Shahrood University of Technology

Journal of Mining and Environment (JME)

Journal homepage: www.jme.shahroodut.ac.ir

Iranian Society of Mining Engineering (IRSME)

Predictive Modeling of Coal Gross Calorific Value Using Conventional and Robust Machine Learning Regression Techniques

Satyajeet Parida[1], Abhishek Kumar Tripathi[1]*, Tarek Salem Abdennaji[2], and Yewuhalashet Fissha[3,4]*

1. Department of Mining Engineering, Aditya University, Surampalem, Andhra Pradesh, India
2. Department of Civil Engineering, College of Engineering, Northern Border University, Arar, Saudi Arabia
3. Department of Electrical and Computer Engineering, National Institute of Technology, Asahikawa College, Asahikawa, Japan
4. Department of Mining Engineering, Aksum University, Aksum, Ethiopia

Article Info

Received 24 February 2025
Received in Revised form 12 May 2025
Accepted 20 June 2025
Published online 21 June 2025

DOI: 10.22044/jme.2025.15823.3043

Keywords

Gross Calorific value
Machine learning regression
Robust regression models
Asia-Pacific coal dataset
RANSAC

Abstract

Coal quality is predominantly determined by its Gross Calorific Value (GCV), which directly influences its economic valuation. Traditional empirical formulas for GCV estimation, though effective, become inefficient and laborious when handling large datasets. To address this, machine learning (ML) techniques offer a robust alternative for accurate and rapid predictions. This study employs seven coal quality parameters. Total Moisture (TM), Ash (ASH), Volatile Matter (VM), Hydrogen (H), Carbon (C), Nitrogen (N), and Sulphur (S), as independent variables to develop predictive models for GCV. Four conventional regression techniques, namely Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Random Forest (RF), and Decision Tree (DT), along with two robust regression models Random Sample Consensus (RANSAC) and Huber Regressor (HR) are explored. The dataset comprises coal samples from five Asia-Pacific countries: China, Indonesia, Korea, the Philippines, and Thailand. Comparative performance analysis reveals that the robust regression models significantly outperform the conventional ML techniques. The RANSAC and Huber Regressor models achieve superior prediction accuracy with R^2 values of 0.9941 and 0.9952, respectively. These findings highlight the potential of robust regression approaches for reliable GCV estimation, facilitating efficient coal quality assessment in large-scale applications.

1. Introduction

In recent times, the emergence of various renewable and green energy technologies with the backing of environmental bodies is a happening phenomenon. Although this phenomenon has been welcomed and tried to be implemented by governments of different nations for a sustainable and cleaner environment; coal is still the most dominant and widely adopted energy option worldwide [1]. In the scenario of India and China, coal is the most widely used source for primary energy consumption [2]. The consumption of primary energy, and the percentage of it; fulfilled by coal, of the top 11 coal consuming countries in the world is presented in Table 1. From the table, it can be clearly seen that more than 50% of the primary energy demand for India and China are fulfilled by coal; only one other country, i.e. South Africa, uses more percentage (71.02%) of coal to fulfil their primary energy requirements. Comparison between total primary energy and coal consumption of the countries is presented in Figure 1.

The primary energy consumption, and the percentage, of that, fulfilled by coal for the countries considered in this study is presented in Table 2. Here, it can be observed that, except Thailand all the other 4 countries rely heavily on coal as the primary source of energy.

Corresponding author: yowagaye@gmail.com (Y. Fissha)

Table 1. Primary energy consumption (in exajoules) by top 11 coal consuming countries [3]

Country	Coal	Oil	Natural gas	Nuclear energy	Hydro-electricity	Renewables	Total primary energy consumption	% of coal in total primary energy consumption
China	82.27	28.5	11.9	3.25	11.74	7.79	145.46	56.55
India	17.54	9.02	2.15	0.4	1.45	1.43	31.98	54.84
US	9.2	32.54	29.95	7.39	2.56	6.15	87.79	10.47
Japan	4.57	6.49	3.76	0.38	0.69	1.13	17.03	26.83
South Africa	3.48	1.02	0.15	0.14	Less Than 0.05	0.11	4.9	71.02
Russia	3.27	6.39	14.81	1.92	1.89	0.04	28.31	11.55
Indonesia	3.26	2.81	1.5	NA	0.17	0.37	8.1	40.24
South Korea	3.03	4.9	2.04	1.42	0.03	0.36	11.79	25.69
Vietnam	2.1	0.98	0.31	NA	0.61	0.08	4.09	51.34
Germany	1.84	4.21	3.12	0.57	0.17	2.21	12.11	15.19
Australia	1.69	1.83	1.47	NA	0.13	0.45	5.57	30.34

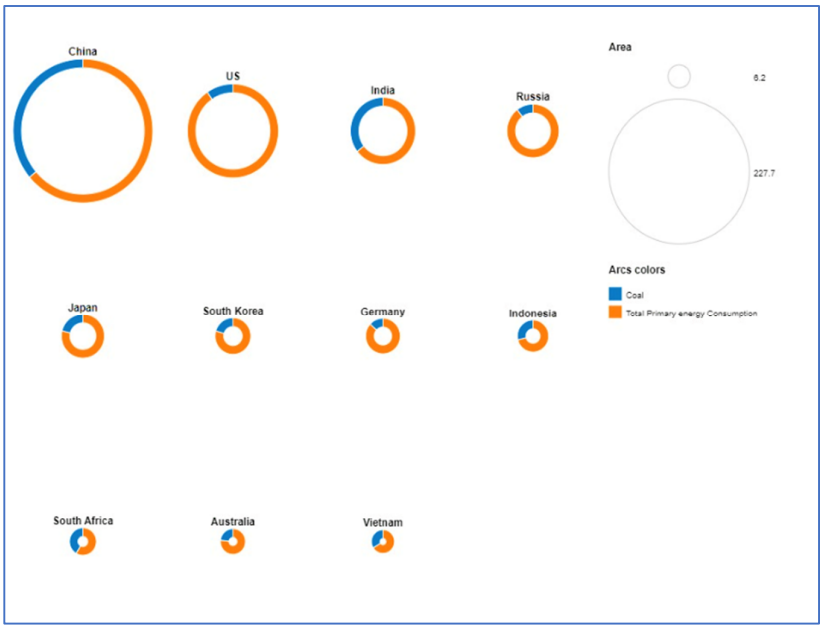

Figure 1. Comparison between total primary energy and coal consumption within and between [3]

Table 2. Primary energy consumption (in exajoules) of the countries considered for this work [3]

Country	Coal	Oil	Natural gas	Nuclear energy	Hydro-electricity	Renewables	Total primary energy consumption	% of coal in total primary energy consumption
China	82.27	28.5	11.9	3.25	11.74	7.79	145.46	56.55
Indonesia	3.26	2.81	1.5	NA	0.17	0.37	8.1	40.24
South Korea	3.03	4.9	2.04	1.42	0.03	0.36	11.79	25.69
Philippines	0.73	0.75	0.14	NA	0.06	0.15	1.82	40.10
Thailand	0.73	2.39	1.69	NA	0.04	0.28	5.12	14.25

Of the countries, which mine coal for economical purposes; almost a little below three quarter of these process it and roughly; 75% of the total coal usage is for power generation in thermal power plants [4]. The economy of the world is heavily influenced by coal, particularly, in India and other Asia Pacific countries rely heavily on coal for economic, and energy demands [5]. With a wide usage in various industries, the proper knowledge of coal quality becomes vital prior to its use in various industrial purposes. The quality of coal, along with some other parameters, is commonly expressed and based on its GCV [6]. Along with quality determination, GCV plays a major role in deciding coal price also, for example, in India, the price for coal consignments are determined by multiplying GCV value with a pricing coefficient (In paise per unit of energy (generally in kilocalorie)) [7].

The calorific value defines the amount of heat energy released during the complete combustion of coal. The gross release heat during the process can

be termed as Gross Calorific Value (GCV), which is a qualitative parameter, and is widely used for evaluating the quality of coal. There are various methods for determining the calorific value of coal. Each one of these has its benefits and limitations. A bomb calorimeter is a common method for measuring heat. It burns coal in high-pressure oxygen [8]. The temperature change in the surrounding water measures the heat released from the coal. This method gives accurate results, but it's complex and takes a lot of time. So, it may not be the best choice in some situations. Dulong's formula is another method that utilises the elemental composition of coal to provide a quicker estimation of the GCV of coal. Still, due to not considering the variation in coal structure, this method provides less accuracy [9]. Proximate analysis gives a quicker but less accurate GCV estimate. It uses elemental factors like moisture; volatile matter, fixed carbon, and ash content. New methods such as microwave-assisted calorimetry allow for fast GCV estimation [10]. They also need less sample material. This method isn't used much in industry or academia. This is because there isn't enough advanced equipment available.

