Predicting Natural Gas Heating Value using Supervised Machine Learning Models*

S. A. Marfo[1], B. Kofie[1], W. A. Owusu[1] and C. B. Bavoh[1]
[1]University of Mines and Technology, P.O Box 237, Tarkwa, Ghana

Marfo, S. A., Kofie, B., Owusu, W. A., and Bavoh, C. B. (2024), "Predicting Natural Gas Heating Value using Supervised Machine Learning Models", *Ghana Mining Journal,* Vol. 24, No. 1, pp. 188-194.

Abstract

Heating value (HHV) is an essential parameter for evaluating natural gas quality. The existence of supervised machine learning models is therefore necessary for HHV prediction to ensure the eco-friendly and efficient utilisation of natural gas. This study aims to develop machine learning models based on the gas composition to accurately predict the HHV of natural gas. Three predictive models namely; decision tree, AdaBoost, and XGBoost models were used in the evaluation. Data samples from Jubilee, TEN, and SGN fields in Ghana were used. The study considered 721 data sets and the performance of each model was evaluated using R^2, RMSE, and MAE. Results obtained highlighted XGBoost model performs better than the other models. This was backed with an R^2 value of 92.9 % and RMSE and MAE error values of 2.002 and 1.195 respectively.

Keywords: XGBoost; Decision tree; AdaBoost Machine learning; Natural gas.

1 Introduction

The heating value of natural gas plays a crucial role in determining its quality and suitability for various applications (Kale, 2022; Kale *et al*., 2022; Nieto *et al*., 2019; Samadi *et al*., 2021). Globally, this source of energy from natural gas is extensively used for power generation and heating purposes in homes and several industrial applications. However, prior to the use of natural gas, its heating value must be well estimated to ensure the amount of heat needed for the specific intended application (Afolabi *et al*., 2022). To determine the HHV of natural gas, three methods are used; direct, indirect, and inferential.

The direct method involves the combustion and measurement of energy released from gas samples in a bomb calorimeter (Ulbig and Hoburg, 2002). The measurements based on composition and quantity of oxygen utilised during combustion are the inferential and indirect methods respectively. Distinctly, these methods are reported to be challenged with issues of incomplete combustion as a result of catalyst poisoning, cooling effects, and apparatus imprecision, cost-intensive, and time-consuming (Afolabi *et al*., 2022). The efficient utilisation of natural gas primarily depends on the precise prediction of its heating value (Afolabi *et al*., 2022).

In recent years, the use of supervised machine learning (ML) models for predicting the heating value of solid, liquid, and gaseous fuels has gained significant attention, owing to their ability to identify data patterns and provide accurate predictions. In a study by Açıkkar and Sivrikaya (2018), the heating value of coal was predicted based on the proximate analysis using an artificial neural network (ANN) model, and the generalisation of the model was found to be excellent. Also, the use of ANN outperforms multiple regression analysis models in the prediction of natural gas heating value. Similarly, Büyükkanber *et al*. (2023) predicted the heating value of coal using Random Forest (RF) and Artificial Neural Network (ANN) ML models. They confirmed that RF and ANN methods provide a satisfactory prediction of coal heating value. Xing *et al*. (2019) proved that machine learning models such as ANN, SVM, and RF perform better than empirical correlations for the prediction of biomass HHV, with RF showing the best prediction efficiency. Elmaz *et al*. (2020) reported that a polynomial regression-based machine learning algorithm could predict HHV of materials than linear regression, decision tree regression, and support vector regression. Li *et al*. (2021) used artificial neural networks and support vector machines, to develop a hybrid AI model to predict the heating value of natural gas by employing a dataset comprising 619 natural gas samples obtained from China. Propitiously, the heating value of natural gas was predicted using support vector machines, with a dataset of 1800 natural gas samples extracted from the Jilin oilfield. From all the studies (Broni-Bediako *et al*., 2023; Taki and Rohani, 2022; Yaka *et al*., 2022), very few machine learning methods have been applied for the prediction of natural gas heating value.

