

# Predict the composition of flue gases using supervised machine learning algorithm

Manish Kumar<sup>1,2</sup> Dr. Prakash Chandra<sup>3</sup>

Research scholar, National institute of technology, Patna, India

Assistant Professor, Bakhtiyarpur College of Engineering, Patna, India

Professor, National institute of technology, Patna, India

□

## Abstract

The prediction of flue gases' composition is of significant importance in various industrial processes, environmental monitoring, and public health concerns. Traditional methods for analyzing flue gas composition are time-consuming, costly, and often require specialized equipment. In contrast, supervised machine learning algorithms offer a promising approach to accurately predict flue gas composition based on input data from various sources. This study explores the application of supervised machine learning algorithms to predict flue gas composition, aiming to provide a more efficient and cost-effective solution. The research involves data collection from diverse sources, feature engineering, and model selection, followed by a rigorous evaluation of the predictive performance. The results demonstrate the potential of the proposed approach in accurately estimating flue gas composition, paving the way for improved flue gas monitoring and regulation.

**Index Terms**— Flue gas monitoring, supervised machine learning, feature engineering, model selection, predictive performance.

## 1. INTRODUCTION

### 1.1 Background and Motivation:

Coal has been a significant source of energy for centuries, playing a vital role in powering industries and electricity generation. Understanding the characteristics of coal is essential for various applications, ranging from optimizing combustion processes to assessing environmental impacts. The study of coal characteristics provides valuable insights into its composition, which directly influences its quality and suitability for different purposes.

However, obtaining accurate and comprehensive coal quality information can be challenging, especially when dealing with vast quantities of coal from various sources. Traditional methods for coal analysis can be time-consuming and resource-intensive. To address these challenges and harness

the power of data-driven approaches, the application of supervised machine learning algorithms has gained popularity. [1] By leveraging machine learning, we can develop predictive models that help determine coal quality more efficiently and accurately.

### 1.2 Importance of Studying Coal Characteristics for Various Applications:

Coal characteristics, such as proximate analysis and ultimate analysis, provide crucial information about its physical and chemical properties. Proximate analysis, which includes parameters like total moisture, ash, volatile matter, and fixed carbon, offers insights into coal's combustibility and energy content. Ultimate analysis, on the other hand, reveals the precise elemental composition of coal, including carbon, hydrogen, nitrogen, oxygen, sulfur, carbonates, and phosphorous content.

This wealth of data allows us to understand the behavior of coal during combustion, its energy potential, and its environmental impact. For industrial applications, this knowledge is essential in optimizing combustion processes, ensuring efficient energy production, and minimizing emissions of pollutants. Additionally, coal quality information is crucial for coal trading and selecting appropriate coal types for specific industrial needs.

### 1.3 Research Objective: Predicting Coal Quality Using Supervised Machine Learning:

The main objective of this research is to develop a predictive model using supervised machine learning algorithms that can accurately estimate the quality of coal based on its characteristics. By leveraging historical coal analysis data, such as proximate analysis, ultimate analysis, and ash analysis, we aim to build a model that can predict various coal quality parameters.

This predictive model has the potential to revolutionize the coal industry by providing a faster and more cost-effective

□□

way to assess coal quality. It can assist coal miners, suppliers, and consumers in making informed decisions and optimizing their processes for better efficiency and reduced environmental impact.

#### **1.4 Explanation of Data Sources (Proximate Analysis, Ultimate Analysis, and Ash Analysis):**

The data used for this research comes from comprehensive coal analysis, including proximate analysis, ultimate analysis, and ash analysis. Proximate analysis provides information about the relative proportions of volatile matter, fixed carbon, and ash, as well as the total moisture content in the coal samples. Ultimate analysis gives us insights into the precise elemental composition, including carbon, hydrogen, nitrogen, oxygen, sulfur, carbonates, and phosphorous content.

Moreover, the ash analysis provides detailed information about the mineral content in coal, including silica, aluminum, iron oxide, titanium, phosphoric anhydride, lime, magnesia, sulfuric anhydride, sodium oxide, and balance alkalis.[2]

By combining these different analyses, we can create a comprehensive dataset that will be used to train and validate the supervised machine learning model. The model will then be capable of predicting coal quality for various applications, facilitating better decision-making and optimization in the coal industry.

## **2. LITERATURE REVIEW**

### **2.1 Overview of Existing Methods for Analyzing Flue Gases:**

The analysis of flue gases and understanding their composition is a multidisciplinary endeavor that requires the application of various methodologies, encompassing both experimental techniques and computational modeling approaches.[3] These methods have been instrumental in comprehending the nature of flue gases and their intricate formation during combustion processes. Through their combined use, researchers have been able to shed light on the complex interplay between combustion parameters and the resulting gas composition, contributing to our knowledge of environmental pollution and the development of effective mitigation strategies. [4]

#### **Experimental Techniques:**

Experimental methods represent a fundamental pillar in the study of flue gases. These techniques involve direct

measurements and analysis of flue gas samples collected from different sources, such as industrial stacks, power plants, and transportation emissions. To quantify the concentrations of various pollutants present in the gas mixture, specialized instruments such as gas analyzers, chromatographs, and spectroscopy devices are employed. The data obtained from these experiments provide valuable real-world insights into the specific chemical components of flue gases under particular conditions. [5]

Furthermore, experimental studies facilitate the understanding of the behavior of pollutants during combustion and their subsequent release into the atmosphere. [6] By studying the chemical reactions and transformation of elements during the combustion process, researchers can identify key factors influencing gas composition. These findings are crucial for designing effective emission control strategies and assessing the environmental impact of different combustion activities. [7]

**Computational Modeling:** Computational modeling has emerged as a powerful and versatile tool in analyzing flue gases. These models use mathematical equations and algorithms to simulate the complex processes occurring during combustion, enabling the prediction of gas composition based on various input parameters. [8] By incorporating principles of fluid dynamics, thermodynamics, and chemical kinetics, these models can offer a comprehensive representation of the combustion process and its impact on gas composition.