Predictive modeling of GCV allows for real-time decisions in coal classification. This means less need for long lab tests. Machine learning models analyze large datasets. They find complex nonlinear relationships among coal's elemental properties. Unlike empirical formulas, ML models can include more factors [11]. These factors are mineral composition, coal rank, combustion traits, and spectral or imaging data. This ensures better prediction accuracy of the developed ML based GCV prediction models. This method boosts accuracy and cuts down on the need for many experiments. As a result, it saves a lot of energy. ML boosts its predictive accuracy as it gets more data. It adapts to various coal types without needing new formulas. Plus, it offers quick estimates without costly lab tests. This makes it a smart and scalable choice for industries [12].

According to Scopus database, a total of 33 articles were published during the years 2004 to 2021, based on the prediction of GCV from either proximate or ultimate analysis [13]. The year wise documents published is presented in Figure 2.

The prediction of GCV against input characteristics derived from coal proximate analysis has been the subject of numerous studies [1] [8] [14] [15] [16] [17]. Similarly, a number of GCV prediction studies have been conducted using the parameters derived from coal's final analysis [18] [19] [20]. GCV, as a function of the parameters derived from both analyses, has also not been predicted by many researchers [21]. Further research work is necessary, because GCV relies on both analysis and its scant literature.

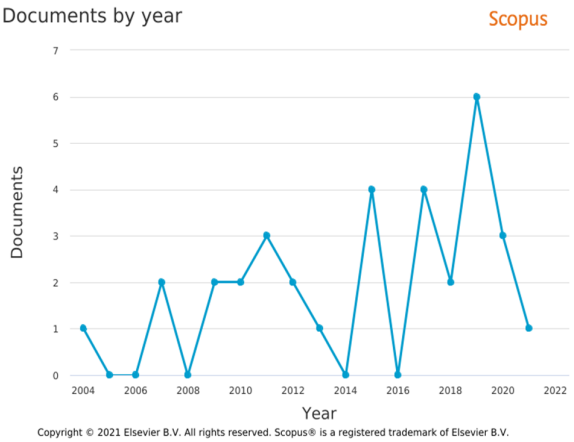

Figure 2. Documents published on prediction of GCV of coal year-wise [2004-2021] (Scopus data)

Chelgani (2021) had carried out the work of estimating GCV from conventional coal properties such as carbon, ash, total moisture, hydrogen, sulphur, volatile matter, nitrogen. The work was conducted using the XGBoost technique of machine learning and each output of the model was interpreted by SHAP [22]. Wen, Jian, and Wang (2017) carried out the prediction of GCV with both model and Gaussian Wavelet Neural Networks (WNNs) by considering moisture (total and coal internal), ash, volatile matter; fixed carbon, sulphur, hydrogen, oxygen, and nitrogen as independent parameters. The modelling was done separately for proximate and ultimate analysis parameters; the MWNN models predicted 99.4 and 83.2 percentage of the GCVs with a deviation of less than 1MJ/kg as a function of ultimate and proximate analysis respectively, whereas the GWNN models predicted the same for 99.4 and 84.2 percentages of the GCVs [23].

A similar approach was adopted by Matin and Chelgani (2016) by taking ash, moisture, and volatile matter, as proximate analysis parameters and carbon, hydrogen, sulphur and nitrogen as ultimate analysis parameters to develop Random Forest (RF) models for prediction of GCV. The developed models predicted the GCVs with R^2 values of 0.97 and 0.99 for proximate and ultimate analysis input parameters, respectively [24].

In this work, the different parameters namely: total moisture (TM), ash (ASH), volatile matter (VM), hydrogen (H), carbon (C), nitrogen (N), and sulphur (S) have been taken together as the

independent variables, to develop, four conventional Machine Learning (ML) regression models, namely, Support Vector Regression (SVR), K Nearest Neighbours (KNN) , Random Forest (RF), and Decision Tree (DT), and two robust regression models, namely, Random Sample Consensus (RANSAC), and Huber Regressor (HR) for predicting GCV of coals of selected Asia Pacific countries.

2. Literature Review

Mahvash Mohammadi and Hezarkhani used satellite data and machine learning models like SVM and random forest to identify alteration zones for mineral exploration. They found that random forest gave slightly better results [25]. Srivastava, Choudhary, and Sharma, tested various machine learning methods. They aimed to predict ground vibration from blasting in mines. Their work showed that ensemble models like gradient boosting were the most accurate [26]. Celik and Genc used Sentinel-2 satellite data. They applied different machine learning algorithms to predict geochemical patterns. Random forest stood out as the best performer for this type of remote sensing task [27]. Alamdari and his team applied ML models to predict how much fuel haul trucks would use in open-pit mines. Their models found the key factors that affect fuel use. They also suggested ways to boost efficiency [28]. Emami Meybodi and co-authors predicted rock fracture toughness (mode-I and mode-II) using ML models like SVR and RF. Their work helps in designing safer and more effective rock engineering operations [29]. Nabavi and colleagues created a hybrid model. They used XGBoost along with metaheuristic optimization methods. This model predicts back-break after blasting. Their model gave better results than traditional approaches [30]. Khamis, El-Rammah, and Salem, used KNN and MLP algorithms. They predicted how quickly a drill could penetrate rock in oil and gas drilling. They found MLP to be slightly more accurate [31]. Mohammadzadeh and his team used machine learning to find areas that might have copper and gold deposits. They improved prediction accuracy with a semi-supervised Bayesian model. This method used fewer labeled data points [32]. Sahoo, tripathy, and Jayanthu, reviewed how ML techniques analyze slope stability. They explained how models like decision trees and neural networks can help assess slope safety more effectively [33]. Ravi Kiran and Karra used ML models to explore how mining impacts air quality. They focused on particulate matter close to a coal mine. Their work supports better pollution monitoring and control strategies [34]. Marquina Araujo and team used machine learning models to estimate copper ore grades in a mining area. Their study helps mining companies improve planning and reduce waste during ore extraction [35]. Nag and Mishra applied ML techniques to the field of mining heritage tourism. They studied visitor data and suggested ways to improve tourism at historical mining sites [36]. Cotrina Teatino and co-authors used ML to predict production levels in open-pit mines in Peru. Their models, like neural networks and boosting methods, gave helpful insights for better mine planning [37]. Madani tested satellite-based band ratio techniques with ensemble ML models. The goal was to predict iron and titanium mineralization in Saudi Arabia. The combination turned out to be very effective [38]. Jahantigh and Ramazi created a geological map using airborne geophysical data. They applied fuzzy c-means, an unsupervised machine learning method, for this task. This approach helped predict rock types in unexplored areas of southern Iran [39]. Elgindy, Nooh, and Wahba created a machine learning model. It helps spot early signs of a "kick" in oil and gas drilling. The model successfully spotted pressure changes before dangerous situations could arise [40].

3. Data and Modelling

This study used six machine learning models to predict the Gross Calorific Value (GCV) of coal namely, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), RANSAC Regression, Huber Regression (HR). These models were chosen for their varied learning methods. They handle noise well and capture complex patterns in the dataset effectively.

- Support Vector Regression (SVR): SVR is a strong regression method. It uses a kernel-based approach to transform data into a higher-dimensional space. This helps it model nonlinear relationships well. It seeks to reduce margin violations within a set tolerance. This makes it very effective for small and medium-sized datasets that have complex patterns.
- K-Nearest Neighbors (KNNs): KNN is a model that predicts outcomes by averaging the values of the k-nearest training samples. It does not assume any specific data distribution. Its simplicity and flexibility make it ideal for regression tasks that have local similarity patterns. However, it can be

costly in terms of computation when dealing with large datasets.

- Random Forest (RF): RF is a method that uses many Decision Trees. It boosts accuracy and helps prevent overfitting. Combining several weak learners improves model stability and generalization. This makes it a strong option for nonlinear regression tasks, such as GCV estimation.
- Decision Tree (DT): DT is a rule-based model that splits data recursively to form a tree-like structure. It offers quick predictions and clear insights. However, it can overfit, hurting its performance when compared to ensemble methods.
- RANSAC Regression: It creates several models by picking random data subsets. Then, it finds the best fit. This method works well for datasets with measurement errors or unusual values.
- Huber Regression (HR)**:** HR is a strong linear regression method. It combines Mean Squared Error (MSE) and Mean Absolute Error (MAE) loss functions. This helps it manage noise and outliers well. Its flexible method is useful when the dataset has extreme variations.

