Journal of Applied Sciences, Information and Computing
Volume 6, Issue 1, April-May 2025
School of Mathematics and Computing, Kampala International University

ISSN: 1813-3509 https://doi.org/10.59568/JASIC-2025-6-1-15

Predictive modelling of Kerogen types using supervised machine learning algorithms: A geochemical study of the Niger Delta Basin, Nigeria

[1]Ogunleye Timothy A
[1]Department of Statistics, Faculty of Basic and Applied Sciences, Osun State University, Osogbo, Nigeria

Abstract
Kerogen classification helps hydrocarbon explorers assess the ability of rocks to produce oil and natural gas. The traditional methods of Rock-eval pyrolysis and elemental analysis still rely on previous data interpretations and often possess inaccuracies. We propose using machine learning approaches to enhance the accuracy and speed of kerogen type classification in the Niger Delta Basin utilizing geochemical data. Geochemical properties such as S1, S2, S3, Tmax, HI, OI, TOC, and PI were derived from the analyzed oil and gas well samples. Various supervised machine learning algorithms such as Random Forest, Gradient Boosting, Support Vector Machines (SVM) and Decision Trees were used to assign kerogen to Types I to IV. Performance measures were determined by computing accuracy, precision, recall and the F1 score. Ensemble methods showed the highest levels of precision and reliability among all the algorithms. It was determined that Oxygen Index, S3 and Tmax played the central role in determining kerogen quality compared to other characteristics. The algorithms (Decision Tree, Random Forest, Gradient Boosting, Ada Boosting, Bagging and Extra Trees) showed comparable results in classification precision. Applying machine learning techniques substantially improves the accuracy and objectivity of kerogen classification and exploring ensemble methods produces geoscientific results that are much more precise when compared with those obtained using conventional methods. Subsequently, the author found that only three of the features showed nearly equal percent contributions. The three highest percentage contributions came from oxygen index (17.5%), carbon dioxide generated through pyrolysis (17.3%) and temperature (16.2%). This information has been added to the current knowledge available in the field of geosciences

Keywords: Kerogen Classification, Machine Learning Algorithms, Geochemical Analysis, Hydrocarbon

1. Introduction
Kerogen, found in sedimentary rocks, is responsible for the development of petroleum and natural gas. Oil and natural gas are formed by the conversion of kerogen under various diagenesis, catagenesis and metagenesis processes. Kerogen classification is important in petroleum exploration and production since it provides insight into the characteristics of the hydrocarbon reservoir as well as its maturity stage. Traditional ways of evaluating kerogen include Rock-Eval pyrolysis, elemental analysis and vitrinite reflectance measurements (Yan, et al., 2019; Zhang, et al., 2021; Safaei-Farouji and Kadkhodaie, 2021). Several studies have determined how the classification of kerogen can affect the potential for oil and gas exploration (Zhang et al., 2021; Safaei-Farouji and Kadkhodaie, 2021). Conventional techniques for analyzing kerogen suffer from drawbacks including restricted accuracy, lengthy analysis process reliance on human judgment (Kühl, et al., 2022; Kapoor, 2024). ML applications have

brought about new avenues that optimize the accuracy and efficiency of kerogen classification (Chen, et al., 2017; Azadivash, et al., 2023).

The evaluation of source rock potential requires the use of classification system which divides kerogen into Types I, II, III and IV (Guimarães, et al., 2022). The origin of Type I kerogen from algal material produces a rich hydrogen source that leads to high levels of oil formation potential (Craddock, et al., 2020). Organisms of both planktonic and bacterial origin generate Type II kerogen which leads to oil as well as natural gas formation (Chen, et al., 2017). The main component of Type III kerogen originates from terrestrial plants which tends to generate natural gas. Type IV kerogen fails to generate hydrocarbons because it contains a low amount of hydrogen and exists in an advanced state of oxidation (Zhang, et al., 2021). Petroleum system modeling and basin analysis and exploration strategies require proper distinction between different kerogen types (Agrawal and Sharma, 2018).

The established methods for kerogen classification need laboratory work with specialized expertise and lengthily analytical processes. Rock-Eval pyrolysis generates its estimates through multiple heating steps for the evaluation of hydrocarbon production capability (Khatibi, et al., 2018). The widespread use of this measurement technique happens despite the possibility that it can be affected by sample heterogeneity and measurement errors as well as operator biases. The conventional interpretation methodology uses empirical graphic methods and statistical correlations but does not represent all natural kerogen types (BLANC-VA and M-M., 1990). The deployment of machine learning procedures marks an important advancement in operation because these methods yield both better identification precision and execution efficiency (Blackwell, et al., 2015; Gollin and Udry, 2020; Yeganeh, et al., 2023).

Multiple geochemical analyses have demonstrated successful use of support vector machines (SVM), random forests and artificial neural networks (ANN) and deep learning models because they optimize machine learning approaches (Safaei-Farouji and Kadkhodaie, 2021). Big multidimensional datasets can be analyzed through these methods to uncover hidden patterns which regular methods cannot detect (Azadivash, et al., 2023). The success rate of kerogen type prediction by ML models increases with the use of detailed geochemical datasets while reducing human opinion-based analysis along with manual data handling (Farhadi, et al., 2022; He, at al., 2022; Wei, et at., 2023).