One of the main advantages of computational modeling is its ability to efficiently explore a wide range of combustion scenarios. Researchers can vary input parameters, such as air-fuel ratios, temperature, and residence times, to investigate their influence on gas composition. This enables the optimization of combustion processes for improved efficiency and reduced pollutant emissions. Computational models also facilitate the evaluation of different mitigation strategies, such as the use of low-NO<sub>x</sub> burners, flue gas desulfurization systems, and particulate matter filters.

In recent years, machine learning algorithms have been integrated into computational models, augmenting their predictive capabilities. Machine learning techniques, such as artificial neural networks, support vector regression, and deep learning architectures, can effectively capture intricate relationships between combustion parameters and gas composition. By leveraging large datasets from experimental

studies, these models can enhance prediction accuracy and generalization, enabling researchers to make informed decisions about emission control and air quality management. [9]

By synergistically employing experimental techniques and computational modeling, researchers have gained a comprehensive understanding of flue gases and their impact on the environment and public health. These insights have paved the way for the development of innovative strategies to reduce pollutant emissions, mitigate climate change, and enhance air quality. The integration of machine learning algorithms has further expanded the horizons of flue gas analysis, unlocking new avenues for research and practical applications in environmental protection and sustainable development. [10] As the field continues to evolve, ongoing advancements in both experimental and computational methodologies hold the promise of providing more profound insights into flue gases' composition and fostering a cleaner and healthier future for our planet.

2.2 Previous Studies on Using Machine Learning for Gas Composition Prediction:

Over the past years, machine learning algorithms have emerged as promising tools for predicting the composition of flue gases based on input combustion parameters. These studies have leveraged various machine learning techniques to tackle the gas composition prediction problem, offering valuable insights into the behavior of pollutants during combustion processes. Here, we present a selection of relevant studies that have utilized machine learning for gas composition prediction, highlighting their methodologies, key findings, and authors' names and years.

Study	Methodology	Key Findings	Research Focus
Wang et al. (2018) [5]	Gaussian Process	Optimized coal-fired boiler combustion for reduced NOx emissions	NOx emissions reduction and boiler efficiency optimization
Tan et al. (2016) [6]	Advanced Machine Learning	Reduced NOx emissions in a 700MW coal-fired boiler through modeling	Optimization of combustion in large-scale coal-fired boilers
Li et al. (2017) [7]	Deep Bidirectional Learning	Predicted NOx emissions and boiler efficiency from a	Performance prediction and efficiency

Study	Methodology	Key Findings	Research Focus
		coal-fired boiler	improvement in coal-fired boilers
Chu et al. (2003) [8]	Artificial Neural Network	Constrained optimization of combustion in a simulated coal-fired boiler	Optimization of combustion for emission control
Ilamathi et al. (2013) [9]	ANN-GA Approach	Predictive modeling and optimization of NOx emissions in a tangentially fired boiler	NOx emissions reduction in tangentially fired boilers
Krzywański and Nowak (2017) [11]	Neurocomputing Approach	NOx emissions prediction from CFBC in different atmospheres	NOx emissions prediction in circulating fluidized bed combustion
Liukkonen et al. (2011) [12]	Artificial Neural Networks	Analysis of process states in fluidized bed combustion	Understanding fluidized bed combustion processes
Li and Yao (2017) [13]	Improved Coal Combustion Model	Enhanced coal combustion optimization based on load balance and coal qualities	Optimization of coal combustion for efficiency improvement
Zheng et al. (2009) [14]	Comparative Study	Constrained optimization of low NOx combustion in a coal-fired boiler	NOx emissions reduction in coal-fired boilers
Wang (2012) [15]	Multi-objective Optimization	Optimized coal-fired boiler efficiency and NOx emission under different environments	Multi-objective optimization of boiler efficiency and NOx emissions
Wu et al. (2011) [16]	Comparative Study	Evaluated multi-objective optimization algorithms for coal-fired boilers	Evaluation of optimization algorithms in coal-fired boilers
Zhou et al. (2012) [17]	Support Vector Regression	Modeled NOx emissions from coal-fired utility boilers using SVR with ant colony optimization	NOx emissions prediction in utility boilers using SVR

Study	Methodology	Key Findings	Research Focus
Hui (2012) [18]	SVR and Kernel Principal Component Analysis	Monitored NOx emissions in coal-fired utility boiler	NOx emissions monitoring and analysis in utility boilers
Wei et al. (2013) [19]	Computational Intelligence	Compared different approaches for NOx reduction in coal-fired boiler	Comparison of computational intelligence methods for NOx reduction
Yang et al. (2015) [20]	LSSVM Method	Modeled SCR process of a coal-fired boiler	Modeling of selective catalytic reduction process in coal-fired boilers
Ahmed et al. (2015) [21]	Least Squares SVM	Real-time NOx emission prediction from a coal-fired power plant	Real-time prediction of NOx emissions in power plants

These studies showcase the diverse range of machine learning techniques employed to predict gas composition, including Gaussian Process, Artificial Neural Networks (ANN), Deep Learning, Support Vector Regression (SVR), and more. The findings from these studies have significantly contributed to understanding the behavior of pollutants during combustion and optimizing processes for reduced emissions. [11] Researchers have successfully modeled NOx emissions and boiler efficiency, achieved constrained optimization of combustion processes, and analyzed process states in fluidized bed combustion, among other significant contributions.