Considering the existing literature on the use of ML methods for fuel heating value prediction, the results are relatively good, however, there are issues of low model prediction accuracies. Hence, studying the prediction efficiency of different machine learning

*Manuscript received March 02, 2024
Revised version accepted June 05, 2024
https://doi.org/10.4314/gm.v22i1.3

models that are reliable and stable is needed. Also, critical aspects of the existing models' poor performances are related to their inability to handle small data sets. To deal with these challenges, it's best to develop new heating values predictive models based on different ML models that are capable of optimising well with fewer data sets and yielding low overfitting. Therefore, in this present study, the prediction of the heating value of natural gas was modeled by using three machine learning tools namely Adaboost, XGboost, and decision tree models. These machine learning tools were selected for their ability to minimise predictions overfitting and cope with small data sets. The models in this work are very useful to simply predict the heating value of natural gas in the gas industry.

2 Resources and Methods Used

2.1 Data

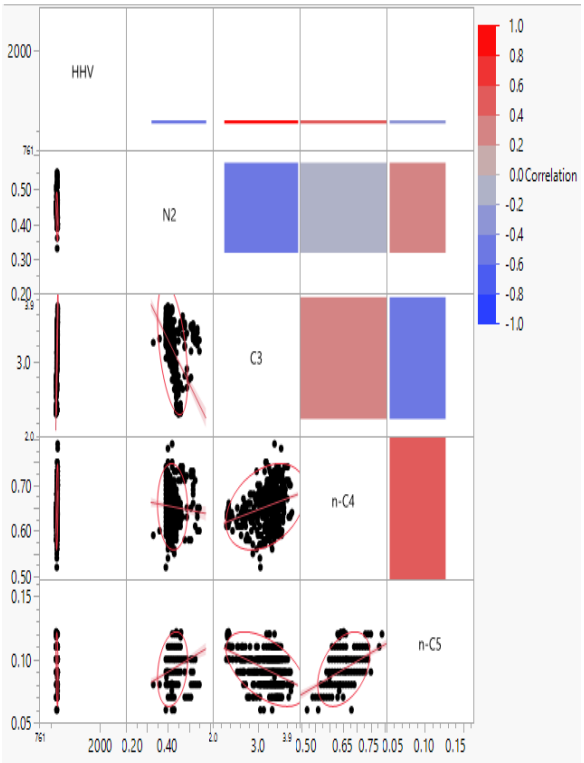

Fig. 1. Distribution and descriptive statistics of the data used for the model development

2.2 Methods

2.2.1 Decision Tree

Decision tree learning is a process of building trees from a dataset for classification (categorical output) or regression (continuous value output) tasks (Ahmadi and Chen, 2019; Öğrenmesi *et al*., 2020). A root node, branches, internal nodes, and leaf nodes make up its hierarchical tree structure. A decision tree starts with a root node that does not have branches. The incoming branches from the root node feed the internal nodes; the decision nodes. Assessment of these nodes is done based on the given attributes to create homogeneous subsets, which are denoted by leaf nodes. Leaf nodes reflect all potential outcomes. The findings of the decision rule to generate the branches or segments beneath the root node are based on a technique that extracts the relationship between the target variable in the data and the input variables.

The comingled gas data used in the study was obtained from Ghana National Gas Company (GNGC), specifically from the Jubilee, TEN, and SGN fields. A total of 750 datasets were used and first subjected to treatment. The data was treated by checking their correlation matrix detection of collinearity and removal of outliers.

The data obtained consisted of the natural gas heating value (HHV), which is the output variable. The natural gas composition was used as the input variable. Three natural gas compositions were selected for the model development due to their minimal contribution to collinearity. The HHV, nitrogen (N_2), propane (C_3), normal butane (n-C_4), and normal pentane (n-C_5) compositions were used to develop the models in this work. Table 1 and Figure 1 show the basic statistics of the data used.