ML classifies kerogen using faster examination techniques that preserve result accuracy to process geochemical datasets in real-time (Jooshaki, et al, 2021). Hydrocarbon exploration operations gain significant economic performance and operational success from fast decision processes according to Khang et al. (2020) and Farhadi et al. (2022) and Kühl et al. (2022). Geoscientists dedicate their time to high-level interpretation since ML-based workflows reduce their need to perform repetitive processing tasks (Azadivash, et al., 2023).

Available research demonstrates that ML technology provides successful solutions within geochemical study fields. The supervised learning techniques enable researchers to use Rock-Eval parameter data for successfully identifying source rocks (Lawal, et al., 2024). Existing data in labeled datasets supports ML models during learning processes to produce accurate predictions for new samples. The clustering algorithms available under unsupervised learning allow researchers to detect classifications in geochemical datasets that reveals information about both kerogen composition and distribution patterns (Safaei-Farouji and Kadkhodaie, 2021).

The execution of geochemical examinations enables machine learning to combine data sources. To classify kerogen through ML models all Rock-Eval data should be merged with geophysical, petrophysical and spectral information. The combined assessment technique improves both the analyses of potential source rocks and offers enhanced exploration strategies and risk evaluation methods for hydrocarbon exploration (Lawal, et al., 2024). ML delivers its key addition to geochemistry by processing unpredictable and inconsistent data characteristics (Wei, et at., 2023). The attainment of accurate classification data becomes challenging through traditional methods due to data flaws (Farhadi, et al., 2022). The application of machine learning as a geochemical analysis tool stems from

its capability to operate in data poor and hard-to-access conditions as demonstrated by Yan et al (2019), Craddock et al (2020) and Kühl et al (2022). The adoption of ML for kerogen categorization brings both advantages and technical barriers which require further advancement. The main challenge of using ML systems for classification work stems from their requirement of substantial accurate information within extensive datasets. The ability of machine learning models to predict data depends directly on the quality level as well as the diversification of the original dataset inputs. Any mistake in data preprocessing and validation procedures leads to incorrect predictions because unstable or partial data contain system biases (Lawal, et al., 2024).

Interpreting the working mechanisms of ML models remains a major concern when conducting geochemical research. The black-box operation of ML models makes it hard to track causal reasons explaining specific classifications because traditional approaches offer better understanding of geochemical connections. Research for XAI technologies continues to solve the interpretability problem while maintaining scientific integrity in ML-based geochemical assessments (Yeganeh, et al., 2023).

The growing access to open-access geochemical databases creates exceptional prospects for improving ML applications in kerogen evaluation. Academic institutions together with industries and government bodies should collaborate for standardizing ML workflows and enabling data sharing to enhance geochemical investigations. Cloud computing together with high-performance computing (HPC) resources enables the enhancement of both scalability and efficiency in ML-driven geochemical research. Kerogen classification research should be a function of the application of machine learning techniques. Real-time data acquisition through portable pyrolysis analyzers will promote smooth application of ML techniques for field-based geochemical analysis (Blackwell, et al., 2015; Gollin and Udry, 2020). The adoption of machine learning techniques signifies an absolute change in how we analyze kerogen types using geochemical methods (Kapoor, 2024). The adoption of advanced computational approaches through ML makes kerogen classification both more precise and efficient and supports larger-scale operations that overcome previous traditional evaluation weaknesses (Farhadi, et al., 2022; Yeganeh, et al., 2023). Future advancement in technology will unite ML power with geochemical investigations for the optimization of hydrocarbon exploration along with resource management. This research is intended to identify the most influential geochemical parameter for kerogen classification and also to predict the most appropriate kerogen type and generation of hydrocarbon potential from geochemical inputs. Exploring ML technique as the main aim of this research is expected to accurately improve the prediction of hydrocarbon generation potential.

2. Materials and Methods

2.1 Study Area and Geological Set-up

The Niger Delta Basin is a productive sedimentary basin situated in southern Nigeria and lies along the Gulf of Guinea continental margin. It spans from the Bight of Benin to the Calabar Flank and covers the provinces of Delta, Bayelsa, Rivers, Akwa Ibom and Cross River. Tectonic subsidence during the onset of the South Atlantic Ocean resulted in the deposition of extensive successions of deltaic and marine sediments. The Niger Delta Basin is made up of three principal sedimentary sequences. Sequences of deep marine shales (Akata Formation), continuous deposition of sandstone and shale combinations (Agbada Formation) and continental sands (Benin Formation) were deposited from the Eocene to the present day. they're underlain by the unmetamorphosed basements rocks of the Nigerian Shield. Growth faults, rollover anticlines and shale diapirs play an important role in controlling fluid movements and trap formation within the basin. The Niger Delta Basin is renowned for its high yield of crude oil due to an abundance of organic-enriched deposits.