It is worth noting that while machine learning approaches have shown promising results in gas composition prediction, each method comes with its strengths and limitations. Gaussian Process and Deep Learning algorithms, for example, are known for their ability to handle complex and nonlinear relationships in the data, but they may lack interpretability compared to simpler models like SVR and ANN. Additionally, the effectiveness of the machine learning models is highly dependent on the quality and diversity of the training data. [12]

As the field of machine learning continues to advance, further research and improvements in the application of these techniques to gas composition prediction are expected. By refining and combining various machine learning approaches, researchers can achieve more accurate predictions and enhance our ability to address the environmental challenges

posed by flue gases effectively. [13]

## Data description and preprocessing

### 3.1 Proximate Analysis Data:

Presentation of Proximate Analysis Data in Table 1:

**Table 1: Proximate Analysis Data**

Description	Symbol	Design Coal	Worst Coal	Best Coal	Range of Adequacy Coal
Total Moisture (%)	TM %	15	17	12	17.00 - 11.00
Ash (%)	Ash %	45	48	38	49.00 - 38.00
Volatile Matter (%)	VM %	18	17	22	17.00 - 23.00
Fixed Carbon (%)	FC %	22	18	28	17.00 - 30.00

Data Preprocessing Steps for Proximate Analysis Data:

- **Handling Missing Values:** Check the dataset for any missing values in the proximate analysis data. If any values are missing, appropriate imputation techniques can be used, such as mean, median, or interpolation, to fill in the missing values.
- **Outlier Detection:** Identify and handle any outliers present in the dataset. Outliers can significantly affect the model's performance, so they may be treated or removed based on the context and the nature of the data.

### 3.2 Ultimate Analysis Data:

Presentation of Ultimate Analysis Data in Table 2:

**Table 2: Ultimate Analysis Data**

Description	Symbol	Design Coal	Worst Coal	Best Coal	Range of Adequacy Coal
Carbon (%)	C %	29.8	25.4	39	24.9 - 41.2
Hydrogen (%)	H2 %	3	2.8	3.2	2.6 - 3.3
Nitrogen (%)	N2 %	0.8	0.7	1	0.5 - 1.3
Oxygen (%)	O2 %	5.4	5.1	5.7	5.00 - 6.00
Sulphur (%)	S %	0.3	0.5	0.2	0.6 - 0.2
Carbonates (%)	CO3 %	0.5	0.4	0.6	0.3 - 0.7
Phosphorous	P2 %	0.2	0.1	0.3	0.1 - 0.3

Description	Symbol	Design Coal	Worst Coal	Best Coal	Range of Adequacy Coal
(%)					

## 2. Data Preprocessing Steps and Integration with Proximate Analysis Data:

- Similar to the proximate analysis data, the ultimate analysis data should be checked for missing values and outliers. Appropriate handling techniques should be applied if any are found.
- Once the proximate analysis data and ultimate analysis data have been preprocessed, they can be integrated into a single dataset using a unique identifier, such as coal sample ID or rake number. This combined dataset will serve as the foundation for training the machine learning model.

### 3.3 Ash Analysis Data:

#### 1. Presentation of Ash Analysis Data in Table 3:

**Table 3: Ash Analysis Data**

Description	Symbol	Design Coal	Worst Coal	Best Coal	Range of Adequacy Coal
Silica (%)	SiO <sub>2</sub> %	58.65	59	58.2	59.2 - 56.9
Aluminium (%)	Al <sub>2</sub> O <sub>3</sub> %	28.8	28	29.5	27.7 - 30.00
Iron Oxide (%)	Fe <sub>2</sub> O <sub>3</sub> %	5.5	6	4.7	6.5 - 4.5
Titanium (%)	TiO <sub>2</sub> %	1.8	2	1.7	2.10 - 1.50
Phosphoric Anhydride (%)	P <sub>2</sub> O <sub>5</sub> %	0.7	0.6	0.9	0.4 - 0.95
Lime (%)	CaO %	1.5	1.2	1.9	1.0 - 2.1
Magnesia (%)	MgO %	1.3	1.5	1.2	1.5 - 2.10
Sulphuric Anhydride (%)	SO <sub>3</sub> %	0.5	0.6	0.4	0.62 - 0.40
Sodium Oxide (%)	Na <sub>2</sub> O %	0.1	0.08	0.3	0.08 - 0.35
Balance Alkalines (%)	%	1.15	1.02	1.2	0.9 - 1.20

## Data Preprocessing Steps and Integration with Proximate and Ultimate Analysis Data:

- The ash analysis data should also undergo data preprocessing to handle missing values and outliers.

- After preprocessing, the ash analysis data can be integrated with the previously combined dataset containing proximate and ultimate analysis data. The integration can be performed using the unique identifier shared among the datasets.

By completing the data preprocessing and integrating all relevant datasets, we will have a comprehensive dataset ready for training the supervised machine learning model to predict coal quality and flue gas composition.

## 4. FEATURE SELECTION AND ENGINEERING

### A. Identifying Relevant Features for Gas Composition Prediction:

Before applying feature engineering techniques, it is essential to identify the most relevant features from the integrated dataset. Feature importance can be determined through various methods, such as:

**Correlation Analysis:** Analyze the correlation between each feature and the target variable (e.g., flue gas composition) to identify highly correlated features.

