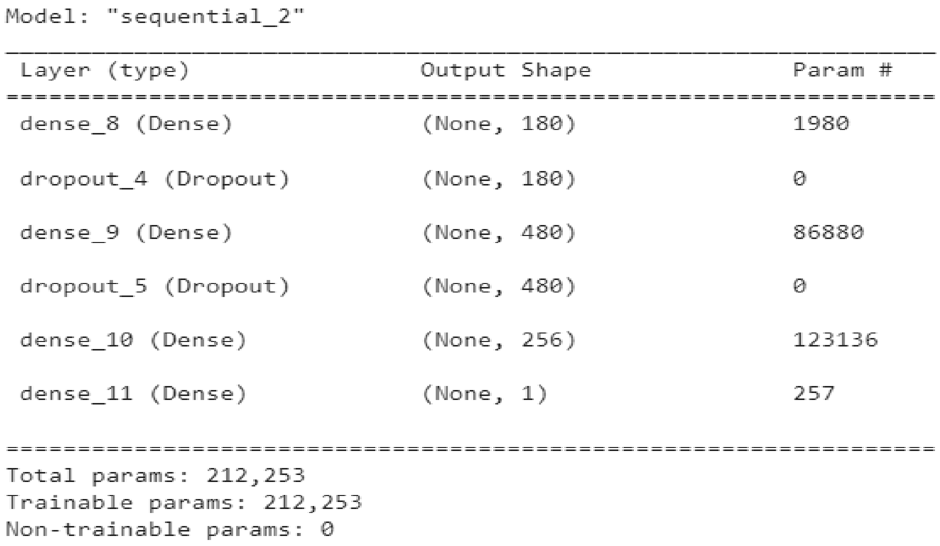


Fig. 2. Diagram of a Deep Neural Network

The model consisted of an input layer with 180 neurons, two hidden layers with 480 and 425 neutrons. A dropout layer with 0.2 dropout was applied between the input layer and hidden layers. The output layer also had a single neuron to output a single prediction. The input and hidden layers used the ReLU activation while the output used a linear activation. Fig. 3 presents the deep learning network model summary.



```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_8 (Dense)             (None, 180)               1980

 dropout_4 (Dropout)         (None, 180)               0

 dense_9 (Dense)             (None, 480)               86880

 dropout_5 (Dropout)         (None, 480)               0

 dense_10 (Dense)            (None, 256)               123136

 dense_11 (Dense)            (None, 1)                 257

=================================================================
Total params: 212,253
Trainable params: 212,253
Non-trainable params: 0
```

Fig. 3. Deep Neural Network Model Summary

The model compilation step is a crucial preparatory step that optimises the network's configuration for efficient training and metrics reporting. This step involves:

i. Choosing an optimizer algorithm, such as SGD, Adam, or RMSprop, to adjust the weights of the network during training for minimising the loss function.
ii. Specifying the Loss Function: Defining the loss function that quantifies the difference between predicted and actual outputs. The choice of loss function depends on the problem type (classification, regression, etc.).

iii. Defining Evaluation Metrics: Setting evaluation metrics like accuracy, precision, or recall to monitor the network's performance during training without directly influencing the training process.
iv. Compiling the Model: Using the compile () method to combine the optimizer, loss function, and evaluation metrics, preparing the model for efficient training.

After compilation, the model was primed for efficient training using the fit () method. This step ensured that the ANN was equipped with the right optimization settings, loss function, and metrics for successful training and convergence, ultimately enhancing the model's predictive capabilities. Table 2 shows the model compilation for ANN. The model was trained on the training set (consisting of 10 features with 1617 records) and training labels (HHV values), for 150 epochs. The training was completed in 42.1 seconds. Fig. 4 presents the model training flowchart for ANN.

Table 2. ANN Model Compilation Configuration

Optimizer	Loss	Metrics	Epochs
Adam	Mean Squared Error	MAE, Root Mean Squared Error	150

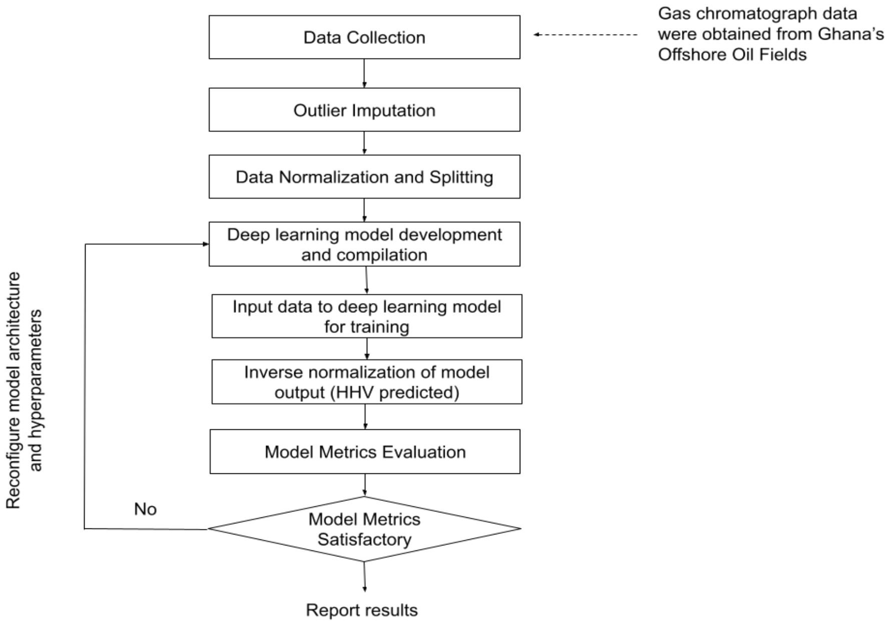

Data Collection
Gas chromatograph data were obtained from Ghana's Offshore Oil Fields
Outlier Imputation
Data Normalization and Splitting
Deep learning model development and compilation
Input data to deep learning model for training
Inverse normalization of model output (HHV predicted)
Model Metrics Evaluation
Model Metrics Satisfactory
No
Reconfigure model architecture and hyperparameters
Report results

Fig. 4. Flowchart of a Deep Neural Network Training Process

Adaptive Boost

Adaptive Boost (AdaBoost) is a popular machine learning algorithm that is used for classification and regression problems. It is an ensemble learning method that combines multiple weak learners (base classifiers) to form a strong classifier. AdaBoost is a boosting algorithm, which means it works by increasing the weight of samples that are misclassified by the previous base classifiers. The model was trained by fitting a Regressor on the provided dataset and then fitting additional copies of the regressor on the same dataset where the weights of instances were adjusted according to the error of the current prediction. As such, subsequent regressors focused more on wrongly predicted data points to make better predictions. Hyperparameter tuning was performed to provide the best model performance. Bayesian Optimization was used to obtain the most optimal hyperparameters, after several different ranges were explored, the most optimal combination was chosen to train the model. Fig. 5

shows the flowchart for the Bayesian optimization. Table 3 presents the values of the optimal hyperparameters for the Adaptive Boosting Model.