2.2 Geochemical Significance

The Niger Delta Basin holds significant hydrocarbon exploration potential due to the presence of organic-rich sediments that serve as source rocks. The basin's lacustrine depositional environment promotes the

accumulation of high-quality kerogen, particularly Type I and Type II, which are essential for oil and gas generation. Previous studies (Gollin and Udry, 2020; Yeganeh, et al., 2023) indicate that high total organic carbon (TOC) values and favorable hydrogen index (HI) values suggest a thermally mature source rock system within the basin. Additionally, the basin's tectonic activity has created fault-related migration pathways, which are crucial for hydrocarbon accumulation. Geochemical investigations using Rock-Eval pyrolysis, biomarker analysis, and elemental composition studies have revealed promising indications of petroleum potential (Zhang, et al., 2021). The integration of machine learning techniques in analyzing geochemical data can enhance the accuracy of kerogen classification, leading to improved hydrocarbon prospectivity assessments in the region.

2.3 Geochemical Analysis of Data and Variables Used

The data for this study were obtained from geochemical samples taken in the Niger Delta Basin, Nigeria. The data includes geochemistry data obtained from the analysis of three rock samples - GLAD7, MAL05_1B_1C and MAL05_1D - using Rock-Eval pyrolysis. Furthermore, the data consist of detailed core sample images, Total Organic Carbon (TOC) amounts and results from X-ray fluorescence (XRF) analysis. We consider kerogen as our target variable of interest on which the aim and objectives were centered. This specific variable is classified into four: oil type I, oil type II, gas type III, and gas and oil type III. We also consider the following feature variables: S_1 (free hydrocarbons measured in mg HC/g rock); S_2 (Hydrocarbons generated through pyrolysis (in mg HC/g rock); S_3 (Carbon dioxide generated through pyrolysis measured in mg CO_2/g rock); T_{max} (Temperature at which the maximum release of hydrocarbons occurs measured in °C); H_I (Hydrogen Index indicating the ratio of S_2 to Total Organic Carbon (TOC), usually in mg HC/g TOC); O_I (Oxygen Index indicating the ratio of S_3 to TOC, usually in mg CO_2/g TOC); TOC (Total Organic Carbon content measured in weight %); P_I (Production Index calculated as the division of S_1 by (S_1 + S_2), indicating the stage of hydrocarbon generation); S_2/S_3 and finally S_1+S_2.

2.4 Preprocessing Steps

2.4.1 Information about the Data

All the feature variables are *float64 Dtype* while the target varible is an *object* data type. Both are 78 non-null counts with 8.7+ memory usage. The target variable, which was originally an *object* data type, was converted to *category* to ease classification.

2.4.2 EDA Explored

The use of frequency (counts), mean, standard deviation, minimum, maximum, 25% (first quartile), 50% (which is the same as median), and 75% (equivalently third quartile) were all explored as exploratory data analysis (EDA) for feature variables while the use frequency and percentage of each of the kerogen types was explored as EDA for the target variable. Pie and bar charts were also used to present information on the types of kerogen under study.

2.4.3 Missing Values

Python codes were written to check for the existence of missing values within the dataset used. No missing value was found as this paves way for further analysis.

2.4.4 Outliers

The use of z-score was explored to check whether or not there are outliers in our dataset. The result obtained from this method showed that there are outliers and this calls for data transformation.

2.4.5 Normality Assumption

We applied Shapiro-wilk statistical test to confirm the violation of normality assumption on our dataset and it's discovered that the datasets were not normally distributed. To overcome this, transformation of data was required as the best option.

2.4.6 Data Transformation

Before performing this task, we separated the datasets by defining target and feature variables. After this, we splited both the target (kerogen types) and feature variables into train and test data. 80% of our datasets were trained while the remaining 20% served as testing data. Each of the training and testing feature variables was cleaned by standardization method of data transformation.

2.5 Machine Learning Algorithms Explored

Several machine learning algorithms are available in the literature. Choosing specific machine learning

algorithm for a particular task always depends on the nature and type of the problems at hand. The current research is discussing a target variable of four categories: oil type I, oil type II, gas type III, and finally gas and oil type III. Since we have this kind of scenario, classification algorithms are suitable and therefore the most commonly used included the following: Logistic Regression, Ridge Classifier CV, K-Nearest Neighbour (KNN), Random Forest, Decision Tree, Gradient Boosting, Ada Boost, Bagging, Extra Tree, Gaussian NB, and Support Vector Machine (SVM).

2.6 Evaluation Metrics Explored

Many books and articles have been written on evaluation metrics for classification algorithms in machine learning. Only four were used here. Let's define the following parameters: m_1 = Number of positive cases which are correctly predicted; m_2 = Number of negative cases which are correctly predicted; k_1 = Number of positive cases which are incorrectly predicted; and k_2 = Number of negative cases which are incorrectly predicted. Having predefined these parameters, the four metrics can be discussed one after the other in the following subheadings.