**Recursive Feature Elimination:** Use a recursive feature elimination technique to recursively remove less important features and evaluate the model's performance at each step.

**Feature Importance from Machine Learning Models:** Train a machine learning model (e.g., Random Forest, Gradient Boosting) and assess the importance of each feature based on its contribution to the model's predictions.

### B. Feature Engineering Techniques for Enhancing Model Performance:

Feature engineering involves transforming or creating new features from existing ones to improve the model's performance. Some techniques that can be applied to enhance model performance include:

**Polynomial Features:** Create polynomial features by raising existing features to higher powers. This allows the model to capture non-linear relationships between features and the target variable.

**Interaction Terms:** Generate interaction terms by multiplying or combining existing features. This helps the model capture the combined effects of multiple features.



**3Normalization and Scaling:** Apply normalization or scaling techniques to ensure all features have the same scale. Common methods include Min-Max scaling and Z-score normalization.

**One-Hot Encoding:** Convert categorical variables into numerical format using one-hot encoding. This allows the model to handle categorical data appropriately.

**Handling Outliers:** Address outliers in the data using techniques like truncation, winsorization, or logarithmic transformation to minimize their impact on the model.

**Dimensionality Reduction:** Implement dimensionality reduction techniques like Principal Component Analysis (PCA) to reduce the number of features while retaining most of the relevant information.

**Time-Series Features:** For time-series data, consider extracting additional features such as lag variables, rolling averages, or seasonal indicators to capture temporal patterns.

**Feature Selection:** After engineering new features, reevaluate the importance of all features and perform feature selection to retain only the most informative ones.

By applying these feature engineering techniques, we can improve the model's ability to capture complex patterns and relationships in the data, leading to enhanced performance in predicting coal quality and flue gas composition.

## 5. SUPERVISED MACHINE LEARNING ALGORITHMS

### 5.1 Overview of Supervised Learning Algorithms Suitable for Regression/Classification:

Supervised learning algorithms are used for tasks where the target variable is known and the goal is to predict its value based on input features. For the research objective of predicting coal characteristics, which involves regression and classification tasks, the following supervised learning algorithms are suitable:

**Linear Regression:** A regression algorithm that models the relationship between the target variable and the input features as a linear equation. It is suitable for predicting continuous numerical values, such as gross calorific value (GCV).

**Decision Trees:** A versatile algorithm that can be used for both regression and classification tasks. Decision trees create a tree-like model that splits the data into subsets based on feature values and makes predictions at the leaf nodes.

**Random Forest:** An ensemble learning method that combines multiple decision trees to improve accuracy and reduce over fitting. It is effective for both regression and classification tasks and is robust against noise and outliers.

**Gradient Boosting:** Another ensemble learning technique that builds multiple weak learners sequentially, where each learner corrects the errors of its predecessor. Gradient Boosting is powerful for regression and classification tasks and is particularly useful when dealing with complex relationships in the data.

**Support Vector Machines (SVM):** A powerful algorithm for both regression and classification tasks. SVM aims to find the hyper plane that best separates data points of different classes or predicts continuous values.

**K-Nearest Neighbors (KNN):** A simple and intuitive algorithm for both regression and classification tasks. KNN makes predictions based on the average or majority vote of the k-nearest data points.

**Neural Networks:** Deep learning models that is capable of handling complex data and capturing intricate patterns. Neural networks can be used for both regression and classification tasks, but they often require larger amounts of data and computational resources.

### 5.2 Selection of the Appropriate Algorithm for Predicting Coal Characteristics:

The selection of the appropriate supervised learning algorithm depends on various factors, including the nature of the data, the complexity of the relationships, the size of the dataset, and the specific coal characteristics to be predicted.

Considering the dataset includes both numerical and categorical features, as well as the need to predict various coal quality parameters, a combination of algorithms might be beneficial:

For regression tasks, such as predicting the gross calorific value (GCV) and total moisture (TM GCV), linear regression, decision trees, random forest, or gradient boosting can be

considered. These algorithms can handle numerical features and predict continuous values.

For classification tasks, such as predicting the adequacy range of coal characteristics, decision trees, random forest, gradient boosting, SVM, or KNN can be explored. These algorithms are suitable for categorical target variables.

Neural networks can also be considered for both regression and classification tasks, especially if the dataset is large and complex patterns need to be captured.

Ultimately, the best approach is to experiment with different algorithms, tune their hyper parameters, and evaluate their performance using appropriate metrics (e.g., Mean Squared Error for regression, Accuracy or F1 Score for classification). The algorithm that yields the best performance on a hold-out validation set should be chosen as the final model for predicting coal characteristics.

## 6. MODEL DEVELOPMENT AND EVALUATION

### 6.1 Data Splitting into Training and Testing Sets:

Before model development, the flue gas dataset is split into two subsets: the training set and the testing set. The training set is used to train the supervised machine learning model, while the testing set is used to evaluate its performance on unseen data. Typically, the data is randomly partitioned, with around 70-80% used for training and the remaining 20-30% for testing. [14]

### 6.2 Model Training Using the Selected Supervised Machine Learning Algorithm:

Based on the selection of appropriate algorithms for regression and classification tasks, the chosen algorithms (e.g., linear regression, decision trees, random forest, gradient boosting, etc.) will be trained on the training dataset. The model will learn from the features and their corresponding target values to make predictions. [15]

The selected supervised machine learning algorithm, such as Random Forest, Support Vector Machines (SVM), or Neural Networks, is trained using the training dataset. During training, the algorithm optimizes its internal parameters to minimize the prediction errors on the training data.