The chosen models offer both linear and nonlinear learning options. This mix allows for a thorough exploration of coal GCV prediction. SVR, KNN, and RF effectively capture complex relationships. RANSAC and HR boost robustness by reducing the impact of outliers.

We used Python's Scikit-learn library to implement all machine learning models. This library is popular for machine learning applications. We did data preprocessing, model training, hyperparameter tuning, and evaluation in Jupyter Notebook. This setup helps with reproducibility and makes model deployment more efficient. This method accurately predicted the Gross Calorific Value of coal. It also made the model stronger against noise and changes in coal composition.

The data used for this work is extracted from the World Coal quality (WCQ) inventory maintained by U.S. Geological Survey. Though, GCV data, for a lot of countries are available in the inventory, for this work, the data of China, Indonesia, Korea, Philippines, and Thailand have been taken. The purpose of selecting data from such a widely distributed geographical area is to create an ML model, which is, robust, not dependent on homogenous data collected from similar mines or regions, and can handle outliers effectively. Because of this non-homogeneity in the data, a lot of outliers were found which in turn aided in fulfilling the goal of this work. The entries with NAN values were completely taken off the data set used for modelling. The description of the data is presented in Table 3.

Table 3. Description of data used

Attributes/Variables	TM (%Wt)	ASH (%Wt)	VM (%Wt)	H (%Wt)	C (%Wt)	N (%Wt)	S (%Wt)
Count	80	80	80	80	80	80	80
Mean	20.112	13.893	27.99788	3.341125	51.30675	0.781875	2.124625
Std	13.97776	8.638256	8.679361	0.988358	17.51361	0.335971	2.369157
Min	0.61	1.85	2.5	0.41	16.82	0.1	0.2
Max	43.44	47.33	44.3	5.24	82.84	1.94	10.81
25%	5.545	7.485	26.4575	2.8375	38.8	0.5625	0.5575
50%	22.31	13.155	29.715	3.6	47.925	0.705	0.94
75%	32.8625	19.1025	31.685	4.0025	68.345	1.0025	3.635

The data used is also presented in Figure 3, if analysed the outliers present in the data can be clearly visible. Further, before going for modelling, correlations between the variables have been established and is presented in the corresponding correlogram in Figure 4.

Six predictive models were developed, including four conventional machine learning models—SVR, KNN, RF, and DT and two robust machine learning models, namely Random Sample Consensus (RANSAC) and Huber Regressor (HR). Machine learning algorithms play a crucial role in predicting the Gross Calorific Value (GCV) of coal by identifying complex relationships between input parameters and combustion characteristics. In this study two type of model such as conventional supervised learning and robust regression are used. Each predictive modeming method contributes for enhancing the prediction accuracy. Support Vector Machine (SVM) is a widely used conventional machine learning algorithm that effectively handles both classification and regression problems [41]. Its preference in predictive modeling stems from its ability to maximize the margin between data points, leading to higher accuracy [42] [43]. In GCV estimation, SVM helps establish non-linear relationships between coal properties and calorific value [44]. Similarly, K-

Nearest Neighbors (KNN), a non-parametric supervised learning algorithm, relies on the proximity of similar data points to make predictions. Its effectiveness in regression problems makes it a suitable choice for approximating GCV based on coal composition [45]. Decision Tree (DT), another supervised learning approach, is widely applied for both classification and regression tasks. It creates an interpretable hierarchical structure for decision-making, making it particularly useful for understanding how coal properties influence GCV [46]. Unlike categorical variable decision trees that predict discrete outcomes, this study utilizes DT for continuous variable prediction, making it well-suited for GCV estimation [47]. Random Forest (RF) extends the capabilities of Decision Trees through an ensemble learning approach, where multiple decision trees collectively improve prediction accuracy and generalization. RF is particularly advantageous in GCV modeling as it reduces overfitting and enhances robustness by averaging multiple predictions, capturing intricate dependencies between coal constituents and calorific value [48].

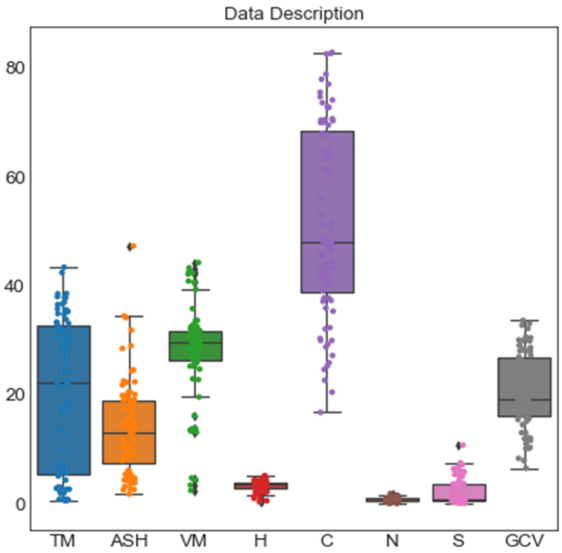

Figure 3. Graphical representation of data used

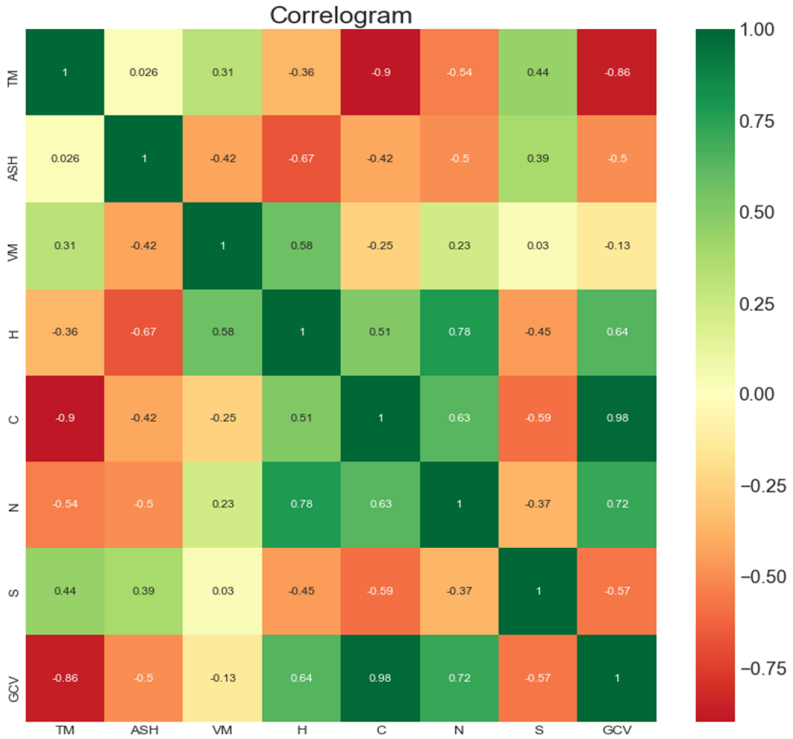

Figure 4. Correlations between the variables

While these conventional models perform well under normal conditions, their accuracy may be affected by the presence of outliers in the dataset. If the input data contains significant outliers in either the independent or dependent variables, robust regression algorithms are more suitable for modelling [49]. Huber Regression, a well-known robust regression algorithm, identifies outliers and assigns them lower weight than other data points, ensuring that extreme values do not excessively influence the model [50]. This makes Huber Regression highly effective for datasets with

moderate outliers. Another robust approach, Random Sample Consensus (RANSAC), separates data into inliers and outliers and fits the model using only the inliers, making it particularly useful for handling datasets with a significant number of anomalies [51]. By integrating both conventional and robust machine learning techniques, this study ensures a more accurate and reliable prediction of GCV, effectively addressing the complexities associated with coal quality assessment [52]. The common work flow of the models used is presented in Figure 5.

The workflow or mechanism of modelling were same for all the models; the difference in performance were because of the algorithms of the models developed. The entire study data used were separated in two categories such as input features and target variable, and inserted into two data frames namely X and Y. The dataset used in this study consisted of 80 entries, which were divided into 85% for training (68 entries) and 15% for testing (12 entries) to ensure a reliable model evaluation. However, we chose the 85–15 split as the baseline. This split fits well with common industry standards for real-time quality monitoring. In these workflows; the focus is often on maximizing training data rather than validation size. This is especially true for smaller datasets. The results from all splits showed that the model.