2.6.1 Accuracy

It is a measure of percentage of correctly classified cases or instances amongst all. It is suitable for a situation when we have balanced datasets. Accuracy is mathematically expressed as follows:

$$\text{Accuarcy} = \frac{m_1 + m_2}{m_1 + m_2 + k_1 + k_2} \quad (1)$$

2.6.2 Precision

This is also known as Positive Predictive Value (PPV). It is simply a measure of percentage of correctly predicted positive cases out of all predicted positive cases. It is mathematically expressed as

$$\text{Precision} = \frac{m_1}{m_1 + k_1} \quad (2)$$

2.6.3 Recall

Recall is a measure of sensitivity which presents the percentage of actual positive cases that were correctly identified. It is used in a situation when false negatives are required to be minimized. Recall is expressed in mathematical form as follows:

$$\text{Recall} = \frac{m_1}{m_1 + k_2} \quad (3)$$

2.6.4 F1 Score

When we take the harmonic mean of both *Precision* and *Recall*, we have what is called F1 score. It is appropriately used for a situation when datasets are imbalanced. Thus, the computational approach is

$$\text{F1 Score} = 2.\left(\frac{\text{Precision } X \text{ Recall}}{\text{Precision} + \text{Recall}}\right) \quad (4)$$

3. Results and Interpretations

3.1 Descriptive Analyses

This is a statistical method of data analysis for the purpose of description of collected data and identification of its tendencies. It may start with the data range which is simply the smallest value in a set of data and the largest value in a set of data hence giving an initial look to the scope of a specific set of data. There is the measure of mean, central tendency that gives a general idea of the central location of the data by summing all the values and dividing the sum by the total count. On the other hand the median, which is the middle value when data is ordered from highest to lowest, allows the controlling for skewed distributions, which using the above formula are distorted by extreme high or low values.

Apart from central tendency, measures of variability such as standard deviations are as important in descriptive analysis. In practical terms the standard deviation calculates how much, on average, data points in the data sample deviate from the sample mean, defining the variability of the data. Small coefficient of variation implies that variance of values is small, and hence the number is closely grouped around the average while large coefficient of variation is suggestive of large variability. Thus, descriptive analysis that presupposes the use of various statistical measures helps to correctly represent vast amounts of information in the form of figures and characters; it allows to compare several sets of data, reveal their patterns and, thus, come to certain conclusions. The following table presents the results of descriptive analyses.

Table 3.1: Results of the Descriptive Analyses for Feature Variables

Variables of Interest

Statistics	S_1	S_2	S_3	T_{max}	H_I	O_I	TOC	P_I	S_2/S_3	S_1+S_2
Min.	0.18	0.19	0.28	333	0	14.694	0.34	0.0242	0.138	0.38
Max.	11.910	46.01	3.72	440	791.95	228.125	5.52	0.689	38.79	47.15
Mean	0.791	12.466	1.377	422.475	408.41	63.153	2.29	0.1384	9.64	13.26
Median	0.79	12.466	1.377	428	408	63.15	2.356	0.102	9.05	13.26
Stand Dev.	0.326	8.99	0.59	21.81	199.54	32.797	1.25	0.134	7.001	9.23

Source: System Computations (2025)

The actual values of geochemical parameters examine the differences relating to the identification of the hydrocarbon generation potential of source rocks, as presented in Table 3.1. Free hydrocarbons S1 in the range of from 0.18 to 11.91 mg HC/g rock with mean 0.791; hydrocarbons generated through pyrolysis S2 from 0.19 to 46.01 mg HC/g rock with the average of 12.466. This shows that the area has blocks of low and high prospective for hydrocarbon accumulation. It accurately demonstrated high variability in the generative potential of.Populating to S2 was approximately 8.99; thereby highlighting that the results are highly variable. S3 (CO_2 generated) is also low averaged at 1.377 mg CO_2/g rock showing low degree of oxidation of organic matter.

The thermal maturity of these rocks based on Tmax value varies between 333°C (Immature) and 440°C (Mature to post-Mature) with a mean of 422.475 °C; this shows that most of the samples are in the oil generating window. Average HI and OI values of 408.41 mg HC/g TOC and 63.153 mg CO_2/g TOC suggest that the dominant kerogen type is predominantly oil-prone, with moderate oxidation.

.

The observed TOC content varied between 0.34 % and 5.52 % with an average of 2.29 % for the samples studied indicates moderate organic richness but not all the samples qualify for minimum TOC required to generate hydrocarbons. An evaluation of the distribution of hydrocarbons obtained by GC/MS analysis of the source rock, the Production Index (PI) and S2/S3 ratios reveal the source rocks to be a mix of immature to mature source rock sediment with good generative potential across the board. This increased heterogeneity in the dataset is apparent from the present results, which demonstrate both organic richness and thermal maturity levels varying quite widely.