Table 1. Range of Data Used in This Study

Data Type	Mean	Minimum	Maximum
HHV(Btu/scf)	1128.50	1104.12	1142.13
N_2 (%)	0.4219	0.33	0.55
C_3 (%)	3.1554	2.32	3.71
n-C_4 (%)	0.6524	0.52	0.78
n-C_5 (%)	0.0922	0.06	0.1218

2.2.2 AdaBoost

AdaBoost is an ensemble learning machine learning algorithm used for regression, though it also has potential for classification problems *(*Gavrishchaka *et al*., 2018). AdaBoost uses both strong and weak learning methods for effective predictions. It belongs to the boosting algorithms and operates on the principles of increasing the sample weight of previous base classifiers that have poor convergence.

2.2.3 Extreme Gradient Boost (XGBoost)

The XGBoost is an upgraded version of the gradient boosting algorithm. This model is also used for both regression and classification problems and operates based on scalable machine learning algorithm

techniques. The predictive performance of XGBoost is enhanced by the sequential training of an ensemble of decision trees. In this study, Extreme Gradient Boosting (XGBoost) is one of the types of boosting algorithms that were used to predict the heating value of natural gas (Acharya and Bahadur, 2021; Nwachukwu *et al*., 2018).

2.2.4 Model Development

The models considered in this work are the Decision Tree, Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost). After selecting the models developed using Jupyter Notebook software. The model prediction process was initiated by the selection of input and output variables from the collected natural gas data. The HHV was assigned as the output values while N_2, C_3, n-C_4, and nC_5 were used as input values.

The scaling feature used is normalisation which involves the scaling down of dataset values within a fixed range. This feature enhances the fairness of the training by preventing inputs with higher values from kicking out inputs with lower values. The dataset was normalised to (0,1) interval. The equation for normalisation is shown in Equation 1 (Bavoh *et al*., 2023).

$$X = \frac{X_{actual} - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X is the normalised value, X_{actual} is the actual feature value, X_{min} is the minimum value of the feature and X_{mas} is the maximum value feature.

2.2.5 Model Training and Performance Evaluation

In this work, the HHV of natural gas was predicted using three algorithms (AdaBoost, XGBoost, and decision tree models) based on the percentage composition of the gas. The dataset consisted of 750, which was later reduced to 721 after preprocessing and outliers exclusion. A total of 505 (70%) samples were used for training and 216 (30%) samples were reserved for performance testing of all the developed models.

Each algorithm was fitted on the baseline model to perform a repeated evaluation of the training data to fully comprehend its properties, identify relationships within the dataset, and fine-tune itself for good model development. With each algorithm, multiple weak learners were sequentially trained, with each learner focusing on the mistakes made by the previous ones. The performance of each was improved by systematically selecting and tuning their hyperparameters.

The Decision Tress model parameters used for training and predictions in this work consist of binary trees with at least 10 instances in leaves and 5 instances in internal nodes at a maximum depth of 30.

An optimal hyperparameters of 100 trees, a tree depth of 6, and a learning rate of 0.3 were used for the Extreme Gradient Boost model. While 50 estimators, 1.0 learning rate, and a square loss function were used for the training and prediction of the AdaBosst model.

The developed models from the training were examined on the testing data which acts as unseen data to evaluate the generalisation of the developed models' abilities in predicting the heating value of the gas. The evaluation was conducted during the training, and the testing stages for each model were developed using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). The equation used to estimate the RMSE, MAE, and R^2 in this study are shown in Eq.s (2) to (4) (Bavoh *et al*., 2023).

$$RMSE = \sqrt{\frac{1}{n}\sum_1^n \left(HHV_{exp} - HHV_{pre}\right)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_1^n \left|HHV_{exp} - HHV_{pre}\right| \quad (3)$$

$$R^2 = 1 - \frac{\sum_1^n \left(HHV_{exp} - HHV_{pre}\right)^2}{\sum_1^n \left(HHV_{exp} - HHV_{mean}\right)^2} \quad (4)$$

where HHV_{exp}, HHV_{pre}, and HHV_{mean} are the experimental, predicted, and mean natural gas heating values. *n* is the number of data samples.