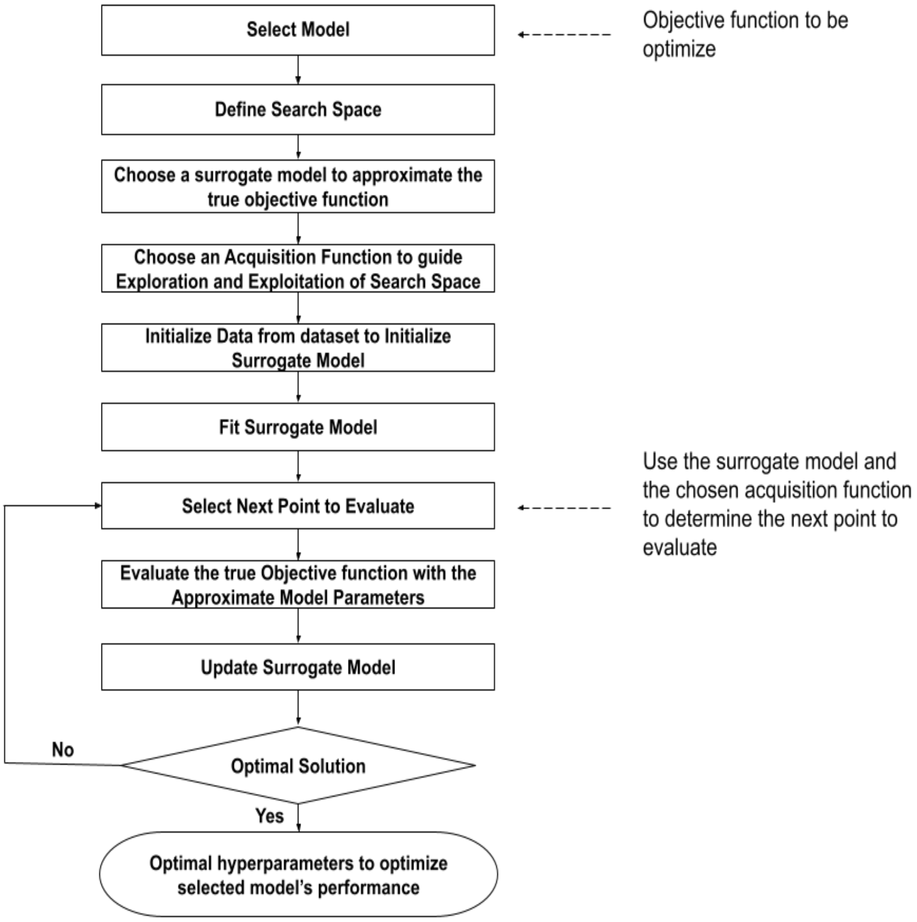

Select Model
Objective function to be optimize
Define Search Space
Choose a surrogate model to approximate the true objective function
Choose an Acquisition Function to guide Exploration and Exploitation of Search Space
Initialize Data from dataset to Initialize Surrogate Model
Fit Surrogate Model
Select Next Point to Evaluate
Use the surrogate model and the chosen acquisition function to determine the next point to evaluate
Evaluate the true Objective function with the Approximate Model Parameters
Update Surrogate Model
No
Optimal Solution
Yes
Optimal hyperparameters to optimize selected model's performance

Fig. 5. Flowchart of Bayesian optimization

Table 3. Optimal hyperparameters for AdaBoost Model

Number of Estimators	Random State	Learning Rate
459	27	0.03

At the initial stage of the algorithm, the weights of the samples were initialised to be equal, which means that each sample had the same importance. The weights of the samples were updated at each iteration of the algorithm. A weak learner was a base classifier that was trained on the current sample weights. After training the weak learner, the sample weights were updated to reflect the performance of the weak learner. The samples that were wrongly predicted by the weak learner were given higher weights, while the samples that are correctly predicted are given lower weights. The weight of the weak learner was calculated based on its accuracy. The weight is given by Eq. 8:

$$\alpha_t = \frac{1}{2} ln\, ln \left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{8}$$

where: t is the iteration number and ϵ_t is the misclassification rate of the weak learner. The final classifier is updated by combining the weighted predictions of the weak learners. The prediction for a sample is given by Eq. 9:

$$f(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t\, h_t(x)\right) \tag{9}$$

where, T is the number of weak learners, $h_t(x)$ is the prediction of the t^{th} weak learner, and sign is the sign function that returns 1 for positive values and -1 for negative values. The algorithm repeats the second and fifth steps until a stopping criterion is met, such as a maximum number of iterations or a minimum accuracy threshold.

Extreme Gradient Boost

Extreme Gradient Boost (XGBoost) is an optimised version of the gradient boosting algorithm. It is a scalable machine learning algorithm that is widely used for both regression and classification problems. XGBoost is a tree-based algorithm that works by constructing an ensemble of decision trees that are trained in a sequential manner to improve predictive performance. Fig. 6 shows the model structure for an XGBoost.

Hyperparameter tuning was performed to provide the best model performance. Bayesian Optimization as shown in Fig. 5 was used to obtain the most optimal hyperparameters, after a number of different ranges were explored, the most optimal combination was chosen to train the model.

Table 4. Optimal Hyperparameters for XGBoost Model

Number of Estimators	**Random State**	**Max Depth**
92	52	52

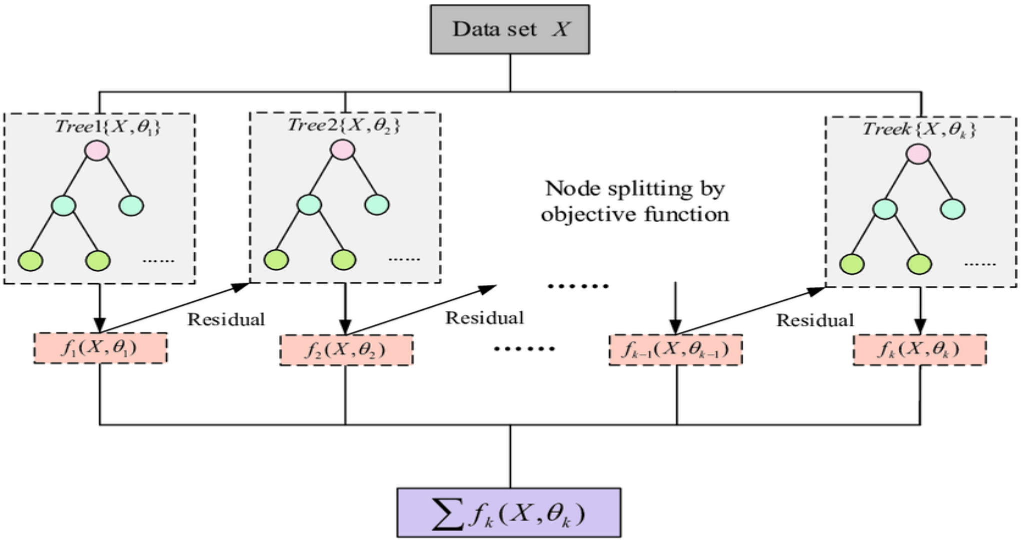


Fig. 6. Diagram of the Model Structure of XGBoost

At the initial stages of the algorithm, the base learners were initialised with a constant value, such as the mean or median of the target variable. A decision tree model was trained on the residuals (difference between the predicted values and the true values) of the previous base learners. The decision tree model was trained to minimise the objective function, which is defined as the sum of the squared residuals. The predictions of the decision tree model were added to the predictions of the previous base learners to form a new set of base learners. The updated base learners formed the new residuals for the next iteration. The weight of the tree model was calculated based on the reduction in the objective function after adding its predictions to the base learners. The weight of the tree model is given by Eq. 10:

$$\eta = learning\ rate \tag{10}$$

$$\text{Objective Function} = \sum_{i=1}^{n} L\big(y_i, \widehat{y_i} + \eta f(x_i)\big)$$

where: L is the loss function, y_i is the true value of the i^{th} sample, $\widehat{y_i}$ is the prediction of the base learners for the i^{th} sample, and $f(x_i)$ is the prediction of the decision tree model for the i^{th} sample. The algorithm repeats the second and fifth steps until a stopping criterion is met, such as a maximum number of iterations or a minimum error threshold.

Evaluation on Training and Testing Datasets

The models were evaluated on the training and testing datasets using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R-Squared (R^2) and Adjusted R-Squared. These performance evaluation tools have been used in earlier studies and can be determined using Eqs. 11-15 [19, 20]. After model development and evaluation, a plot of the predicted heating value and actual heating value was made, and a line of best fit was drawn to determine the equation for prediction.

Root Mean Square Error (RMSE)

The Root Mean Square Error (RSME) measures the standard deviation of the errors. This metric is very important in prediction as it tells how well a regression model can predict the value of a response variable in absolute terms. The predicted heating values from each model were exported to excel to estimate the RMSE for both training and testing datasets using Eq. 11.

$$RSME \sqrt{\frac{1}{n}\sum_{1}^{n}\left(HHV_{act} - HHV_{pre}\right)^2} \tag{11}$$

where n is the number of data samples, HHV_{pre} predicted heating value and HHV_{act} is the actual heating value.

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is the average of the absolute difference between the actual and predicted values in a dataset. The predicted heating values from each model were exported to excel to estimate the Mean Absolute Error for both training and testing datasets using Eq. 12.

$$MAE = \frac{1}{n}\sum_{1}^{n}\left|HHV_{act} - HHV_{pre}\right| \tag{12}$$

where MAE is the mean absolute error, n is the number of data samples, HHV_{pre} predicted value and HHV_{act} is the actual value.

Coefficient of Determination (R^2)

The Coefficient of Determination (R^2) represents the proportion of the variance in the dependent variable (Heating Value) which is explained by the linear regression model. R^2 is always less than 1. The predicted heating values from each model were exported to excel to estimate the R^2 for both training and testing datasets using Eq. 13.

$$R^2 = 1 - \frac{\sum_1^n \left(HHV_{act} - HHV_{pre}\right)^2}{\sum_1^n \left(HHV_{act} - HHV_{avp}\right)^2} \tag{13}$$

where n is the number of data samples, HHV_{pre} predicted heating value, HHV_{act} is the actual value and HHV_{avp} is the average of the predicted heating value.

Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error is the measure of the prediction accuracy of a forecasting method in regression analysis. MAPE is always in percentage (%). The predicted heating values from each model were exported to excel to estimate the MAPE for both training and testing datasets using Eq. 14.

$$MAPE = \frac{\frac{100}{n} \sum_1^n \left|HHV_{act} - HHV_{pre}\right|}{HHV_{act}} \tag{14}$$

Adjusted R^2

Adjusted $R^{2,}$ is a modified form of the coefficient of determination (R^2) which is mainly adjusted for the number of independent variables in the model. Adjusted R^2 will always be less than or equal to R^2. In this project, the predicted heating values from each model were exported to excel to estimate the Adjusted R^2 for both training and testing datasets using Eq. 15.

$$Adj\ R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \tag{15}$$

where Adj R^2 is the Adjusted R^2, n is the number of data samples, k is the number of predictors, and R^2 is the sample R^2.

Results and Discussion

Linear Regression

The model's output for Linear Regression is reported in this part as results. Fig. 7 displays the line plot created using linear regression for the actual and anticipated heating levels. From Fig. 1, the model was not able to accurately predict heating values lower than 1 110 BTU/SCF because the data used had values higher than 1 110 BTU/SCF. As a result of this, the error margin for this model increased. Fig. 8 shows the scatter plot for Actual HHV and Predicted HHV in the Linear Regression Model. A line of best fit was drawn with an R^2 value determined. An equation for prediction was also generated from the model. Eq. 16 is the linear equation generated for the linear regression model.

$$\begin{aligned} HHV = {} & 1116.7221 - 2.8079 \times C1 - 1.4472 \times C2 + 5.2744 \times C3 \\ & + 1.0681 \times IC4 - 1.1064 \times NC4 + 4.2370 \times IC5 + 0.2595 \times NC5 \\ & + 2.8699 \times C6 + + 0.5632 \times N2 - 0.6045 \times C02 \end{aligned} \tag{16}$$

where (CO_2, N_2, C1 … C6) is the composition of the gas sample.