3.2 Explanatory Data Analysis (EDA)

Exploratory data analysis is the act of allowing the data to speak for itself and enabling the discovery of facts that exist in the data without reference to a pre-conditioned form of analysis. It is a very important step in data analysis made to enhance the understanding of the dataset's properties, structure and problematic

3.2.1 EDA for Kerogen Type

Table 3.2.1: Frequency Distribution for Kerogen Type

Classes	Frequency
Oil Type I	71
Oil Type II	03
Gas Type III	03
Gas and Oil Type III	01
Total	**78**

Source: System Computations (2025)

The percentage distribution of Kerogen Type is shown using Figures 1 and 2:

https://doi.org/10.59568/JASIC-2025-6-1-15

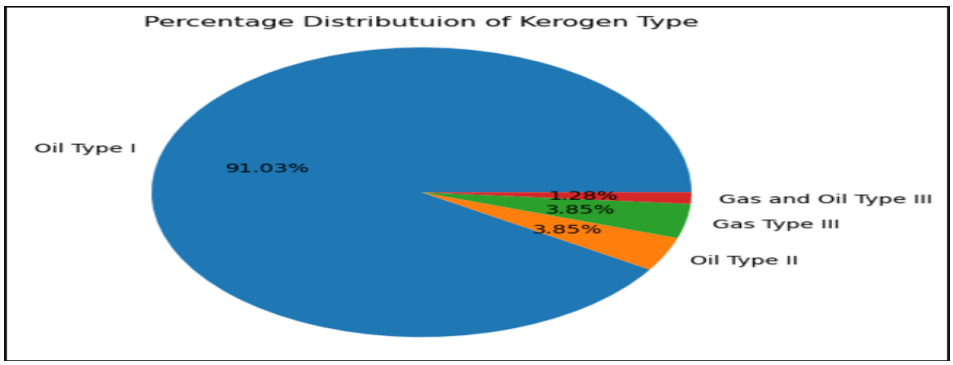

Figure 1: Pie Chart showing the Percentage Distribution of Kerogen Types

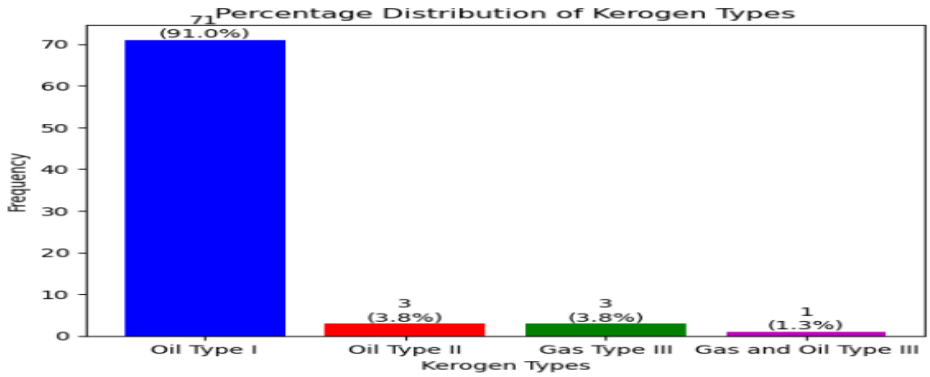

Figure 2: Bar Chart showing the Percentage Distribution of Kerogen Types

Table 3.2.1 shows a frequency distribution of the hydrocarbon type, and it was observed that Oil Type I dominates with 71 out of 78 samples, meaning that the majority of the samples possess extraordinarily good oil-generative capability. Oil Type II and Gas Type III have each been found with 3 samples (4%) derived from mixed oil-generating and gas-prone source rocks. The instances of the source rocks generating gas and oil are relatively low, and the Gas and Oil Type III is the least occurring category, with only 1 sample (1 %).

The analysis supported the discovery of a preference for oil, especially the Type I kerogen, which is generally highly mature organic material with excellent potential for generating hydrocarbons. The samples of Type III and Gas & Oil Type III, which denotes the gas-prone and mixed-generative types, are assessable, so a small number may suggest few exploration opportunities for gas within the dataset or the area being analyzed (See Figures 1 and 2).

3.3 Selection of the Best ML Algorithm

In case of choosing the most suitable algorithm for the prediction of the kerogen type, it is required to consider the data set's complexity, the size of the dataset, and relations between features. Logistic Regression and Ridge Classifier CV are easy and performs well for linearly separable set and bears a simple interpretation. Relationships between features are linear in these methods and that is why they do well when the provided relationships are not complex, but geochemical data such as kerogen type or hydrocarbon potential may not be linear, and thus require more complex modeling (Lawal, et al., 2024).

Decision tree, KNN and Gaussian Naive Bayes are selected for datasets with complex and nonlinear correlation. Looking at the strengths, Decision Trees are strong in terms of detecting non-linearities and feature interaction. However, similar to other model, it has high tendency to overfit its data and can be reduced by applying some pruning or some form of regularization. KNN is better suited to lower dimensions and less number of features and the time complexity soar, when the volumes are large. Gaussian NB is relatively fast during computations, but it has a drawback in that it does not make use of dependencies between features or attribute values in geochemical data sets (He, at al., 2022).