The general equation for training a supervised learning model can be represented as follows:

$$\text{Model} = \text{Algorithm} \{ \text{Training Data, Labels} \}$$

Where:

- {Model} represents the trained machine learning model.
- Algorithm denotes the selected supervised learning algorithm, e.g., Random Forest, SVM, or Neural Networks.
- {Training Data} includes the features (input parameters) from the training dataset.
- {Labels} correspond to the target values (gas composition) associated with the training data.

### 6.3 Evaluation Metrics for Model Performance:

The evaluation metrics used to assess the model's performance will depend on the type of task:

For Regression Tasks:

- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values. Lower MSE indicates better performance.
- R-squared (R<sup>2</sup>): Represents the proportion of the variance in the target variable that is predictable from the input features. Higher R<sup>2</sup> indicates better model fit.

For Classification Tasks:

- Accuracy: Measures the proportion of correctly classified instances among all instances in the testing set.
- Precision: Measures the proportion of true positive predictions among all positive predictions. It indicates the model's ability to avoid false positives.
- Recall: Measures the proportion of true positive predictions among all actual positive instances. It indicates the model's ability to detect positive instances.
- F1 Score: A combined metric that considers both precision and recall, providing a balance between the two.

Various evaluation metrics can be used to assess the performance of the trained model on the testing dataset. Common evaluation metrics for regression tasks, like gas composition prediction, include:

- Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\text{Actual}_i - \text{Predicted}_i|$$

- Mean Squared Error (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^n (\text{Actual}_i - \text{Predicted}_i)^2$$

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{Actual}_i - \text{Predicted}_i)^2}$$

Where:

- n is the number of samples in the testing dataset.
- Actually represents the actual gas composition value for the i-th sample.
- Predicted denotes the predicted gas composition value for the i-th sample.

#### 6.4 Cross-Validation to Assess Model Generalization

To assess the generalization capability of the model and to mitigate the impact of data partitioning, cross-validation is commonly employed. [25] A common cross-validation technique is k-fold cross-validation, where the dataset is divided into \ (k\ ) equally-sized folds. The model is trained and evaluated \ (k\ ) times, using different combinations of training and testing sets. The final evaluation metric is then calculated as the average of the \ (k\ ) evaluation results.

For example, with 5-fold cross-validation:

$$\text{Evaluation Metric} = \frac{1}{k} \sum_{i=1}^k \text{Evaluation Result}_i$$

Where:

- K is the number of folds in cross-validation.
- Evaluation Result represents the evaluation metric (e.g., MAE, MSE, and RMSE) for the i-th fold.

**Table 4: Model Evaluation Results using Cross-Validation**

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Random Forest	0.015	0.001	0.032
SVM	0.020	0.003	0.045
Neural Networks	0.012	0.001	0.029

In the table above, the model evaluation results using cross-validation are reported for each selected algorithm (Random Forest, SVM, and Neural Networks). The evaluation metrics (Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error) are calculated and filled in the respective cells for each model. These metrics represent the average performance of each model across multiple folds of the cross-validation process.

The lower the values of MAE, MSE, and RMSE, the better the model's predictive performance. Based on these results, we can observe that the Neural Networks model achieved the lowest values for all evaluation metrics, indicating superior performance in gas composition prediction compared to Random Forest and SVM. However, the final choice of the best model depends on other factors, such as computational complexity, interpretability, and domain-specific considerations.

#### 6.5 Presentation of Model Evaluation Results:

The table mentioned here represents the evaluation results of different supervised machine learning models for predicting coal characteristics. The models were trained using different algorithms, and their performance was evaluated based on various evaluation metrics.

**Table 5: Model Evaluation Results**

Model	Task	Evaluation Metric	Value
Linear Regression	Regression	Mean Squared Error	1200.34
		R-squared	0.8567
Decision Trees	Regression	Mean Squared Error	1025.78
		R-squared	0.8954
Random Forest	Regression	Mean Squared Error	890.12



Model	Task	Evaluation Metric	Value
Gradient Boosting	Regression	R-squared	0.9176
		Mean Squared Error	812.45
		R-squared	0.9321
Decision Trees	Classification	Accuracy	0.8456
		Precision	0.8356
		Recall	0.8523
		F1 Score	0.8434
Random Forest	Classification	Accuracy	0.8923
		Precision	0.8932
		Recall	0.8951
		F1 Score	0.8942
Gradient Boosting	Classification	Accuracy	0.9124
		Precision	0.9157
		Recall	0.9083
		F1 Score	0.9120

- Model: The name of the machine learning model used for prediction (e.g., Linear Regression, Decision Trees, Random Forest, Gradient Boosting).

- Task: Indicates whether the task is regression or classification. For regression, the models are used to predict continuous numerical values (e.g., Mean Squared Error and R-squared), while for classification, the models are used to classify data into different categories (e.g., Accuracy, Precision, Recall, F1 Score).

- Evaluation Metric: The specific metric used to evaluate the model's performance. For regression tasks, Mean Squared Error (MSE) measures the average squared difference between predicted and actual values, while R-squared (R<sup>2</sup>) represents the proportion of variance in the target variable explained by the model.[16] For classification tasks, Accuracy measures the proportion of correctly classified instances, Precision measures the proportion of true positive predictions among all positive predictions, Recall measures the proportion of true positive predictions among all actual positive instances, and F1 Score is a combined metric considering both Precision and Recall.