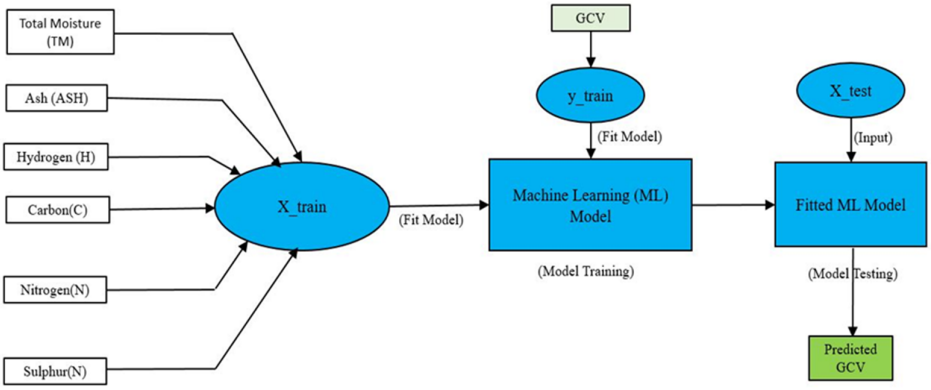

Figure 5. Workflow of the models

evaluation was both fair and reproducible. Before being fed into the machine learning models, the data was standardized using a scaling technique to maintain uniformity and improve model performance.

The processed data was then stored in four distinct arrays: X_train, Y_train, X_test, and Y_test. Here, X_train and Y_train represent the independent and dependent variables for training, respectively, containing 68 samples each. Similarly, X_test and Y_test store the independent and dependent variables for testing, comprising 12 samples. This structured approach ensures that the model learns effectively from the training data, while being evaluated on an unseen test set to validate its predictive accuracy. We tested 80–20 and 5-fold cross-validation splits to check generalization and robustness. We used quantile-based stratification to split the data into equal parts. This helps manage any imbalances in the GCV target variable before we divide it into training and testing sets. This ensured that both sets retained similar distribution characteristics. For regression tasks, use different techniques instead of SMOTE or ADASYN, which are for classification. This stratification made sampling fairer. However, the small changes in R^2 (±0.5%) and MAE (±1%) showed that the practical benefits were limited. Diagnostic plots of the GCV distribution showed our dataset was not very imbalanced. Thus traditional sampling was enough to fine-tune the models; we used both Grid Search and Bayesian Optimization (BO) strategies. Each model followed set hyperparameter ranges from previous research and early tests. For example, SVR models were optimized over $C \in [0.1, 1000]$, epsilon $\in$ [0.01, 1.0], and various kernel types. Random forest models were adjusted with n_estimators between 50 and 500, and max_depth from 3 to 20. We ran bayesian optimization for up to 50 iterations. Early stopping was used if the validation metric (R^2) didn't improve for 10 rounds in a row. Table X shows the hyperparameter space for each model. BO needed fewer iterations than Grid Search to find optimal settings. It's better at exploring complex parameter spaces with fewer evaluations. The hyper parameter search space is presented in Table 4. All models underwent hyperparameter tuning using either grid search or Bayesian optimization techniques. Table X

summarizes the parameters considered, their search ranges, and the best values selected. This step made sure each model worked its best. It improved how well they predicted and how they could be applied in different situations.

Table 4. Hyperparameter search space and optimal values for each model

Model	Hyperparameters tuned	Search range / Options	Best value (selected)
SVR	C, epsilon, kernel	C: [0.1, 1000], ε: [0.01, 1.0], kernel: ['linear', 'rbf', 'poly']	C = 100, ε = 0.1, kernel = 'rbf'
KNN	n_neighbors, weights	n_neighbors: [1, 20], weights: ['uniform', 'distance']	n_neighbors = 5, weights = 'distance'
RF	n_estimators, max_depth, max_features	n_estimators: [50, 500], max_depth: [5, 30], max_features: ['auto', 'sqrt']	n_estimators = 200, max_depth = 20, max_features = 'sqrt'
DT	max_depth, min_samples_split	max_depth: [5, 30], min_samples_split: [2, 10]	max_depth = 15, min_samples_split = 4
RANSAC	max_trials, residual_threshold	max_trials: [50, 500], threshold: [1.0, 5.0]	max_trials = 300, threshold = 2.0
HuberReg	epsilon, alpha	epsilon: [1.0, 2.5], alpha: [0.0001, 1.0]	epsilon = 1.35, alpha = 0.01

4. Model outcomes and its Analysis

The outcomes obtained from the models and their performance evaluation is discussed in this section of the work. The performance of the models were evaluated using few of the usual parameters like coefficient of determinations and errors obtained while predicting. R^2 of each model was determined along with their respective Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The measured evaluation coefficients for each developed models are presented in Table 5. We trained models using the original feature set. We looked at feature selection methods, such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). These methods helped us reduce dimensions and improve prediction accuracy. PCA transformed the input variables into principal components. PCA captured data variance but made it harder to interpret key physical features. Ash, volatile matter, and fixed carbon are important for understanding in this field. Model performance with PCA-transformed features didn't improve much compared to the original set. RFE helped rank and remove less important features with different machine learning models. The analysis showed that every original feature played an important role in the model's performance. No feature was found to be unnecessary or removable without harming accuracy. So, to keep both predictive power and clarity, we retained all the original input features for final modeling and analysis. This method makes sure the predictive models match the physical properties of coal that affect gross calorific value.

Table 5. Evaluation coefficients of the developed models

Model	R^2	MSE	RMSE	MAE	Adjusted R^2	MBE	95% confidence interval
SVR	0.938	2.74645	1.657242	1.26175	0.8823	0.25	(-3.94, 4.44)
KNN	0.9618	2.0519	1.432445	1.265	0.9548	1.0803	(-0.31, 2.48)
RF	0.9319	3.00003	1.732059	1.337917	0.9432	0.425	(-2.38, 3.23)
DT	0.8956	4.251667	2.061957	1.41666	0.9567	1.0426	(-0.37, 2.46)
RANSAC	0.9941	1.566071	1.251428	1.042647	0.9182	0.4898	(-2.88, 3.86)
HR	0.9952	1.631249	1.277204	1.08026	0.9258	0.4588	(-2.76, 3.68)

From the data presented in Table 5, it is evident that the robust models RANSAC and HR have outperformed their conventional counter parts with R^2 values of 0.9941 and 0.9952, respectively. Further, among the conventional models, KNN performed better than the other three with an R^2 value of 0.9618. The worst performing model among all the six was DT with R^2 value of 0.8956. Adjusted R^2 adds to R^2 by considering model complexity. It penalizes predictors that don't help improve performance. This shows that models like KNN and RANSAC are strong. They keep high Adjusted R^2 values, which means they use input features well. Mean Bias Error (MBE) reveals model trends. KNN and DT have a positive MBE, hinting at slight overestimation. In contrast, SVR and HR have low MBE, showing better bias control. Narrow 95% Confidence Intervals (e.g., HR: –2.76 to 3.68) show high prediction reliability and low error dispersion. This confirms the model's consistency. These metrics support choosing strong models like HR and RANSAC instead of just RMSE or MAE. They provide a better overall view of model reliability for predicting coal GCV. Although stacking marginally improved R^2 by ~1.1% and reduced RMSE, the added computational complexity made real-time deployment less feasible in the current scope.

However, the trade-off may be worthwhile in future settings with relaxed runtime constraints.

A few plots were also drawn to further establish the performances of the models and are presented in figures below. First, the regression plots drawn between the actual GCV and predicted GCV of all the models developed are presented in Figures 6 to 11.