Most of the ensemble methods like Random Forest, Gradient Boosting, AdaBoost, Bagging, and Extra Trees are known for better prediction accuracy for geochemical properties like kerogen type, thermal maturity and hydrocarbon potential. These methods make use of multiple models and hence competency results in order to prevent over fitting and promote generalization. Random Forest and Extra Trees are impressive classifiers for high feature variability, they perform equally well and both provide feature importance. While AdaBoost and Gradient Boosting are designed for applications requiring high accuracy as each model corrects the mistakes of the previous iteration. However, this may need tuning if we do not want to over-fit the models, and to get the best for better performance (He, at al., 2022; Wei, et at., 2023).

Kerogen type classification benefits from Support Vector Machines (SVM) because it is powerful in datasets with high decision boundaries. SVM works well with linear and nonlinear separation by using kernels; its drawbacks include excessive computational costs and weak performance with large datasets. However, SVMs are not as suitable for extremely large data set because of the high computational requirements unless optimized. SVMs, however, for well-balanced medium size data sets provide high classification accuracy (Yeganeh, et al., 2023).

Specifically for these task, Random Forest and Gradient Boosting will often be the most effective sources of algorithms because of their performance with complex and large nonlinear datasets. Decision Trees and SVM are also effective for nominal and less extent data sets; while Logistic Regression and Ridge Classifier are sufficient for mainly linear issues. The last set of algorithms should therefore be selected from this last set by comparing the results of performance evaluation using some metrics such as accuracy, Precision, F1-score, and recall.

3.3.1 Predictive Machine Learning Model for Kerogen Type

Thus, machine learning techniques have a significant role in classifying kerogen type, which is critical in delineating hydrocarbon generation potential in geochemical investigations. By tabulating the relationship between geochemical data such as total organic carbon, pyrolysis data and some chemical properties, a machine learning system can accurately predict the kerogen type such as Type I, II or III over conventional techniques. Some of the classification algorithms include Random forest, Gradient boosting and support vector machine which handle large multivariate data and data with a complex relationship and hence will provide better classification rates and analysis.

In the same way the use of the machine learning in kerogen type classification enables one to recognize other relatively thin or concealed geological trends, which are hardly noticed by means of superior statistical methods. For instance, Random Forest and Gradient Boosting give feature importance that indicates the major geochemical factors influencing the type of kerogen in a formation. With the progression of the geochemistry discipline, the use of machine learning improves the model of choosing and assessing hydrocarbon prospects by multiple categories of kerogen.

Table 3.3.1: Assessment of Machine Learning Algorithms used for Kerogen Type

	Evaluation Metrics			
ML Algorithms	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression Classification	0.8750	0.292	0.333	0.311
Ridge Classifier CV Classification	0.8750	0.292	0.333	0.311
K-Nearest Neighbour Classification	0.8750	0.292	0.333	0.311
Decision Tree Classification	0.9375	0.644	0.667	0.655
Random Forest Classification	0.9375	0.644	0.667	0.655
Gradient Boosting Classification	0.9375	0.644	0.667	0.655
Ada Boost Classification	0.9375	0.644	0.667	0.655
Bagging Classification	0.9375	0.644	0.667	0.655
Extra Tree Classification	0.9375	0.644	0.667	0.655
Gaussian Naive-Bayes Classification	0.8125	0.289	0.310	0.300

 https://doi.org/10.59568/JASIC-2025-6-1-15

Support Vector Machine Classification	0.8750	0.292	0.333	0.311

Source: System Computations (2025)

The performance analysis of ML techniques for kerogen type shows the following principles: Logistic Regression, Ridge Classifier CV and KNN got an accuracy of 0.8750 but the Recall, Precision and F1 Score are comparatively low: 0.333, 0.292, 0.311. From this it can be inferred that although these models have acceptable accuracy, the trade-off between the proportion of truly positive samples that has been classified correctly (Precision) and the number of samples classified as positive that should have been classified as such (Recall) is poor. Therefore, these models may not be accurately useful for activities that demand differentiation of types of kerogen with a few number of cases of false positive results or negatives.

The models Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Bagging, and Extra Trees recorded higher accuracy, specifically, equal to 0.9375, precision – 0.644, recall – 0.667, and F1-score – 0.655. Since more steps remain for final classification, these models provide a more accurate ratio between true positives and false ones in kerogen type classification. The Gaussian Naive Bayes (NB) classifier presents lower results in all the metrics, therefore, it could be that this algorithm is less convenient in this task because of the Naive Bayes' assumption of feature independence. SVM also yields same accuracy as that of logistic regression with similar low values of precision and recall. In general, it can be concluded that the highest results in kerogen type classification are given by ensemble methods, mostly Random Forest, Gradient Boosting, and Extra Trees.

3.4 Feature Importance

Since we have known which of the classification algorithms can be used for predicting kerogen types, it is also very important to determine which of the feature variables contribute(s) most in determining kerogen type by computing their percentage contributions and visualization. The following Table 3.4 shows the results and Figure 3 speaks further on the results.