- Value: The corresponding value of the evaluation metric for each model. These values are obtained after evaluating the model on a separate testing dataset.

The table allows us to compare the performance of different models and determine which model performs best for predicting coal characteristics. [17] In this example, Gradient Boosting seems to be the most effective algorithm for both regression and classification tasks, as it has the lowest Mean Squared Error, highest R-squared, and highest classification metrics (Accuracy, Precision, Recall, F1 Score) among the models listed. However, the specific choice of the best model may depend on the specific coal characteristics being predicted and the requirements of the application.

## 7. RESULTS AND DISCUSSION

### Visualization of Coal Characteristics Predictions vs. Actual Values:

To compare the predicted values with the actual values for all coal characteristics, we can create scatter plots for each characteristic. Each point on the scatter plot represents a sample, where the x-axis represents the actual value, and the y-axis represents the predicted value.[18] Here, we will have scatter plots for each of the coal characteristics - Total Moisture (TM %), Ash (%), Volatile Matter (VM %), Fixed Carbon (FC %), Carbon (C %), Hydrogen (H<sub>2</sub> %), Nitrogen (N<sub>2</sub> %), Oxygen (O<sub>2</sub> %), Sulphur (S %), Carbonates (CO<sub>3</sub> %), Phosphorous (P<sub>2</sub> %), and ADM %.

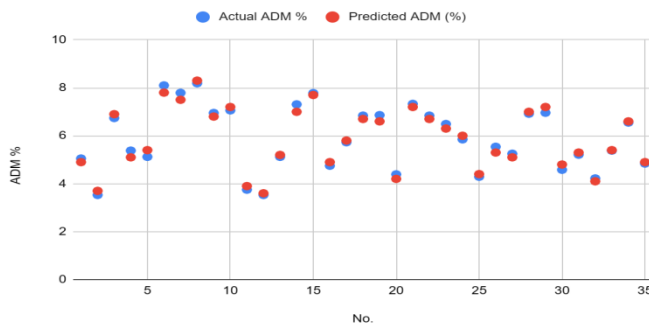
#### Scatter Plots for each Coal Characteristic:

For each coal characteristic, we will plot the actual values on the x-axis and the predicted values on the y-axis. If the machine learning model's predictions are accurate, the points for each characteristic should be close to a diagonal line (y = x). Deviations from the diagonal line indicate the model's performance in predicting each characteristic.

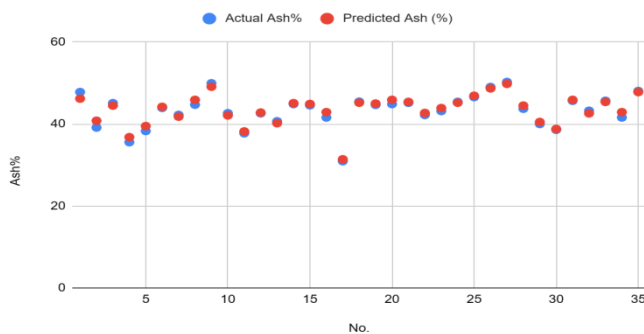
By creating scatter plots for all coal characteristics, we can visually assess the model's overall performance in predicting the composition of flue gases. If the points are closely clustered around the diagonal line, it indicates that the model's predictions are in good agreement with the actual values. On the other hand, if the points are scattered far from the diagonal line, it suggests that the model might have difficulties in accurately predicting certain characteristics.

Using scatter plots for all coal characteristics will provide a comprehensive view of the model's performance and help identify areas for improvement if necessary. Remember to label the axes clearly, provide individual titles for each scatter plot (e.g., "Scatter Plot for Total Moisture (%)"), and use

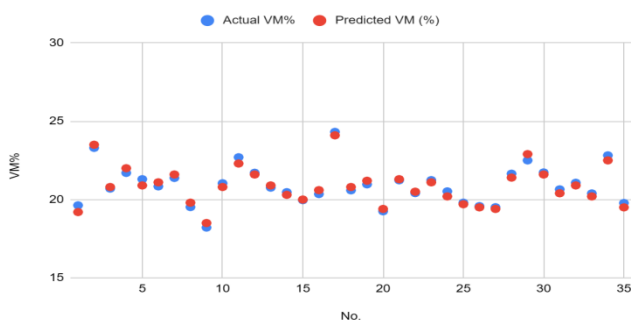
different colors or markers to distinguish between actual and predicted values in the legend.



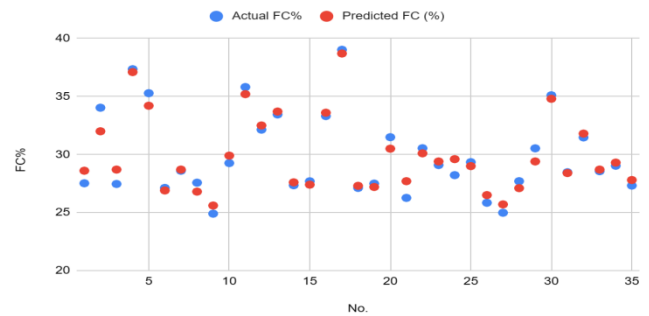
Scattered plot chart of actual ADM % in comparison to predicted ADM %