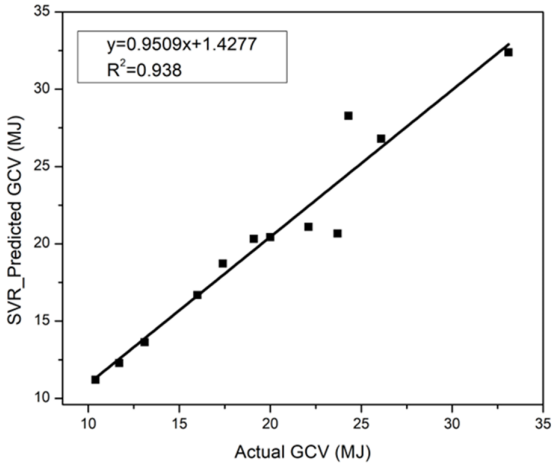

Figure 6. Plot between actual and predicted GCV by SVR

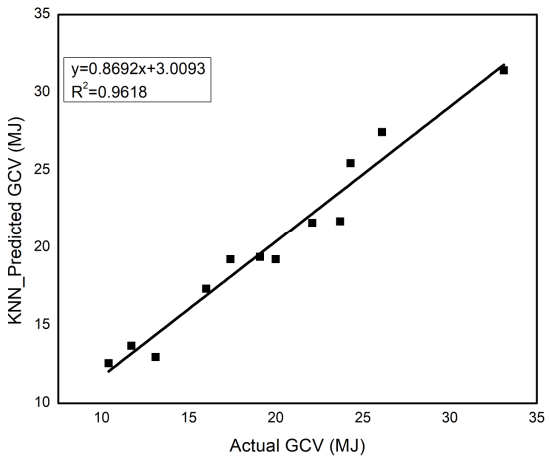

Figure 7. Plot between actual and predicted GCV by KNN

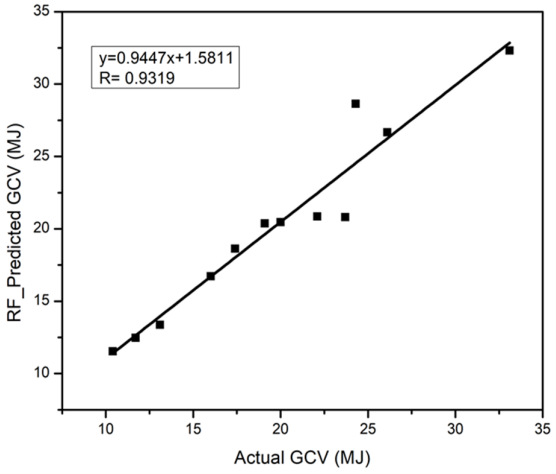

Figure 8. Plot between actual and predicted GCV by RF

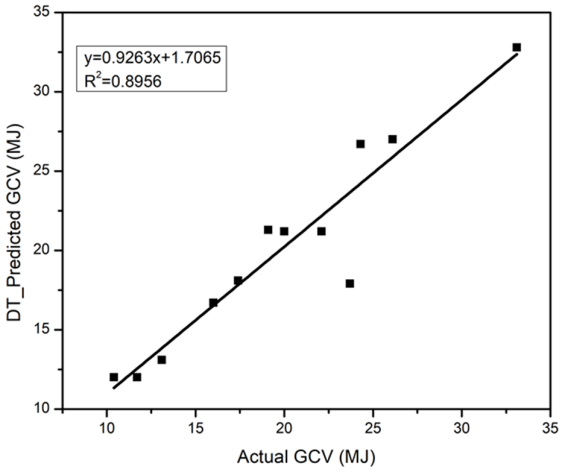

Figure 9. Plot between actual and predicted GCV by DT

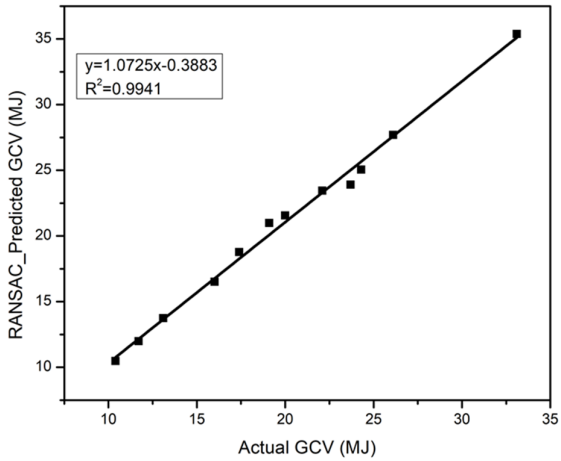

Figure 10. Plot between actual and predicted GCV by RANSAC

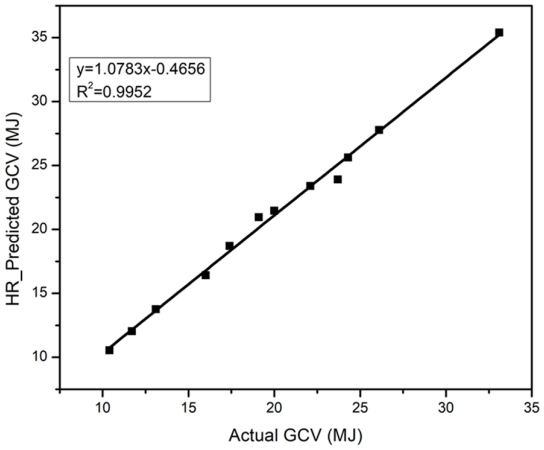

Figure 11. Plot between actual and predicted GCV by HR

From the figures, it is now more evident by observing the regression lines for each of the models that, the robust models RANSAC and HR have performed really well and better than all the other models. To establish a visual comparison, between the actual and predicted GCV, two separate figures, with the sample numbers in the X axis and the actual/predicted GCV in the Y axis have been drawn and presented, with all the conventional models in Figure 11, and the robust models in Figure 12.

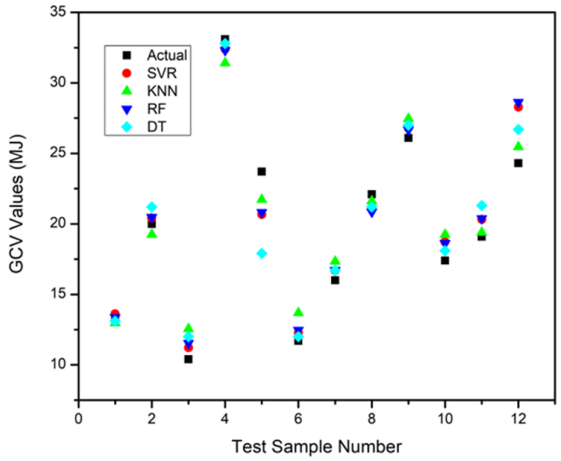

Figure 12. Actual and predicted GCV for conventional models

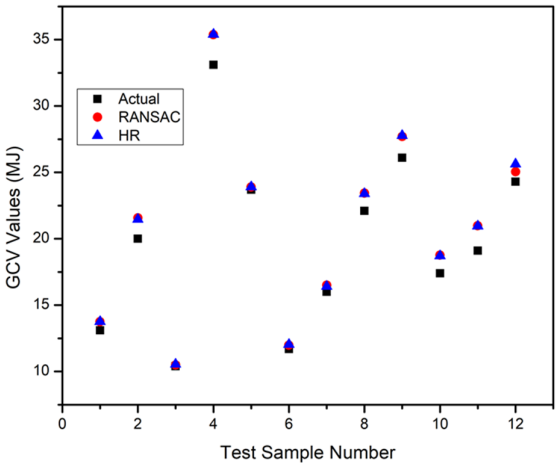

Figure 13. Actual and predicted GCV for robust models

From the above figures, the predicted GCV line plots for the robust models are coinciding or almost matching the actual GCV line plot at most of the instances whereas the same trend is not followed for the conventional models.

Further, in order to compare the errors generated by the models while predicting GCV a residual plot has been drawn and presented in Figure 13.

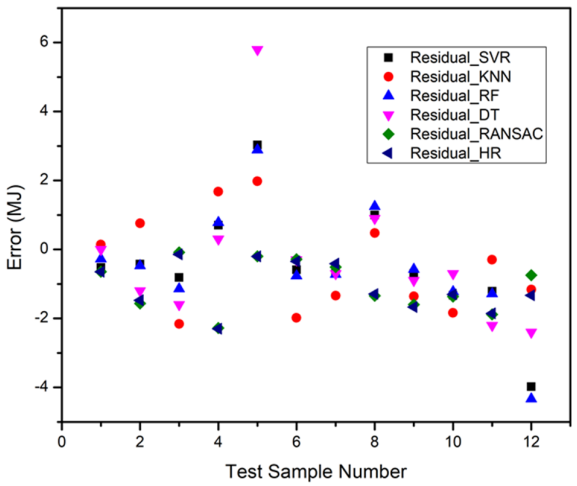

Figure 14. Residuals of the models developed

If Figure 13 is observed closely, the maximum errors for the KNN, HR, and RANSAC models do not exceed ±2 MJ, except for two instances (Sample numbers 3 and 4). For the other three models, the errors deviate more significantly. At the highest deviation from the zero-error line, the errors reach +6 MJ for Sample 5 and -4 MJ for Sample 12. Further analysis of the residuals of the three best-performing models—KNN, HR, and RANSAC—reveals interesting insights. Although KNN has a slightly smaller highest deviation compared to HR and RANSAC, its overall deviations across all instances are larger. This explains why HR and RANSAC perform better overall compared to KNN.