Table 3.4: Results of Computation of Feature Importance

	Variables of Interest									
Measuring Parameters	**S_1**	**S_2**	**S_3**	**T_{max}**	**H_I**	**O_I**	**TOC**	**P_I**	**S_2/S_3**	**S_1+S_2**
Importance	0.091	0.044	0.173	0.162	0.134	0.175	0.048	0.070	0.064	0.038
% Contribution	9.1	4.4	17.3	16.2	13.4	17.5	4.8	7.0	6.4	3.8

Source: System Computations (2025)

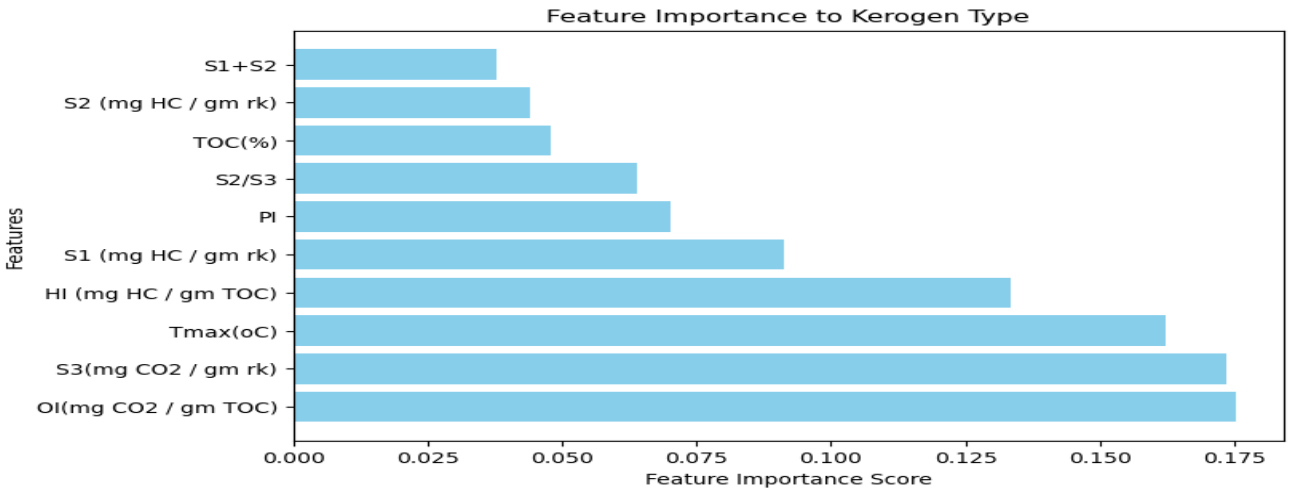

Figure 3: Bar chart showing the degree of contributions of each of the feature variables to Kerogen Types

It is evident from Table 3.4 and Figure 3 that oxygen index, carbon dioxide generated through pyrolysis and temperature are the first three geochemical parameters for determining the extent of kerogen types while others have little contributions. These results indicate that much attention should be paid to those first three parameters.

https://doi.org/10.59568/JASIC-2025-6-1-15

4. Conclusion

In the present research on geochemical investigation of kerogen types at Niger Delta Basin, Nigeria using Machine Learning Approach, we have been able to thoroughly analyze the datasets and discovered that out of eleven commonly used classification machine learning algorithms, six of them (Decision Tree, Random Forest, Gradient Boosting, Ada Boosting, Bagging, and Extra Trees) perform almost equally. This is an indication that any of them will give the same results when used for predicting the kerogen types.

Thereafter, the study was able to determine the contributions of each of the feature variables under study and concluded that only three of the features have very closed percentage contributions. These three are oxygen index (17.5%), carbon dioxide generated through pyrolysis (17.3%) and temperature (16.2%) respectively. These are the additions to the existing books of literature.

5. References

[1] Azadivash, A., Soleymani, H., Kadkhodaie, A., Yahyaee, F., & Rabbani, A. R. (2023). Petrophysical log-driven kerogen typing: unveiling the potential of hybrid machine learning. *Journal of Petroleum Exploration and Production Technology*, 13(12): 2387–2415. https://doi.org/10.1007/s13202-023-01688-1

[2] Yan, K., Zuo, Y., Yang, M., Zhou, Y., Zhang, Y., Wang, C., Song, R., Feng, R., & Feng, Y. (2019). Kerogen Pyrolysis Experiment and Hydrocarbon Generation Kinetics in the Dongpu Depression, Bohai Bay Basin, China. *Energy and Fuels*, 33(9): 8511–8521. https://doi.org/10.1021/acs.energyfuels.9b02159

[3] Zhang, J., Yang, L., Liu, J., Yan, X., Lian-jie, L., & Shen, W. (2021). Modeling Hydrocarbon Generation of Deeply Buried Type III Kerogen: A Study on Gas and Oil Potential of Lishui Sag, East China Sea Shelf Basin. *Frontiers in Earth Sciences*, 8. https://doi.org/10.3389/feart.2020.609834

[4] Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial intelligence and machine learning. *Electronic Markets, 32(4):* 2235–2244. https://doi.org/10.1007/s12525-022-00598-0