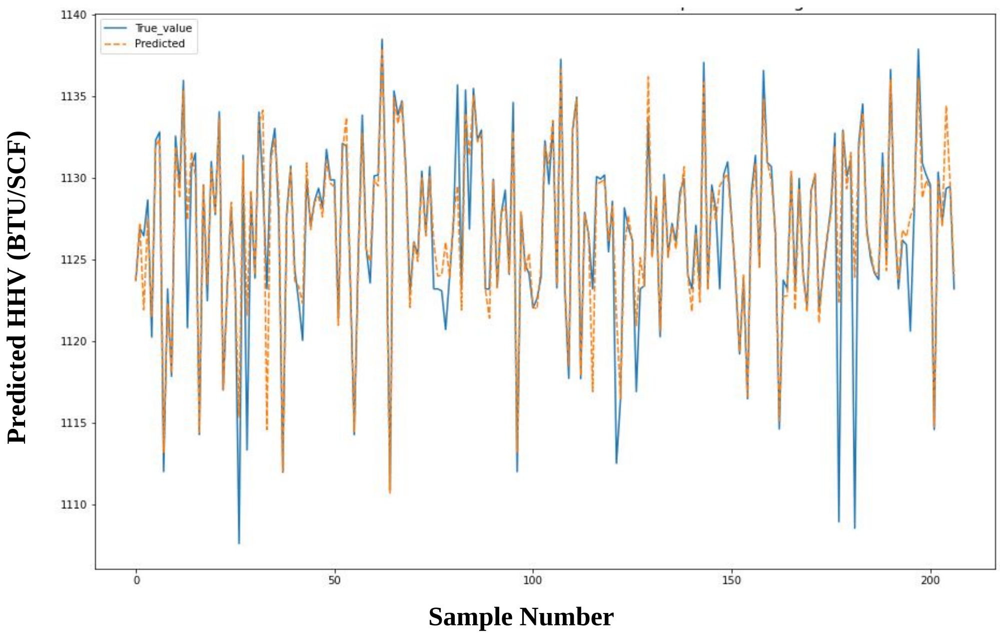


Fig. 7. Line Plot for Actual and Predicted HHV in Multiple Linear Regression

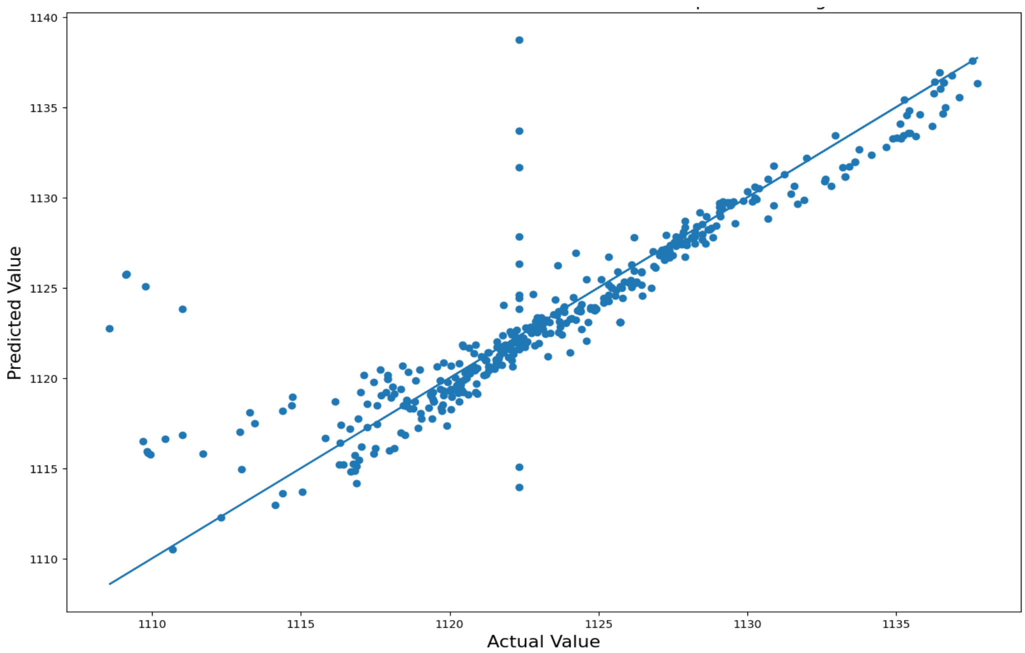


Fig. 8. Scatter Plot for Actual HHV and Predicted HHV in Linear Regression

From Fig. 8, it is seen that most of the values were scattered, and this led to a large difference between the predicted HHV and the actual HHV. Table 5 shows the metric values obtained for Linear Regression Model for both the training and testing dataset.

Table 5. Training and Testing Results for Linear Regression Model

	RMSE	**MSE**	**MAE**	**Adjusted R^2**	**MAPE**	**R^2**
Linear Regression Model			**Training**			
	2.0116	4.0466	1.1343	0.8641	0.53%	0.8650
			Testing			
	2.5343	6.4224	1.2971	0.8055	0.55%	0.8108

From Table 5, the errors for the training dataset were lower than that of the testing dataset. In the training, An R^2 of 86.50% was recorded which shows that the predictor variables were able to explain 86.50% of the variations in the output variable (heating value), whereas in the testing the value of R^2 decreased to 81.08% which shows that the model developed can only explain about 81.08% of the output variable.

AdaBoost Regression

The results obtained from the AdaBoost Regression model are presented in this section. Fig. 9 shows the line plot obtained for the actual heating value and predicted heating value using AdaBoost Regression Model.

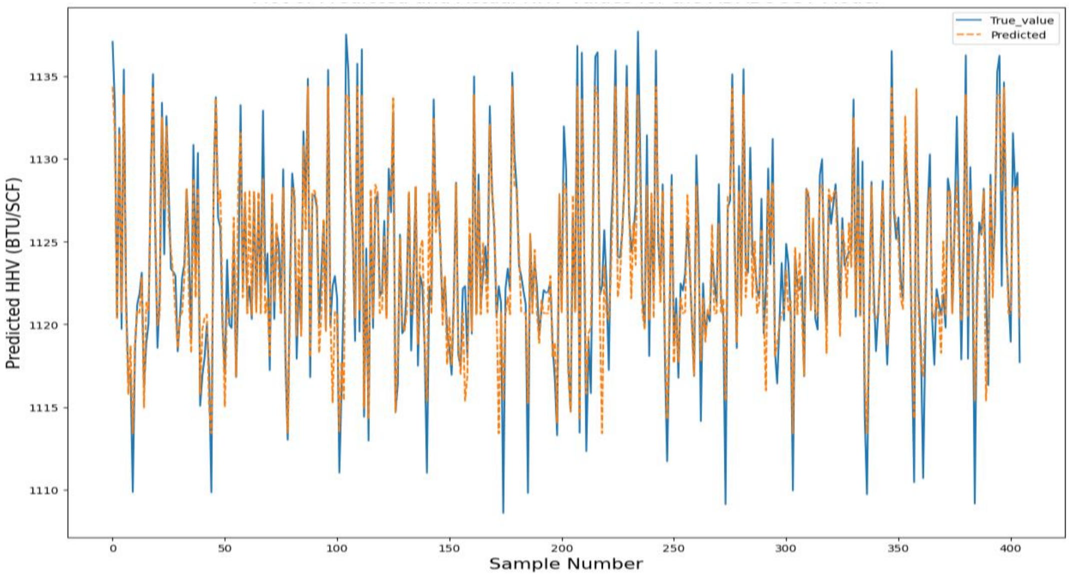


Fig. 9. Line Plot for Actual and Predicted HHV in AdaBoost Regression Model

From Fig. 9, the predicted values and the actual values were not very close together and this resulted in a high error margin between the predicted HHV and the actual HHV. Fig. 10 shows the scatter plot for Actual HHV and Predicted HHV in the AdaBoost Regression Model. A line of best fit was drawn with an R^2 value determined. From Fig. 10, it is evident that most of the points are away from the line of best fit hence there is a high error since the predicted HHV and the actual HHV are not close to each other. Table 6 shows the metric values obtained for the AdaBoost Regression Model for both the training and testing dataset.