3 Results and Discussion

3.1 Decision Tree Model

In this study, a decision tree model was developed, and dataset training and testing were evaluated. *RSME*, *MAE*, and R^2 values obtained on training were 2.211, 1.364, and 0.927 respectively. *RSME*, *MAE*, and R^2 evaluation values obtained on testing data were 2.095, 1.516, and 0.922 respectively.

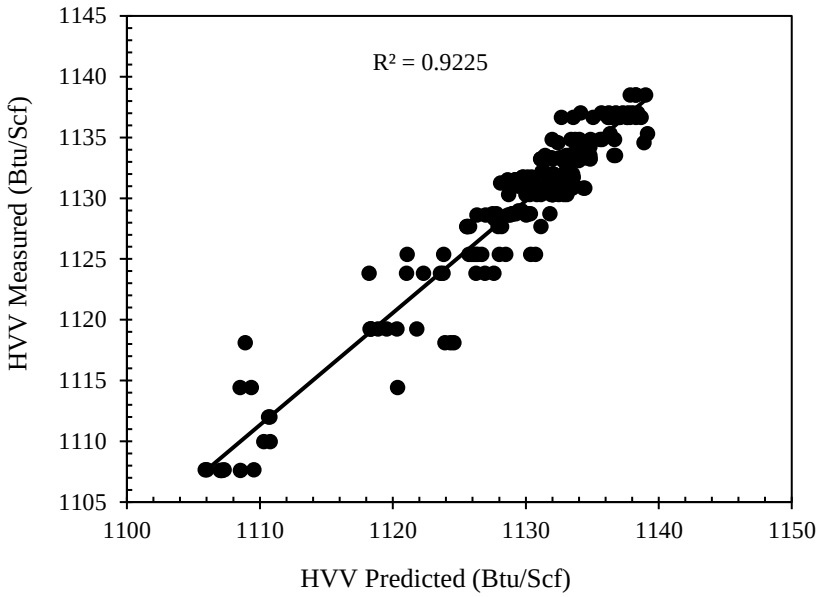

Figure 2. A Plot of experimental and predicted *HHV* values for the Decision Tree Model

The difference in R^2 values of the two datasets indicates the model performance on training to be more effective than testing. Figure 2 shows a line plot of predicted and actual HHV values for the decision tree model. As depicted in the plot, there is a strong correlation between the predicted values generated by the model and the actual HHV values.

3.2 Extreme Gradient Boosting (XGBoost) Model

An XGboost model was developed and evaluated on the training and testing datasets. RSME, MAE, and R^2 values obtained during the training were 1.501, 0.966, and 0.966 respectively. RSME, MAE, and R^2 evaluation values obtained on testing data were 2.002, 1.195, and 0.929 respectively. The R^2 values for the training and the testing do not differ much, indicating the model performs well on both the training and the testing data. Figure 3 shows a line plot of predicted and actual HHV values for the XGBoost model. The majority of predicted HHV values from the XGBoost model closely align with the actual values, indicating a negligible difference in errors between them.

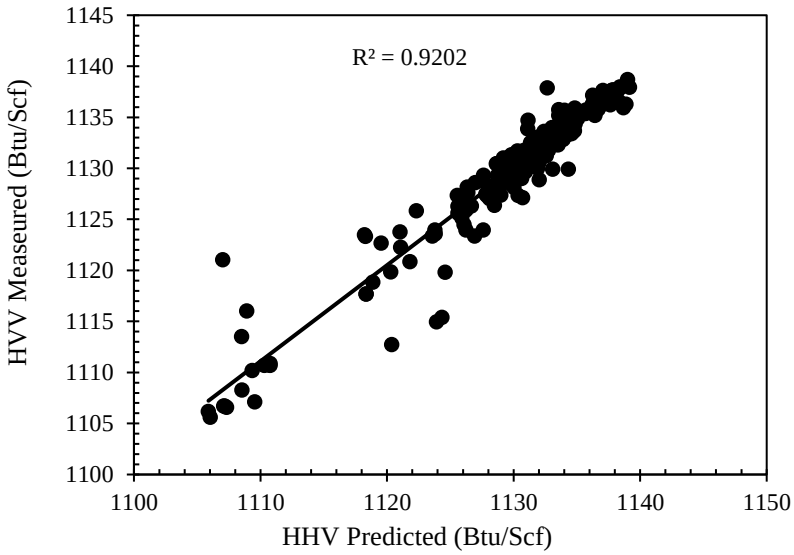

Figure 3. A Plot of experimental and predicted *HHV* values for the XGBoost Model