The comprehensive evaluation of used ML models for GCV prediction can be seen from Taylor diagram, which is presented in Figure 15. This diagram is used to assess the performance of ML models by comparing their output with the observed data. The comparison is performed based on three performance metrics such as standard deviation, correlation coefficient, and root mean square error (RMSE). The radial distance from the origin represents the standard deviation, it indicates the variability of the predicted models. The lesser radial distance shows low variation in the predicted model. The angular position of the Taylor diagram shows the correlation coefficient where models closer to the horizontal axis (correlation = 1) demonstrate stronger agreement with the actual values. The distance of each model from the reference observation represents the RMSE, with

shorter distances indicating higher predictive accuracy. From the diagram, it is evident that Huber Regressor (HR) and RANSAC outperform the other models, achieving the highest correlation, lowest RMSE, and standard deviations closest to the reference data. In contrast, the conventional models, such as KNN, RF, and SVR, show larger deviations and comparatively weaker performance. Overall, the Taylor diagram confirms the robustness and reliability of HR and RANSAC in GCV prediction across diverse coal datasets.

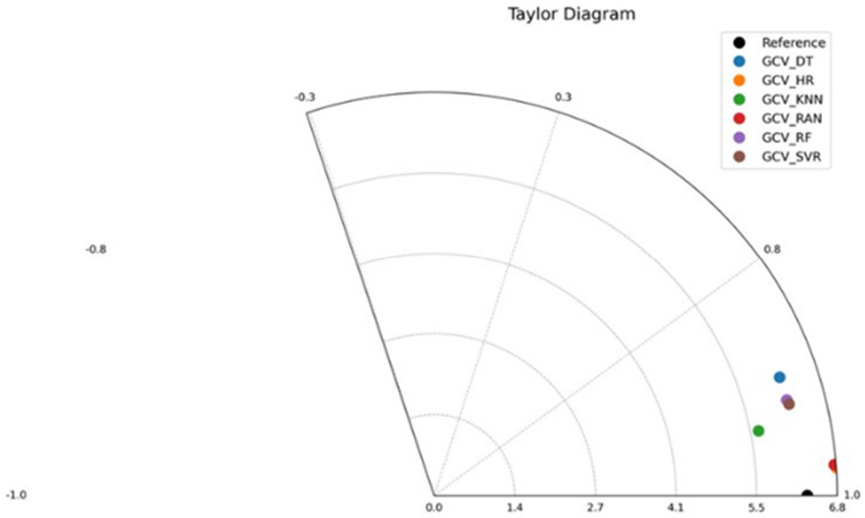

Figure 15. Taylor diagram, showing the performance of machine learning models in predicting GCV based on standard deviation, correlation, and RMSE

The outcomes of this study provide valuable tools for the coal industry to enhance the efficiency and accuracy of Gross Calorific Value (GCV) prediction. The machine learning models, particularly the robust regression techniques such as RANSAC and Huber Regressor, have demonstrated exceptional performance in handling outliers and non-homogeneous data. This makes them highly suitable for diverse coal samples collected from various geographical regions. These models offer a faster, more reliable, and scalable alternative to traditional empirical methods, enabling real-time coal quality assessment, even for large and complex datasets.

The inference time of each model is key for its use in industrial settings. For example, coal blending and dispatch operations require near real-time decisions. In such contexts, even a 500–800 ms delay could introduce inefficiencies in material routing. While Random Forest and stacked models showed higher accuracy, their longer inference times (e.g., ~720 ms) may pose challenges. On the other hand, strong regressors like RANSAC and Huber Regression take less than 250 ms. They provide a good mix of speed and accuracy for real-time use. The mapping of each model with the suitable application scenario is presented in Table 6.

Table 6. Mapping of each ML model with application scenario

Model	Avg. inference time (ms)	Suitable application scenario
RANSAC	240	Real-time blending, edge deployment
HR	225	On-site quality monitoring
RF	720	Offline batch processing, centralized control
SVR	350	Semi-real-time quality control
Stacking	800	Cloud analytics, predictive dashboards

The practical significance of these models is evident in their potential to improve coal classification, optimize pricing strategies, and support more accurate operational decision-making, thus minimizing errors and enhancing resource management. Moreover, accurate GCV prediction contributes to energy production efficiency by enabling power plants to optimize fuel utilization, reduce combustion-related emissions, and promote cleaner and more sustainable energy practices. The proposed models can also be seamlessly integrated into automated

testing systems and industrial processes, offering practical solutions for improving coal supply chain operations and aligning with the evolving needs of the energy and mining sectors. ML models did much better than empirical formulas for prediction accuracy (Table 5). Still, empirical formulas are still important in many real-world situations. Empirical models come from years of coal quality research. They are helpful when computer resources are low or when clear understanding is key. In areas without detailed datasets, empirical formulas give basic predictions. They rely on average trends in composition. ML models work best in situations where sensors can gather lots of data. They also shine when it's important to capture complex relationships. So, the choice between empirical and ML methods depends on specific factors. These include data availability, interpretability needs, and the tech infrastructure.

5. Practical Implication and Scope for Future Work

The machine learning models are highly useful for monitoring coal quality. They help with real-time Gross Calorific Value (GCV) estimation in industrial workflows. This model can be integrated into coal processing plants. Simply embed them in automated quality control systems. Linking predictive models with sensor data allows for real-time input. Parameters like moisture, volatile matter, fixed carbon, and ash content feed into the model. This setup enables instant GCV estimation. This integration boosts decision-making in coal blending. It also optimizes combustion efficiency in power plants. Plus, it improves supply chain management by ensuring consistent fuel quality. These models can also be used in cloud-based platforms. This allows remote access and predictive analytics for mining companies and coal users. This implementation cuts down on slow lab tests. It offers a cost-effective and efficient way to keep coal quality standards in industrial use.

The predictive accuracy of the developed models may be influenced by several potential sources of error. Measurement inconsistencies can come from differences in lab procedures and equipment calibration. These issues may also affect how well the model performs. Regional differences in coal composition can vary due to geological factors. This can cause discrepancies when using the models on different coal types. To tackle these challenges, we need to use better data preprocessing techniques. We also need sensor calibration methods and adaptive modelling approaches. These methods should consider regional differences.

Future research can focus on refining these models to enhance their robustness and adaptability. Deep learning methods, such as neural networks and convolutional models, find patterns and pull-out features. Future studies may include ensemble distillation or adaptive stacking to retain accuracy gains while reducing computational demands for real-time integration. Adding more coal properties, like elemental makeup and thermal behaviour, helps create a better dataset. This improves the accuracy of predicting GCV. This approach takes advantage of the strengths of various algorithms. These advancements can greatly enhance real-time coal quality monitoring. This makes GCV estimation more accurate and useful in various industries.

Future research will enhance the models. This will involve testing with different datasets from various locations. We will use a hold-out validation strategy. In this approach, coal samples from one country or geological zone will be left out during model training. These samples will only be used for independent testing. This method helps evaluate how well models work across different coal-producing regions. External validation will make the predictive models useful beyond the Asia-Pacific region. This supports their use in different industries.

For real-time use, you can embed the trained models into a coal quality monitoring system based on IIoT. Key coal parameters like moisture, ash content, and fixed carbon come from calibrated field sensors. These readings go into a compact edge device, such as an industrial microcontroller or a local server. The machine learning model is pre-trained offline. It runs on a lightweight runtime like ONNX or TensorFlow Lite. This setup allows for quick inference. You can see predicted GCV values on a SCADA interface. They can also be sent to cloud analytics platforms using MQTT or HTTP protocols. This system design provides quick, reliable, and affordable coal quality estimates during operations.