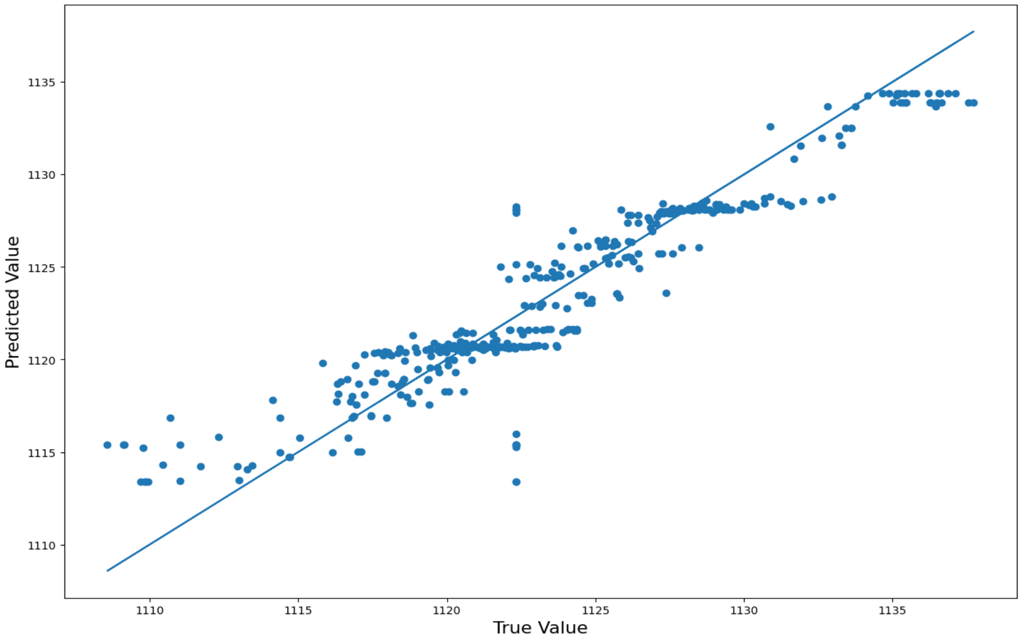


Fig. 10. Scatter Plot for Actual HHV and Predicted HHV in AdaBoost Regression

Table 6. Training and Testing Results for AdaBoost Regression Model

	RMSE	MSE	MAE	Adjusted R2	MAPE	R^2
	Training					
AdaBoost Regression Model	1.9521	3.8105	1.4356	0.8720	0.52%	0.8729
	Testing					
	2.0230	4.0926	1.4559	0.8761	0.55%	0.8794

From Table 6, the errors for the training dataset were lower than that of the testing dataset. In the training, an R^2 of 87.29% was recorded which shows that the predictor variables were able to explain 87.29% of the variations in the output variable (heating value) in the training of the model and this depicts that the model performed very well in the training, whereas in the testing the value of R^2 increased to 87.94% which shows that the model developed can only explain about 87.94% of the output variable which is better than that of the Linear Regression. Fig. 11 shows the feature importance of the AdaBoost Regression Model. It can be seen from Fig. 11 that, C_3 had the highest importance in the prediction of heating value for the AdaBoost Model whereas CO_2 had the least importance. This is because, the presence of CO_2 in a gas mixture reduces the heating value of the gas. The data used had a very low percentage of the CO_2 in the gas stream and this makes it have a low influence on the heating value. Also, C_3 has a very high influence on heating value, this is because the heating value of propane (C_3) is high and having more propane in the gas stream will affect the heating value of the mixture. The gas sample had more propane in the mixture than other heavier fractions (C_3+).

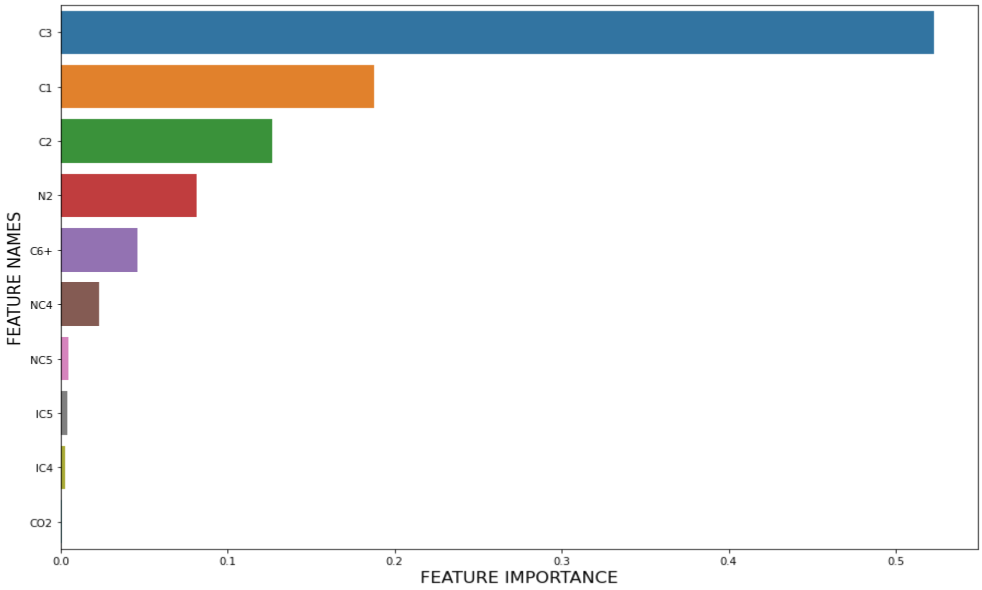


Fig. 11. Feature Importance for the AdaBoost Regression Model

Extreme Gradient Boosting Regressor Model (XGBoost)

The results obtained from the XGBoost model are presented in this section. Fig. 12 shows the line plot obtained for the actual heating value and predicted heating value using XGBoost Regressor Model.

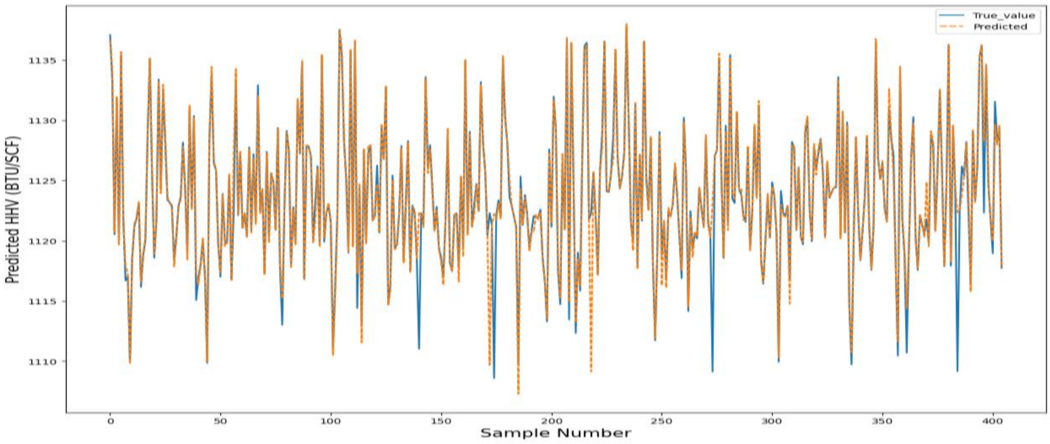


Fig. 12. Line Plot for Actual and Predicted HHV in XGBoost Regressor Model

From Fig. 12, the predicted values and the actual values were a little close together as compared to the AdaBoost Regression and Linear Regression because this model gave out the best (R^2) value amongst all other models and the least error. Fig. 13 shows the scatter plot for Actual HHV and Predicted HHV in the AdaBoost Regression Model. A line of best fit was drawn with an R^2 value determined. From Fig. 13, most points lay on the line of best fit which in turn increased the R^2 for this model and hence made it a better predictor. Table 7 shows the metric values obtained for the XGBoost Regressor Model for both the training and testing dataset.