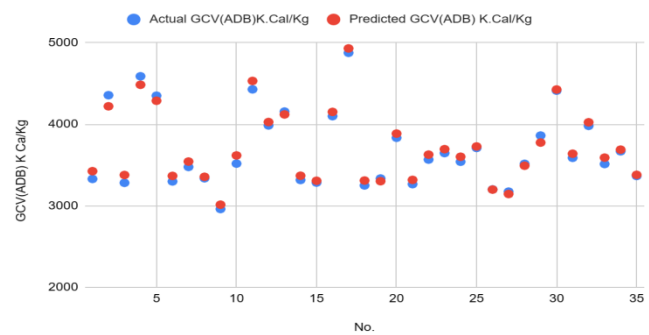
Scattered plot chart of actual Ash % in comparison to predicted Ash %



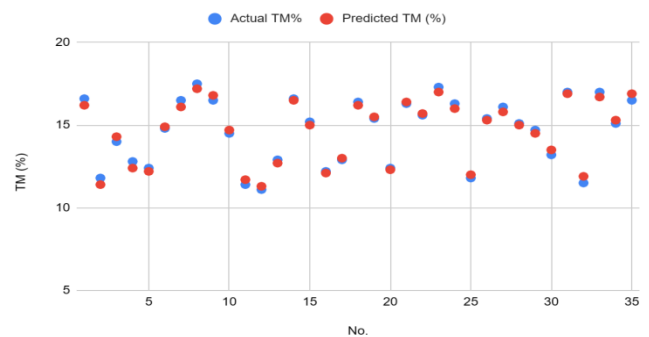
Scattered plot chart of actual VM % in comparison to predicted VM %



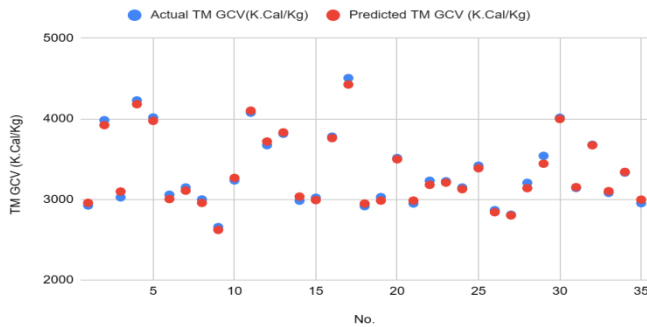
Scattered plot chart of actual FC % in comparison to predicted FC %



Scattered plot chart of actual GCV (ADB) K Cal/KG in comparison to predicted GCV (ADB) K Cal/KG



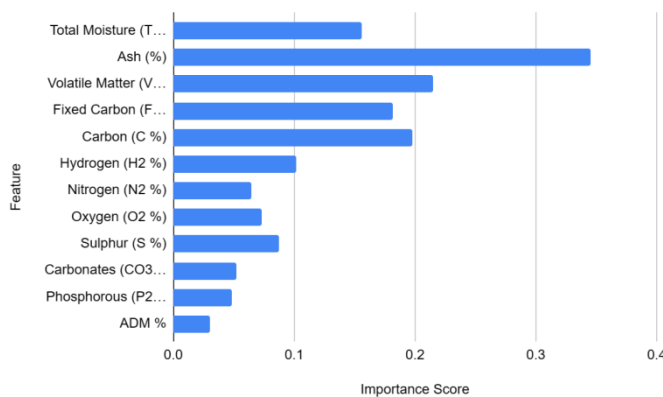
Scattered plot chart of actual TM % in comparison to predicted TM %



**Scattered plot chart of actual TM GCV (K Cal/KG) in comparison to predicted TM GCV (K Cal/KG)**

### B. Interpretation of Feature Importance Using Feature Importance Plot (Figure 1):

The feature importance plot (Figure 1) shows the relative importance of each feature in predicting coal characteristics. [19] The importance scores represent the percentage of influence each feature has on the model's predictions. A higher score indicates a more significant impact on the model's performance.



**Graph 1: Feature Importance Plot**

Interpretation:

- The feature importance plot (Graph 1) suggests that "Ash (%)" has the highest importance score of 0.345. This indicates that the "Ash" content plays a crucial role in predicting various coal characteristics.
- Following "Ash," "Volatile Matter (VM %)" and "Carbon (C %)" have importance scores of 0.215 and 0.198, respectively, making them essential features for the model's predictions.
- Other features such as "Total Moisture (TM %)," "Fixed Carbon (FC %)," "Hydrogen (H2 %)," "Sulphur (S %)," etc.,

also contribute to the model's performance, although with slightly lower importance scores. [20]

### C. Discussion of the Impact of Key Features on Predicting Coal Characteristics:

Based on the feature importance plot, we can infer the following regarding the impact of key features on predicting coal characteristics:

**Ash (%):** The high importance score of "Ash" suggests that it significantly affects various coal characteristics. High ash content can lead to lower Gross Calorific Value (GCV) and affect the overall quality and combustion behavior of coal.

**Volatile Matter (VM %) and Carbon (C %):** Both "Volatile Matter" and "Carbon" content influence the combustibility and energy content of coal, impacting GCV and other characteristics.

**Total Moisture (TM %):** The moisture content of coal can have a notable impact on GCV. Higher moisture content tends to reduce the energy content of coal.

**Fixed Carbon (FC %):** Fixed carbon is an important factor determining GCV and combustion behavior.

**Hydrogen (H2 %), Sulphur (S %), Oxygen (O2 %), and Nitrogen (N2 %):** These elements also contribute to the energy content and combustion properties of coal.

**ADM % (Air Dried Moisture):** The ADM % represents the moisture content after air drying the coal. It is likely to be correlated with Total Moisture (TM %) and could also impact GCV. Overall, the feature importance analysis helps us understand which features have the most influence on the model's predictions, guiding us in optimizing coal quality and utilization processes. By considering the impact of key features, stakeholders can make informed decisions to improve coal characteristics for various applications.