6. Conclusions

This study successfully demonstrates the application of machine learning techniques for predicting the Gross Calorific Value (GCV) of coal using hybrid properties—Total Moisture (TM), Ash (ASH), Volatile Matter (VM), Hydrogen (H), Carbon (C), Nitrogen (N), and Sulphur (S). Six regression models, including Support Vector

Regression (SVR), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Random Sample Consensus (RANSAC), and Huber Regressor (HR), were developed and compared. To assess model robustness, outliers were deliberately included by incorporating data from non-homogeneous countries across the Asia-Pacific region. The results reveal that the robust regression models, RANSAC and Huber Regressor (HR), significantly outperformed the conventional models, achieving exceptional R^2 values of 0.9941 and 0.9952, respectively. Their superior performance can be attributed to their ability to effectively handle outliers and data variability. While conventional empirical methods for estimating GCV remain useful, the machine learning models presented in this study offer faster, more reliable, and scalable solutions, especially for large datasets. The robust regression models, in particular, demonstrate strong potential for accurate GCV prediction across diverse and non-homogeneous data, making them suitable for real-world applications in coal quality assessment from various geographical regions or mines.

Data Availability

The datasets generated during and/or analysed during the current work are available from the corresponding author on reasonable request.

References

[1]. Mondal, C., Pandey, A., Pal, S. K., Samanta, B., & Dutta, D. (2022). Prediction of gross calorific value as a function of proximate parameters for Jharia and Raniganj coal using machine learning based regression methods. *International Journal of Coal Preparation and Utilization*, *42*(12), 3763-3776.

[2]. Sharvini, S. R., Noor, Z. Z., Chong, C. S., Stringer, L. C., & Yusuf, R. O. (2018). Energy consumption trends and their linkages with renewable energy policies in East and Southeast Asian countries: Challenges and opportunities. Sustainable Environment Research, *28*(6), 257-266.

[3]. Petroleum, B. (2021). Statistical Review of World Energy. https://www. BP. com/content/dam/BP/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review. *BP-stats-review-2021-full-report. pdf*.

[4]. Looney, B. (2021). Statistical Review of World Energy globally consistent data on world energy markets. *and authoritative publications in the field of energy. Rev. World Energy Data*, *70*, 8-20.

[5]. Shukla, A. K., Sudhakar, K., & Baredar, P. (2017). Renewable energy resources in South Asian countries: Challenges, policy and recommendations. *Resource-Efficient Technologies*, *3*(3), 342-346.

[6]. Majumder, A. K., Jain, R., Banerjee, P., & Barnwal, J. P. (2008). Development of a new proximate analysis based correlation to predict calorific value of coal. *Fuel*, *87*(13-14), 3077-3081.

[7]. Kumar, P., Chakravarty, S., & Majumder, A. K. (2025). Relationship between petrological characteristics and gross calorific value of coal. *Fuel*, *380*, 133180.

[8]. Vilakazi, L., & Madyira, D. (2025). Estimation of gross calorific value of coal: A literature review. *International Journal of Coal Preparation and Utilization*, *45*(2), 390-404.

[9]. Zhu, W., Xu, N., & Hower, J. C. (2025). Unveiling the Predictive Power of Machine Learning in Coal Gross Calorific Value Estimation: An Interpretability Perspective. *Energy*, 134781.

[10]. Yang, H., Liu, J., Zhang, B., Cheng, T., Zou, D., & Lv, X. (2024). Mechanism of microwave-assisted coal desulfurization with urea peroxide. *Process Safety and Environmental Protection*, *192*, 1127-1137.

[11]. Munshi, T. A., Jahan, L. N., Howladar, M. F., & Hashan, M. (2024). Prediction of gross calorific value from coal analysis using decision tree-based bagging and boosting techniques. *Heliyon*, *10*(1).

[12]. Pimpalkar, A. S., & Gote, A. C. (2024, December). Utilization of artificial intelligence and machine learning in the coal mining industry. In *AIP Conference Proceedings* (Vol. 3188, No. 1). AIP Publishing.

[13]. Feng, Q., Zhang, J., Zhang, X., & Wen, S. (2015). Proximate analysis based prediction of gross calorific value of coals: A comparison of support vector machine, alternating conditional expectation and artificial neural network. *Fuel Processing Technology*, *129*, 120-129.

[14]. Welcome.(n.d.). https://www.scopus.com/results/results.uri?sort=plff&src=s&st1=prediction+AND+of+AND+gross+AND+calorific+AND+value+AND+of+AND+coal&nlo=&nlr=&nls=&sid=cf19baf84bbc129ba397700ce940edd0&sot=b&sdt=b&sl=82&s=TITLEABSKEY%28prediction+AND+of+AND+gross+AND+calorific+AND+value+AND+of+AND+coal%29&cl=t&offset=21&origin=resultslist&ss=plf-f&ws=r-f&ps=r-f&cs=r-f&cc=10&txGid=836f4c083362aed09f196a23c153d0d6

[15]. Açıkkar, M., & Sivrikaya, O. (2018). Artificial neural networks for estimation of the gross calorific value of Turkish lignite coals. In *3rd International*

Mediterranean Science and Engineering Congress (IMSEC 2018) (pp. 1075-1079).

[16]. Akhtar, J., Sheikh, N., & Munir, S. (2017). Linear regression-based correlations for estimation of high heating values of Pakistani lignite coals. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *39*(10), 1063-1070.

[17]. Kumari, P., Singh, A. K., Wood, D. A., & Hazra, B. (2019). Predictions of gross calorific value of Indian coals from their moisture and ash content. *Journal of the Geological Society of India*, *93*(4), 437-442.

[18]. Tan, P., Zhang, C., Xia, J., Fang, Q. Y., & Chen, G. (2015). Estimation of higher heating value of coal based on proximate analysis using support vector regression. *Fuel Processing Technology*, *138*, 298-304.

[19]. Yilmaz, I., Erik, N. Y., & Kaynar, O. (2010). Different types of learning algorithms of artificial neural network (ANN) models for prediction of gross calorific value (GCV) of coals. *Scientific research and essays*, *5*(16), 2242-2249.

[20]. Chelgani, S. C., Mesroghli, S., & Hower, J. C. (2010). Simultaneous prediction of coal rank parameters based on ultimate analysis using regression and artificial neural network. *International Journal of Coal Geology*, *83*(1), 31-34.

[21]. Boumanchar, I., Charafeddine, K., Chhiti, Y., M'hamdi Alaoui, F. E., Sahibed-Dine, A., Bentiss, F., ... & Bensitel, M. (2019). Biomass higher heating value prediction from ultimate analysis using multiple regression and genetic programming. *Biomass Conversion and Biorefinery*, *9*, 499-509.

[22]. Sözer, M., Haykiri-Acma, H., & Yaman, S. (2022). Prediction of calorific value of coal by multilinear regression and analysis of variance. *Journal of Energy Resources Technology*, *144*(1), 012103.

[23]. Chelgani, S. C. (2021). Estimation of gross calorific value based on coal analysis using an explainable artificial intelligence. *Machine Learning with Applications*, *6*, 100116.

[24]. Wen, X., Jian, S., & Wang, J. (2017). Prediction models of calorific value of coal based on wavelet neural networks. *Fuel*, *199*, 512-522.

[25]. Mahvash Mohammadi, N., & Hezarkhani, A. (2020). A comparative study of SVM and RF methods for classification of alteration zones using remotely sensed data. *Journal of Mining and Environment*, *11*(1), 49-61.

[26]. Srivastava, A., Choudhary, B. S., & Sharma, M. (2021). A comparative study of machine learning methods for prediction of blast-induced ground vibration. *Journal of Mining and Environment*, *12*(3), 667-677.

[27]. Celik, T., & Genc, B. (2021). A Comparative Study on Machine Learning Algorithms for Geochemical Prediction Using Sentinel-2 Reflectance Spectroscopy. *Journal of Mining and Environment*, *12*(4), 987-1001.

[28]. Alamdari, S., Basiri, M. H., Mousavi, A., & Soofastaei, A. (2022). Application of machine learning techniques to predict haul truck fuel consumption in open-pit mines. *Journal of Mining and Environment*, *13*(1), 69-85.

[29]. Emami Meybodi, E., Hussain, S. K., Fatehi Marji, M., & Rasouli, V. (2022). Application of machine learning models for predicting rock fracture toughness mode-I and mode-II. *Journal of Mining and Environment*, *13*(2), 465-480.

[30]. Nabavi, Z., Mirzehi, M., Dehghani, H., & Ashtari, P. (2023). A hybrid model for back-break prediction using XGBoost machine learning and metaheuristic algorithms in Chadormalu iron mine. *Journal of Mining and Environment*, *14*(2), 689-712.