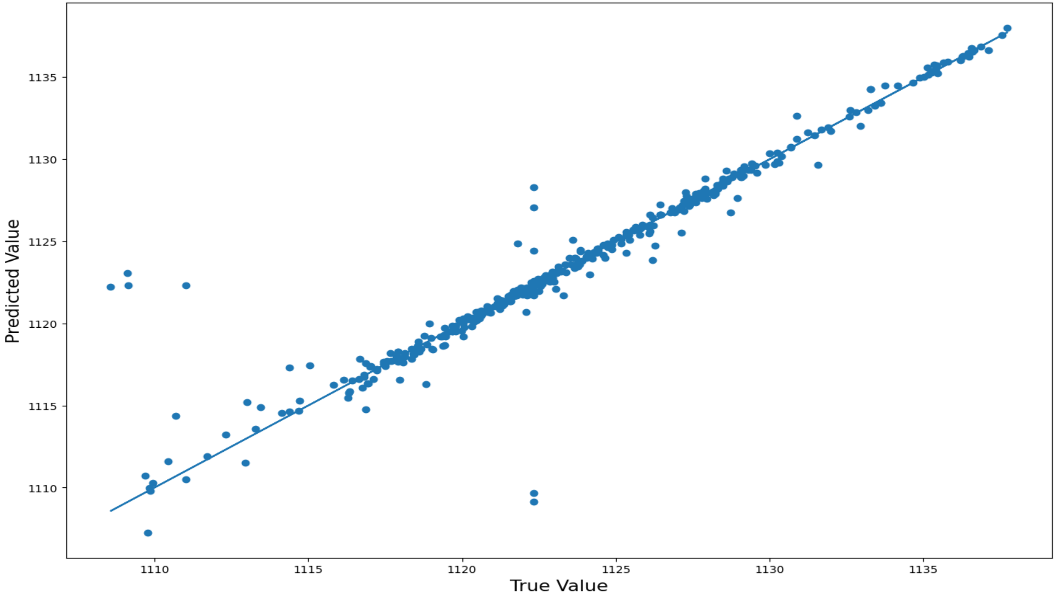


Fig. 13. Scatter Plot for Actual HHV and Predicted HHV in XGBoost Regressor Model

Table 7. Training and Testing Results for XGBoost Regressor Model

	RMSE	MSE	MAE	Adjusted R2	MAPE	R^2
XGBoost Regressor	**Training**					
	0.2761	0.0763	0.0234	0.9974	0.54%	0.9975
	Testing					
	1.7302	2.9934	0.5393	0.9093	0.57%	0.9118

From Table 7, the errors for the training dataset were lower than that of the testing dataset. In the training, an R^2 of 99.75% was recorded which shows that the predictor variables were able to explain 99.75% of the variations in the output variable (heating value) in the training of the model and this means that the model performed very well in the training, whereas in the testing the value of R^2 decreased to 91.18% which shows that the model developed can only explain about 91.18% of the output variable which is better than that of the AdaBoost Regressor Model. Fig. 14 shows the feature importance in the XGBoost Regressor Model. It can be seen from Fig. 14 that, C3 had the highest importance in the prediction of heating value for the XGBoost Model whereas CO2 had the least importance.

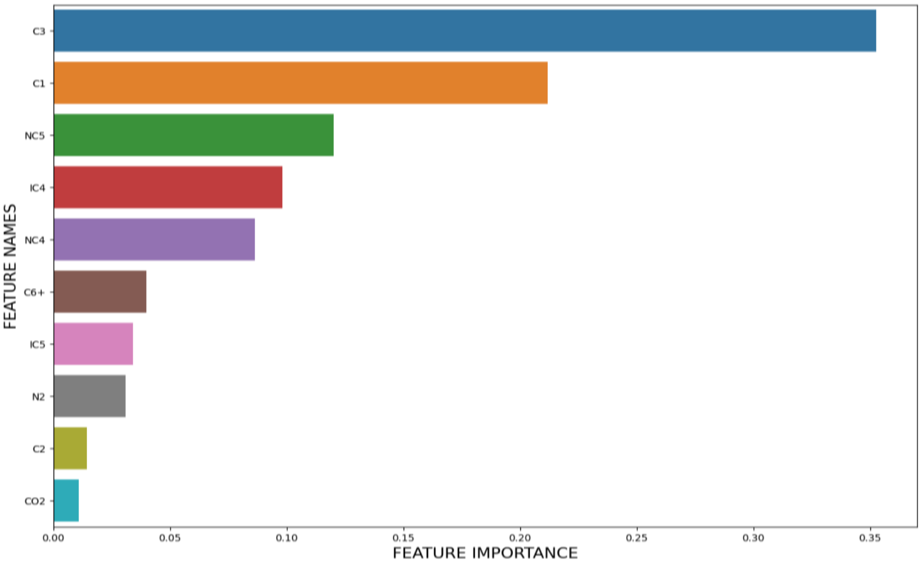


Fig. 14. Feature Importance for the XGBoost Regressor Model

Artificial Neural Networks (ANN)

The results obtained from the Artificial Neural Network (ANN) model are presented in this section. Fig. 15 shows the line plot obtained for the actual and predicted heating values using Artificial Neural Networks.

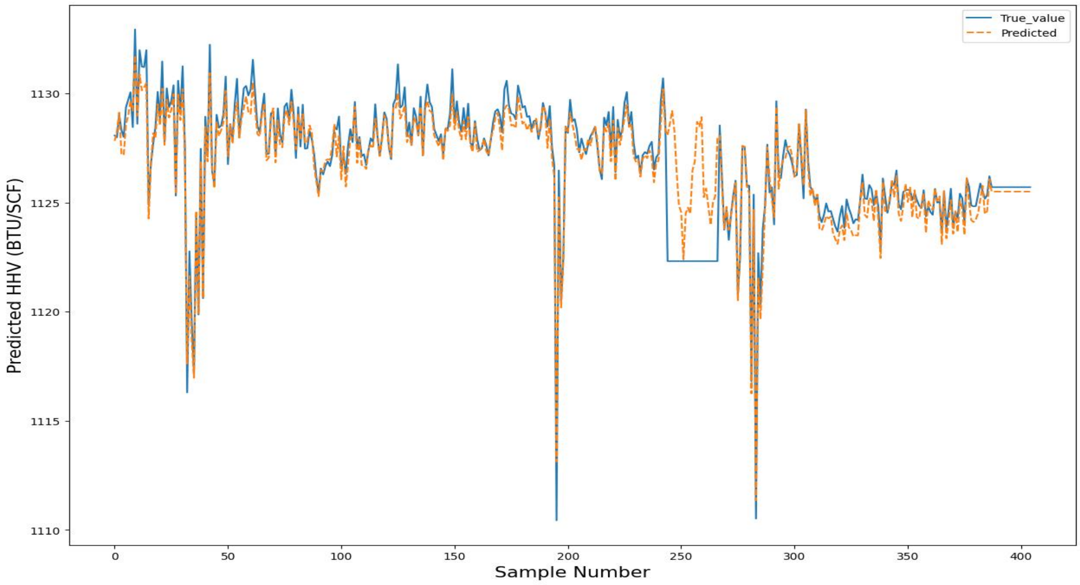


Fig. 15. Line Plot for Actual and Predicted HHV in ANN Model

From Fig. 15, the ANN model gave a moderate prediction for the heating values which were close to the actual heating values. Fig. 16 shows the scatter plot for Actual HHV and Predicted HHV in the Artificial Neural Networks Model. A line of best fit was drawn with an R^2 value determined. From Fig. 16, it is seen that most of the points lie on the line of best fit but not as good as compared to other models. Table 8 shows the metric values obtained for the ANN Model for both the training and testing dataset.