3.3 Adaptive Boosting (Adaboost) Model

An Adaboost model was developed and training and testing evaluation was performed on datasets. *RSME*, *MAE*, and R^2 values obtained during the training were 1.137, 0.336, and 0.981 respectively. *RSME*, *MAE*, and R^2 evaluation values obtained on testing data were 2.732, 1.404, and 0.867 respectively. The R^2 value is higher for the training dataset than the testing dataset, indicating that the model's performance is better on the training data than on the testing data. Figure 4 shows a line plot of predicted and actual *HHV* values for the AdaBoost model. Nevertheless, most of the values predicted by the AdaBoost model were close to the actual value.

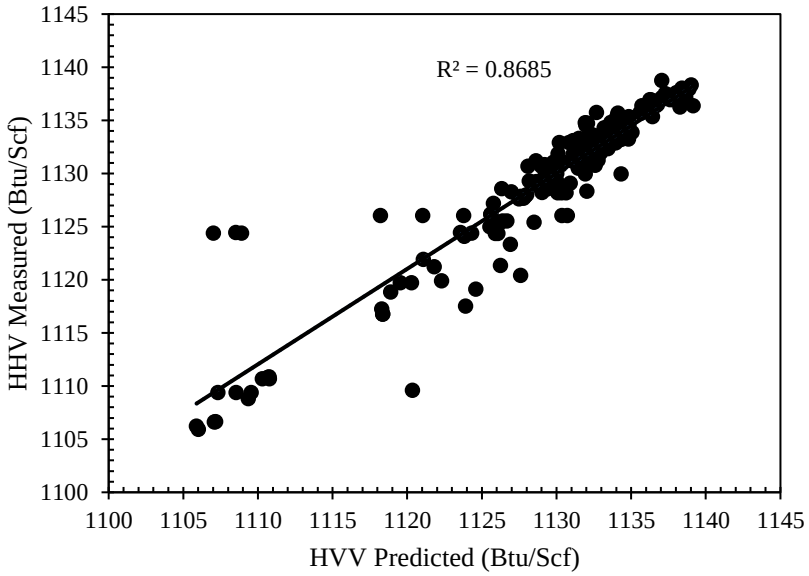

Figure 4. A Plot of experimental and predicted *HHV* values for the AdaBoost Model

Table 2. Comparative results of the developed models

Model	Training			Testing		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Decision Tree	2.211	1.364	0.927	2.095	1.516	0.922
XGBoost	1.501	0.903	0.966	2.002	1.195	0.929
AdaBoost	1.137	0.336	0.981	2.732	1.404	0.867

3.4 Comparison of Developed Models

Table 2 shows the comparative results of the three developed models. It can be established that all three models reasonably predicted the heating value of natural gas. The XGBoost model stands out to be the best-performing model with the highest accuracy or R^2 value of 92.9 % and the lowest RSME and MAE error values of 2.002 and 1.195 respectively on the testing data. The prediction accuracy of the models in this work is comparable to the results reported by Broni-Bediako *et al*. (2023). They proposed that XGBoost outperforms AdaBoost in predicting the heating value of natural gas. In this work, the performance of XGBoost shows higher testing R^2 values and slightly higher RMSE in the study by (Broni-Bediako *et al*., 2023). However, our model uses 4 input variables while theirs uses 10 input variables. This confirms the simplification of our models without compromising the model prediction accuracy.