[31]. Khamis, Y. E., El-Rammah, S. G., & Salem, A. M. (2023). Rate of penetration prediction in drilling operation in oil and gas wells by k-nearest neighbors and multi-layer perceptron algorithms. *Journal of Mining and Environment*, *14*(3), 755-770.

[32]. Mohammadzadeh, M., Mahboubiaghdam, M., Jahangiri, M., & Nasseri, A. (2023). Machine Learning Predictive Approaches for Cu-Au Mineral prospectivity Map in Sonajil, NW of Iran: an Improvement by a Bayesian Semi-supervised Algorithm. *Journal of Mining and Environment*, *14*(4), 1321-1342.

[33]. Sahoo, A. K., Tripathy, D. P., & Jayanthu, S. (2024). Application of machine learning techniques in slope stability analysis: A comprehensive overview. *Journal of Mining and Environment*, *15*(3), 907-921.

[34]. Ravi Kiran, P., & Karra, R. (2024). Analysis of Concentration of Ambient Particulate Matter in the Surrounding Area of an Opencast Coal Mine using Machine Learning Techniques. *Journal of Mining and Environment*, *15*(3), 961-976.

[35]. Marquina Araujo, J. J., Cotrina Teatino, M. A., Mamani Quispe, J. N., Noriega Vidal, E. M., Vega Gonzalez, J. A., Vega-Gonzalez, J., & Cruz-Galvez, J. (2024). Copper Ore Grade Prediction using Machine Learning Techniques in a Copper Deposit. *Journal of Mining and Environment*, *15*(3), 1011-1027.

[36]. Nag, A., & Mishra, S. (2024). Revitalizing mining heritage tourism: A machine learning approach to tourism management. *Journal of Mining and Environment*, *15*(4), 1193-1225.

[37]. Cotrina Teatino, M. A., Marquina Araujo, J. J., Noriega Vidal, E. M., Mamani Quispe, J. N., Ccatamayo Barrios, J. H., Gonzalez Vasquez, J. A., & Arango

Retamozo, S. M. (2024). Predicting open pit mine production using machine learning techniques: A case study in peru. *Journal of Mining and Environment*, *15*(4), 1345-1355.

[38]. Madani, A. A. (2024). Evaluation of Band Ratio Technique for Prediction of Iron-Titanium Mineralization Using Ensemble Machine Learning Model: A Case Study from Khamal area, Western Saudi Arabia. *Journal of Mining and Environment*, *15*(4), 1357-1371.

[39]. Jahantigh, M., & Ramazi, H. R. (2025). Integration of Airborne Geophysics Data with Fuzzy c-means Unsupervised Machine Learning Method to Predict Geological Map, Shahr-e-Babak Study Area, Southern Iran. *Journal of Mining and Environment*, *16*(1), 273-289.

[40]. Elgindy, M. Y., Nooh, A. Z., & Wahba, A. M. (2025). A New Proposed Model for Early Kick Detection in Drilling Operation Using Machine Learning. *Journal of Mining and Environment*, *16*(2), 439-451.

[41]. Matin, S. S., & Chelgani, S. C. (2016). Estimation of coal gross calorific value based on various analyses by random forest method. *Fuel*, *177*, 274-278.

[42]. Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning* (pp. 207-235). Boston, MA: Springer US.

[43]. Wan, C. H., Lee, L. H., Rajkumar, R., & Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. *Expert Systems with Applications*, *39*(15), 11880-11888.

[44]. Kumar, M. A., & Gopal, M. (2009). Least squares twin support vector machines for pattern classification. *Expert systems with applications*, *36*(4), 7535-7543.

[45]. Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, *24*(12), 1565-1567.

[46]. Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, *8*(4), 341-357.

[47]. Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (pp. 99-111). Springer Singapore.

[48]. Tripathy, D. P., Parida, S., & Khandu, L. (2021). Safety risk assessment and risk prediction in underground coal mines using machine learning techniques. *Journal of The Institution of Engineers (India): Series D*, *102*(2), 495-504.

[49]. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *8*(4), e1249.

[50]. Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection*. John wiley & sons.

[51]. D'Urso, P., Massari, R., & Santoro, A. (2011). Robust fuzzy regression analysis. *Information Sciences*, *181*(19), 4154-4174.

[52]. Fotouhi, M., Hekmatian, H., Kashani-Nezhad, M. A., & Kasaei, S. (2019). SC-RANSAC: spatial consistency on RANSAC. *Multimedia Tools and Applications*, *78*, 9429-9461.

انجمن مهندسی معدن ایران
نشریه مهندسی معدن و محیط زیست
نشانی نشریه: www.jme.shahroodut.ac.ir
دانشگاه صنعتی شاهرود

مدل‌سازی پیش‌بینی ارزش حرارتی ناخالص زغال‌سنگ با استفاده از تکنیک‌های رگرسیون یادگیری ماشینی مرسوم و قوی

ساتیا جیت پاریدا[۱]، آبیشک کومار تریپاتی[۱]، طارق سالم عبدالنجی[۲] و یووهالاشت فیشا[۳،۴*]

۱ .گروه مهندسی معدن، دانشگاه آدیتیا، سورامپالم، آندرا پرادش، هند
۲ .گروه مهندسی عمران، دانشکده مهندسی، دانشگاه نورترن بوردر، عرعر، عربستان سعودی
۳ .گروه مهندسی برق و کامپیوتر، موسسه ملی فناوری، کالج آساهیکاوا، آساهیکاوا، ژاپن
۴. گروه مهندسی معدن، دانشگاه آکسوم، آکسوم، اتیوپی

اطلاعات مقاله

تاریخ ارسال: ۲۰۲۵/۰۲/۲۴
تاریخ داوری: ۲۰۲۵/۰۵/۱۲
تاریخ پذیرش: ۲۰۲۵/۰۶/۲۰

DOI: 10.22044/jme.2025.15823.3043

کلمات کلیدی

ارزش حرارتی ناخالص
رگرسیون یادگیری ماشین
مدل‌های رگرسیون قوی
مجموعه داده‌های زغال‌سنگ آسیا و اقیانوسیه
RANSAC

چکیده

کیفیت زغال سنگ عمدتاً توسط ارزش حرارتی ناخالص (GCV) آن تعیین می‌شود که مستقیماً بر ارزش اقتصادی آن تأثیر می‌گذارد. فرمول‌های تجربی سنتی برای تخمین GCV، اگرچه مؤثر هستند، اما هنگام کار با مجموعه داده‌های بزرگ، ناکارآمد و پر زحمت می‌شوند. برای پرداختن به این موضوع، تکنیک‌های یادگیری ماشین (ML) جایگزین قوی برای پیش‌بینی‌های دقیق و سریع ارائه می‌دهند. این مطالعه از هفت پارامتر کیفیت زغال سنگ استفاده می‌کند. رطوبت کل (TM)، خاکستر (ASH)، مواد فرار (VM)، هیدروژن (H)، کربن (C)، نیتروژن (N) و گوگرد (S)، به عنوان متغیرهای مستقل برای توسعه مدل‌های پیش‌بینی برای GCV. چهار تکنیک رگرسیون مرسوم، یعنی رگرسیون بردار پشتیبان (SVR)، K-نزدیکترین همسایه‌ها (KNN)، جنگل تصادفی (RF) و درخت تصمیم‌گیری (DT)، همراه با دو مدل رگرسیون قوی اجماع نمونه تصادفی (RANSAC) و رگرسیون هوبر (HR) بررسی شده‌اند. این مجموعه داده‌ها شامل نمونه‌های زغال سنگ از پنج کشور آسیا و اقیانوسیه است: چین، اندونزی، کره، فیلیپین و تایلند. تحلیل عملکرد مقایسه‌ای نشان می‌دهد که مدل‌های رگرسیون قوی به طور قابل توجهی از تکنیک‌های مرسوم یادگیری ماشین بهتر عمل می‌کنند. مدل‌های RANSAC و Huber Regressor به ترتیب با مقادیر R^2 0.9941 و ۰.۹۹۵۲ به دقت پیش‌بینی بالاتری دست می‌یابند. این یافته‌ها پتانسیل رویکردهای رگرسیون قوی را برای تخمین GCV قابل اعتماد برجسته می‌کنند و ارزیابی کارآمد کیفیت زغال سنگ را در کاربردهای در مقیاس بزرگ تسهیل می‌کنند.

نویسنده مسئول مکاتبات: yowagaye@gmail.com (ی. فیشا)