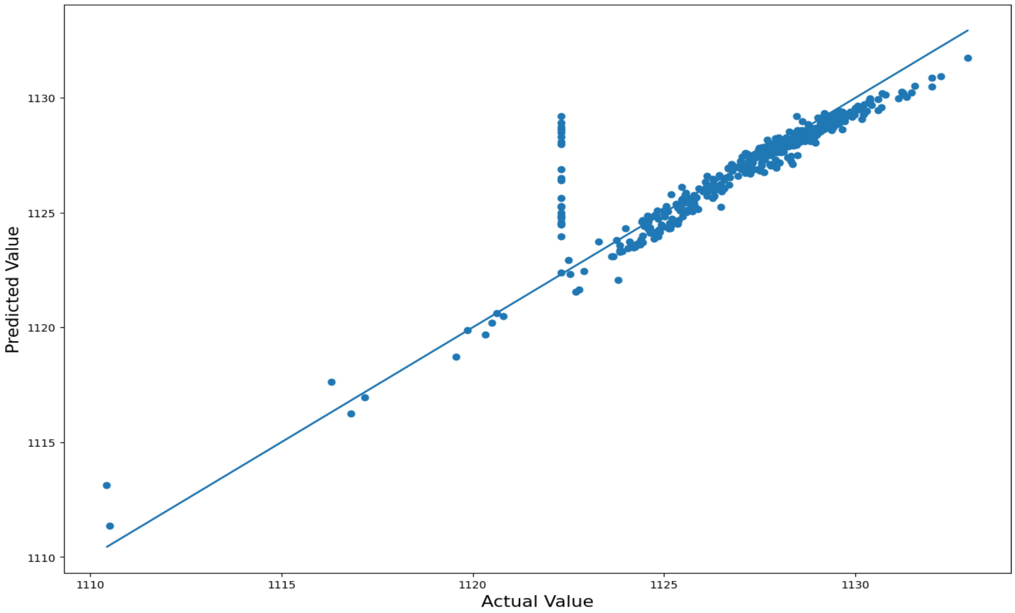


Fig. 16. Scatter Plot for Actual HHV and Predicted HHV in ANN Model

Table 8. Training and Testing Results for ANN Model

	RMSE	MSE	MAE	Adjusted R2	MAPE	R^2
	Training					
ANN Model	0.8366	0.6999	0.3781	0.9789	0.03%	0.9790
	Testing					
	1.1588	1.3425	0.6149	0.8229	0.05%	0.8273

From Table 8, the errors for the training dataset were lower than that of the testing dataset. In the training, an R^2 of 97.90% was recorded which shows that the predictor variables were able to explain 97.90% of the variations in the output variable (heating value) during the training of the model and this means that the model performed very well in the training than in testing, whereas in the testing the value of R^2 decreased to 82.73% which shows that the model developed can only explain about 82.73% of the output variable. Fig. 17 shows the training and validation loss for ANN Model

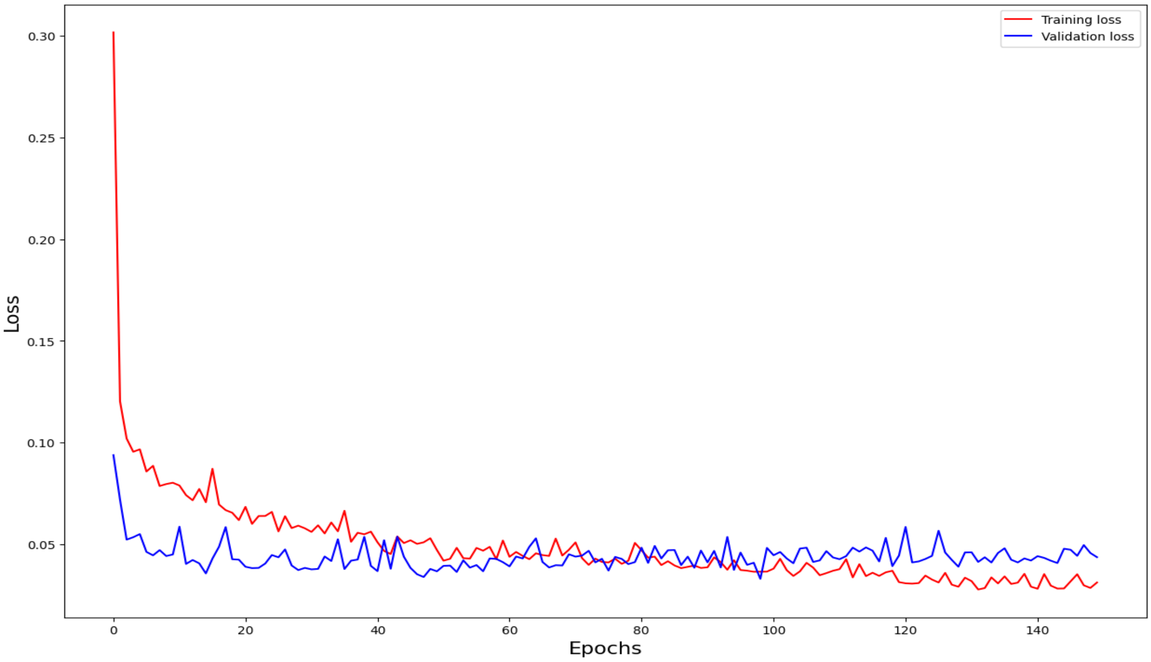


Fig. 17. Training and Validation Loss in ANN Model

Comparison of Models Used

Tables 5, 6, 7 and 8 show the metric values obtained for all four models. Comparatively, all four models performed better during the training than the testing of the data. A model with the least error is preferable since the estimation of heating value plays a vital role in the economics of the industry. For a model, the lower the RMSE, MAE, and MAPE, the higher the accuracy of the model. The model with the lowest RMSE was XGBoost regressor and this gave it a very good R^2 value compared to the other model. The R^2 value simply defines how the independent variables can explain the variations in the dependent variable. XGBoost gave an R^2 of 99.75% and 91.18% respectively for both training and testing, and this simply means that in training, XGBoost was able to explain almost all the variations in the heating value and for testing, 91% of variations were accounted for which is a good score.