4 Conclusions

In this study, three models were used to predict the heating values of natural gas from three Ghanaian fields. The results show that the heating value of natural gas can be predicted effectively using supervised machine learning models that is Adaboost, XGboost, and Decision Tree models. For all the developed machine learning methods in this work, the Extreme Gradient Boosting models best predicted the heating value of natural gas with reliable accuracy and RMSE of 95.86% and 1.6572, respectively. This represents about a 200% reduction in prediction error compared with Decision Tree, Adaptive Boosting. The natural gas heating value prediction accuracy for Decision Tree and Adaptive Boosting is found to be similar with R^2 between 81.82% - 83.81%. Thus, natural gas heating value software could adopt the Extreme

Gradient Boosting model for efficient prediction accuracy in academic and industrial applications.

Acknowledgements

The authors thank the University of Mines and Technology, Tarkwa for providing the facilities for this work.

References

Acharya, P. V., and Bahadur, V. (2021), "Thermodynamic features-driven machine learning-based predictions of clathrate hydrate equilibria in the presence of electrolytes", *Fluid Phase Equilibria*, Vol. *530*, No. 2021, pp. 112894.

Açıkkar, M., and Sivrikaya, O. (2018), "Artificial neural networks for estimation of the gross calorific value of Turkish lignite coals", *3rd International Mediterranean Science and Engineering Congress (IMSEC 2018)*, *November*, pp. 1075–1079.

Afolabi, I. C., Epelle, E. I., Gunes, B., Güleç, F., and Okolie, J. A. (2022), "Data-Driven Machine Learning Approach for Predicting the Higher Heating Value of Different Biomass Classes", *Clean Technologies*, Vol. *4*, No. 4, pp. 1227–1241.

Ahmadi, M. A., and Chen, Z. (2019), "Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs", *Petroleum*, Vol. *5*, No. 3, pp. 271–284.

Bavoh, C. B., Sambo, C., Quainoo, A. K., and Lal, B. (2023), "Intelligent prediction of methane hydrate phase boundary conditions in ionic liquids using deep learning algorithms", *Petroleum Science and Technology*, Vol. *0*, No. 0, pp. 1–20.

Broni-Bediako, E., Oware, S., and Asante-Okyere, S. (2023), "A New Approach in Predicting the Higher Heating Value of Natural Gas from Ghana's Oil Fields", *Journal of Chemical and Petroleum Engineering*, Vol. *57*, No. 2, pp. 321–341.

Büyükkanber, K., Haykiri-acma, H., and Yaman, S. (2023), "Calorific value prediction of coal and its optimization by machine learning based on limited samples in a wide range", *Energy*, Vol. *277*, No. December 2022, pp. 127666.

Elmaz, F., Yücel, Ö., And Mutlu, A. Y. (2020), "Machine learning based approach for predicting of higher heating values of solid fuels using proximity and ultimate analysis", *International Journal of Advances in Engineering and Pure Sciences*, Vol. *32*, No. 2, pp. 145–151.

Gavrishchaka, V. V., Yang, Z., (Rebecca) Miao, X., and Senyukova, O. (2018), "Advantages of hybrid deep learning frameworks in applications with limited data", *International Journal of Machine Learning and Computing*, Vol. *8*, No. 6, pp. 549–558.

Kale, S. B. (2022), "Prediction of Gross Calorific Valueof Coal using Machine Learning Algorithm", *International Journal for Research Trends and Innovation*, Vol. *7*, No. 11, pp. 547–554.

Kale, S. B., Shinde, V. A., and Koshti, V. S. (2022), "Prediction of Gross Calorific Value of Coal using Machine Learning Algorithm", *International Journal for Research in Applied Science & Engineering Technology*, Vol. *10*, No. 7, pp. 4960–4973.

Li, J., Guo, Y., Zhang, X., and Fu, Z. (2021), "Using Hybrid Machine Learning Methods to Predict and Improve the Energy Consumption Efficiency in Oil and Gas Fields", *Mobile Information Systems*, Vol. *2021*, pp. 7.

Nieto, P. J. G., Gonzalo, E. G., Lasheras, F. S., Sánchez, J. P. P., and Fernández, P. R. (2019), "Journal of Computational and Applied Forecast of the higher heating value in biomass torrefaction by means of machine learning techniques", *Journal of Computational and Applied Mathematics*, Vol. *357*, pp. 284–301.