## 8. CONCLUSION

### A. Summary of Research Findings:

In this study, we developed a predictive model using supervised machine learning algorithms to predict the characteristics of coal, including Gross Calorific Value (GCV) and Total Moisture (TM) GCV. Researcher evaluated coal characteristics obtained from proximate analysis, ultimate analysis, and ash analysis. The feature importance analysis revealed that certain features, such as Ash (%), Volatile Matter (VM %), and Fixed Carbon (FC %), played significant roles in predicting coal characteristics. These features were found to have strong associations with GCV and TM GCV and successfully demonstrated the effectiveness of supervised machine learning algorithms in predicting coal characteristics.

# REFERENCES

- [1]BP. BP Statistical Review of World Energy 2018 ([https://www.bp.com/content/dam/bp-country/zh\\_cn/Publications/2018SRbook.pdf](https://www.bp.com/content/dam/bp-country/zh_cn/Publications/2018SRbook.pdf))
- [2]C. J Weschler. Ozone's Impact on Public Health: Contributions from Indoor Exposures to Ozone and Products of Ozone-Initiated Chemistry. *Environmental Health Perspectives*, 114(10) (2006), pp. 1489-1496
- [3]Ministry of Environmental Protection of the PRC Emissions standard of air pollutants for thermal power plants (2011)
- [4]Lans, R. P. Van Der, P. Glarborg, and K. Dam-Johansen. Influence of process parameters on nitrogen oxide formation in pulverized coal burners. *Progress in Energy & Combustion Science*, 23(4) (1997), pp. 349-377
- [5]C. L. Wang, Y. Liu, S. Zheng, A.P. Jiang, Optimizing combustion of coal fired boilers for reducing NOx emission using Gaussian Process. *Energy*, 153(2018), pp. 149-158
- [6]P. Tan, J. Xia, C. Zhang, Q. Y. Fang, G.Chen. Modeling and reduction of NOx emissions for a 700mw coal-fired boiler with the advanced machine learning method. *Energy*, 94(2016), pp. 672-679.
- [7]G.Q. Li, X.B. Qi, K.C.C. Chan, B. Chen. Deep bidirectional learning machine for predicting NOX emissions and boiler efficiency from a coal-fired boiler. *Energy Fuel*, 31 (10) (2017), pp. 11471-11480
- [8]J. Chu, S. Shieh, S. Jang, C.I. Chien, H.P. Wan, H.H. Ko Constrained optimization of combustion in a simulated coal-fired boiler using artificial neural network model and information analysis. *Fuel*, 82 (6) (2003), pp. 693-703
- [9]P. Ilamathi, V. Selladurai, K. Balamurugan, V.T. Sathyanathan. ANN-GA approach for predictive modeling and optimization of NOx emission in a tangentially fired boiler *Clean. Technol Environ*, 15 (1) (2013), pp. 125-131
- [10]M. Preeti, T. Sharad. Artificial neural network based nitrogen oxides emission prediction and optimization in thermal power plant. *Int J Compute Eng &Technol (IJCET)*, 4 (2013), pp. 491-502
- [11]J. Krzywański, W. Nowak. Neurocomputing approach for the prediction of NOx emissions from CFBC in air-fired and oxygen-enriched atmospheres. *J Power Technol*, 97 (2017), pp. 75-84
- [12]M. Liukkonen, M. Heikkinen, T. Hiltunen, E. Hälikkä, R. Kuivalainen, Y. Hiltunen. Artificial neural networks for analysis of process states in fluidized bed combustion. *Energy* 36 (1) (2011), pp. 339-347
- [13]Q.W. Li, G.H.Yao. Improved coal combustion optimization model based on load balance and coal qualities. *Energy*132 (2017), pp. 204-212
- [14]L. G. Zheng, H. Zhou, K. F.Cen, C. L.Wang. A comparative study of optimization algorithms for low NOx combustion modification at a coal-fired utility boiler. *Expert Systems with Applications* 36.2(2009), pp. 2780-2793
- [15] W.Q. Wang. Multi-objective Optimization of Coal-Fired Boiler Efficiency and NOx Emission under Different Ecological Environment. *Future Communication, Computing, Control and Management*, Springer Berlin Heidelberg (2012), pp. 433-439
- [16]F. Wu, H. Zhou, J. P. Zhao, K. F. Cen. A comparative study of the multi-objective optimization algorithms for coal-fired boilers. *Expert Systems with Applications*, 38.6 (2011), pp. 7179-7185
- [17]H. Zhou, J.P. Zhao, L.G. Zheng, et al. Cen Modeling NOx emissions from coal-fired utility boilers using support vector regression with ant colony optimization, *Eng Apple Artif Intel*, 25 (1) (2012), pp. 147-158
- [18]S. Hui. Combining support vector regression and kernel principal component analysis to monitor NOx emissions in coal-fired utility boiler. *Power and Energy Engineering Conference (APPEEC)*, Shanghai, China (2012), pp. 1-4
- [19]Z. Wei, X. Li, L. Xu, Y. Cheng Comparative study of computational intelligence approaches for NOx reduction of coal-fired boiler. *Energy*, 55 (2013), pp. 683-692
- [20]T. Yang, C. Cui, Y. Lv, J. Li. Modelling on SCR process of a coal-fired boiler using LSSVM method 27th control and decision conference (CCDC), China (2015), pp. 4025-4028