Conclusion

An accurate prediction model is needed in Ghana's gas industry to predict the heating value and control issues with overbilling and underbilling between aggregators and off-takers in the event of faulty GC. This paper offers an alternative approach in predicting the heating value of natural gas from Ghana's oil fields using machine learning techniques. After thorough experimentation, it was observed that Artificial Neural Network, AdaBoost Regressor, XGBoost Regressor, and Linear Regression can be used to forecast the heating value with an accuracy (R^2) of 82.73%, 87.94%, 91.18%, and 81.08%, respectively. In the analysis, XGBoost Regressor performed better with a high accuracy (R^2) of 91.18% and the least error amongst all other models hence would be the best fit at as a machine learning model for evaluating the heating value of a gas stream in the event of the GC failing. The R^2 value simply defines how the independent variables can explain the variations in the dependent variable (HV). The Linear Regression model showed the least results amongst all models however, it can be used in prediction with 81% accuracy. The mathematical formula obtained for linear regression can be used for predicting the heating value of natural gas by accounting for the error. For most models, propane (C_3) played the highest role in predicting the heating value and CO_2 recorded the least contribution. This is because the presence of CO_2 in a gas mixture reduces the heating value of the gas. The data used had a very low percentage of the CO_2 in the gas stream and this made it have a low influence on the heating value. Also, C_3 has a very high influence on heating value, this is because the heating value of propane (C_3) is high and having more propane in the gas stream will affect the heating value of the mixture.

Nomenclature

AdaBoost	Adaptive Boost
ANFIS	Nero-Fuzzy Inference System
ANN	Artificial Neural Network
GC	Gas Chromatograph
GPR	Gaussian Processes Regression
HV	Heating Value
HHV	Higher Heating Value
HHV_{act}	Actual Heating Value
HHV_{pre}	Predicted Heating Value
HHV_{avp}	Average of the Predicted Heating Value.
LR	Linear Regression
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MP-ANN	Multilayer Perceptron Artificial Neural Network
RBF-ANN	Radial Bias Function Artificial Neural Network
RMSE	Root Mean Square Error
SVM	Support Vector Machines
XGBoost	Extreme Gradient Boosting Regressor Model
VIF	Variance Inflation Factor

References

[1] Holloway S. Storage of fossil fuel-derived carbon dioxide beneath the surface of the earth. Annual review of energy and the environment. 2001; 26: 145-166. https://doi.org/10.1146/annurev.energy.26.1.145.

[2] Baker RW and Lokhandwala K. Natural gas processing with membranes: An overview. Ind. Eng. Chem. Res. 2008; 47(7):2109-2121. https://doi.org/10.1021/ie071083w

[3] Faramawy S, Zaki T and Sakr, AE. Natural gas origin, composition, and processing: A review.

Journal of Natural Gas Science and Engineering. 2016; 34:34-54. https://doi.org/10.1016/j.jngse.2016.06.030.

[4] Perera F. Pollution from Fossil-fuel Combustion is the Leading Environmental Threat to Global Pediatric Health and Equity: Solutions Exist. Int. J. Environ. Res. Public Health. 2018; 15(1):16. https://doi.org/10.3390/ijerph15010016.

[5] Dale S. BP Statistical Review of World Energy. United Kingdom: British Petroleum Company; 2022.

[6] Mokhatab S, Poe, WA and Speight J. Natural gas compression. Handbook of natural gas transmission and processing. Burlington: Gulf Professional Pub.; 2006.

[7] Siirola, JJ. Natural gas as a chemical industry fuel and feedstock: past, present, future (and Far Future). Eastman Chemical Company. Retrieved from http://egon.cheme.cmu.edu/esi/docs/pdf/SiirolaNaturalGas.pdf.

[8] Veza I, Irianto I, Panchal H, Paristiawan PA, Idris M, Fattah IMR, Putra NR, Silambarasan R. Improved prediction accuracy of biomass heating value using proximate analysis with various ANN training algorithms. Results in Engineering. 2022; 16:1-6. https://doi.org/10.1016/j.rineng.2022.100688.

[9] Sheng C and Azevedo JLT. Estimating the higher heating value of biomass fuels from basic analysis data. Biomass Bioenergy, 2005; 28 (5):499-404. https://doi.org/10.1016/j.biombioe.2004.11.008.

[10] Elmaz1 F, Yücel, Ö and Mutlu, AY. Machine learning based approach for predicting of higher heating values of solid fuels using proximity and ultimate analysis. Int. J. Adv. Eng. Pure Sci. 2020; 32(2): 145-151. https://doi.org/10.7240/jeps.558378.

[11] Afolabi IC, Epelle IE, Gunes B, Okolie JA. Data-driven machine learning approach for predicting the higher heating value of different biomass classes. Clean Technologies. 2022: 1227-1241. https://doi.org/10.3390/cleantechnol4040075.

[12] Yu W and Chen, C. Predicting the heating value of rice husk with neural network. Advances in Intelligent Systems and Computing. 2014; 279. https://doi.org/10.1007/978-3-642-54927-4_84.

[13] Ewing L. Fundamentals of Gas Chromatography, Gas Quality and Troubleshooting. Indiana Avenue, Oklahoma: Chandler Engineering Company LLC; 2001.

[14] Ayaburi J and Bazilian M. Economic benefits of natural gas production: The case of Ghana’s Sankofa Gas Project. Energy for Growth Hub. 2020:1-2.

[15] Steinar F, Reidar S and Victor H. Online Gas Chromatograph: A Technical and Historical Overview – Design and Maintenance Advice to Achieve an accurate End Results. 28th International North Sea Flow Measurement Workshop. 2010:1-15.

[16] Xing J, Luo K, Wang H, Gao Z, Fan J. A comprehensive study on estimating higher heating value of biomass from proximate and ultimate analysis with machine learning Approaches. Energy. 2019; 188: 116077. https://doi.org/10.1016/j.energy.2019.116077

[17] Taki M and Rohani, A. Machine learning models for prediction of the Higher Heating Value (HHV) of Municipal Solid Waste (MSW) for waste-to-energy Evaluation, Case Stud. Therm. Eng. 2022; 31: 101823. https://doi.org/10.1016/j.csite.2022.101823.

[18] Birgen C, Magnanelli E, Carlsson P, Skreiberg O, Mosby J and Becidan, M. Machine learning based Modelling for Lower heating value prediction of municipal solid waste. Fuel. 2021; 283:1-8. https://doi.org/10.1016/j.fuel.2020.118906

[19] Taki M and H. Farhadi H. Application of Artificial Neural Network Models (MLP and RBF) and Support Vector Machine (SVM) to Estimate the Shadow in Flat-plate Solar collectors in Iran, Iran Biosyst. Eng. 2021; 52 (2):197–209.

[20] Qian X, Lee S, Soto A and Chen G. Regression Model to Predict the Higher Heating Value of Poultry Waste from Proximate Analysis. Resources. 2018; 7(3):1-14. https://doi.org/10.3390/resources7030039.

How to cite: Broni-Bediako E, Oware S, Asante-Okyere S. A New Approach in Predicting the Higher Heating Value of Natural Gas from Ghana’s Oil Fields. Journal of Chemical and Petroleum Engineering. 2023; 57(2): 321-341.