Nwachukwu, A., Jeong, H., Pyrcz, M., and Lake, L. W. (2018), "Fast evaluation of well placements in heterogeneous reservoir models using machine learning", *Journal of Petroleum Science and Engineering*, Vol. *163*, pp. 463–475.

Öğrenmesi, M., Kullanarak, A., Yakıtların, K., Isı, Ü., and Yücel, Ö. (2020), "Machine Learning Based Approach for Predicting of Higher Heating Values of Solid Fuels Using Proximity and Ultimate Analysis"Vol. *32*, No. 2, pp. 145–151.

Samadi, S. H., Ghobadian, B., and Nosrati, M. (2021), "Environmental Effects Prediction of higher heating value of biomass materials based on proximate analysis using gradient boosted regression trees method", *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, Vol. *43*, No. 6, pp. 672–681.

Taki, M., and Rohani, A. (2022), "Case Studies in Thermal Engineering Machine learning models for prediction the Higher Heating Value (HHV) of Municipal Solid Waste (MSW) for waste-to-energy evaluation"Vol.

31, No. October 2021, pp. 1–13.
Ulbig, P., and Hoburg, D. (2002), "Determination of the calorific value of natural gas by different methods", *Thermochimica Acta*, Vol. *382*, No. 1–2, pp. 27–35.
Xing, J., Luo, K., Wang, H., Gao, Z., and Fan, J. (2019), "A comprehensive study on estimating higher heating value of biomass from proximate and ultimate analysis with machine learning approaches", *Energy*, Vol. *188*, pp. 116077.
Yaka, H., Akin, M., Yucel, O., and Sadikoglu, H. (2022), "A comparison of machine learning algorithms for estimation of higher heating values of biomass and fossil fuels from ultimate analysis", *Fuel*, Vol. *320*, No. August 2021, pp. 123971.

Authors

Engr Solomon Adjei Marfo is a Senior Lecturer at the Chemical and Petrochemical Engineering Department of University of Mines and Technology, Tarkwa, Ghana. He holds PhD in Petroleum Engineering from the University of Port Harcourt, Nigeria, MEng Degree in Mining (Petroleum Engineering) from the University of Belgrade, Serbia and BSc in Chemical Engineering from the Kwame Nkrumah University of Science and Technology, KNUST, Kumasi, Ghana. He is a member of the Ghana Institution of Engineering and Technology (PE.IET GH), a member of the Society of Petroleum Engineers (SPE), a Registered Environmental Specialist (RES) with the National Registry of Environmental Professionals (NREP) of USA. His research interests include Fuel an Lubricant Quality Analysis, Crude Oil Analysis, Application of Artificial Intelligence in Petroleum, Petrochemical, Refining Engineering.

Bismark Kofie holds a Bachelor of Science in Petroleum Engineering from the University of Mines and Technology in Tarkwa, Ghana. His research interests focus on Fuel Energy content optimisation, Drilling Techniques, Reservoir Characterisation and integration of Supervised Machine Learning in Petroleum.

Winnie Ampomaa Owusu is PhD student in Refining and Petrochemical Engineering at the University of Mines and Technology, Tarkwa, Ghana (UMaT). She holds a BSc in Petroleum Engineering from UMaT. Her research interests include Biofuels, AI in Biofuel Production, Utilisation of Agro-Wastes in Oilfield Applications, Hydraulic fracturing, and Fuel Analysis. She aims to contribute to sustainable energy solutions in the Refining, Petrochemical, and Petroleum Industries through innovative research and technology integration.

Cornelius Borecho Bavoh is a Lecturer at the Chemical and Petrochemical Engineering at the University of Mines and Technology, Tarkwa, Ghana. He holds the degrees of MSc and PhD from Universiti Tecknologi PETRONAS, Malaysia. He is a member of the Society of Petroleum Engineers and the Malaysia Board of Technologist. His research and consultancy work cover projects related to Drilling fliuds, Gas hydrate, and CO_2 capture and separation, Fuel characterisation, etc.