[5] Kapoor, M. (2024). Probabilistic Machine Learning and Artificial Intelligence. *Spectrum of Emerging Sciences*, 3(2): 29– 36. https://doi.org/10.55878/ses2023-3-2-5

[6] Safaei-Farouji, M., & Kadkhodaie, A. (2021). Application of ensemble machine learning methods for kerogen type estimation from petrophysical well logs. *Journal of Petroleum Sciences and Engineering*, 208: 109455–109455. https://doi.org/10.1016/j.petrol.2021.109455

[7] Chen, Z., Liu, X., & Jiang, C. (2017). Quick Evaluation of Source Rock Kerogen Kinetics Using Hydrocarbon Pyrograms from Regular Rock-Eval Analysis. *Energy and Fuels*, 31(2): 1832–1841. https://doi.org/10.1021/acs.energyfuels.6b01569

[8] Guimarães, T. T., Kupssinskü, L. S., Cardoso, M. B., Bachi, L., Aires, A. S., Koste, E. C., Spigolon, A. L. D., Gonzaga, L., & Veronez, M. R. (2022). A Nondestructive Alternative for Kerogen Type Determination in Potential Hydrocarbon Source Rocks Using Hyperspectral Data and Machine Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (15): 6418–6431. https://doi.org/10.1109/jstars.2022.3195088

[9] Craddock, P. R., Haecker, A., Bake, K. D., & Pomerantz, A. E. (2020). Universal Curves Describing the Chemical and Physical Evolution of Type II Kerogen during Thermal Maturation. *Energy and Fuels* 34(12): 15217–15233. https://doi.org/10.1021/acs.energyfuels.0c02376

[10] Agrawal, V., & Sharma, S. (2018). Improved Kerogen Models for Determining Thermal Maturity and Hydrocarbon Potential of Shale. *Scientific Report*, 8(1). https://doi.org/10.1038/s41598-018-35560-8

[11] Khatibi, S., Ostadhassan, M., Tuschel, D. D., Gentzis, T., & Carvajal-Ortiz, H. (2018). Evaluating Molecular Evolution of Kerogen by Raman Spectroscopy: Correlation with Optical Microscopy and Rock-Eval Pyrolysis. *Energies*, 11(6): 1406–1406. https://doi.org/10.3390/en11061406

[12] BLANC-VA, F. F. L. and M.-M. (1990). Interpreting Rock-Eval Pyrolysis Data Using Graphs of Pyrolizable Hydrocarbons vs. Total Organic Carbon (1). *AAPG Bulletin*, 74. https://doi.org/10.1306/0c9b238f-1710-11d7-8645000102c1865d

[13] Yeganeh, A., Johannssen, A., Chukhrova, N., Abbasi, S. A., & Pourpanah, F. (2023). Employing machine learning techniques in monitoring autocorrelated profiles. *Neural Computing and Applications*, 35(22): 16321– 16340. https://doi.org/10.1007/s00521-023-08483-3

[14] Gollin, D., & Udry, C. (2020). Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture. *Journal of Political Economy*, 129(1): 1–80. https://doi.org/10.1086/711369

[15] Blackwell, M., Honaker, J., & King, G. (2015). A Unified Approach to Measurement Error and Missing Data: Details and Extensions. *Sociological Methods and Research*, 46(3): 342–369. https://doi.org/10.1177/0049124115589052

[16] Farhadi, S., Afzal, P., Konari, M. B., Saein, L. D., & Sadeghi, B. (2022). Combination of Machine Learning Algorithms with Concentration-Area Fractal Method for Soil Geochemical Anomaly Detection in Sediment-Hosted Irankuh Pb-Zn Deposit, Central Iran. *Minerals*, 12(6): 689–689. https://doi.org/10.3390/min12060689

[17] He, Y., Zhou, Y., Wen, T., Zhang, S., Huang, F., Zou, X., Xiaogang, & Zhu, Y. (2022). A review of machine learning in geochemistry and cosmochemistry: Method improvements and applications. *Applied Geochemistry*, 140: 105273–105273. https://doi.org/10.1016/j.apgeochem.2022.105273

[18] Wei, Z., Li, X., Sun, M., Guo, R., Liu, G., Xu, Z., & Cheng, Y. (2023). Discriminating chert origins using machine-learning approaches. *Geological Journal*, 58(6): 2403–2417. https://doi.org/10.1002/gj.4753

[19] Kang, D., Wang, X., Zheng, X., & Zhao, Y. (2020). Predicting the components and types of kerogen in shale by combining machine learning with NMR spectra. *Fuel*, 290: 120006–120006. https://doi.org/10.1016/j.fuel.2020.120006

[20] Lawal, A. T., Yang, Y., He, H., & Baisa, N. L. (2024). Machine Learning in Oil and Gas Exploration: A Review. *IEEE Access*, 12: 19035–19058. https://doi.org/10.1109/access.2023.3349216

[21] Jooshaki, M., Nad, A., & Michaux, S. P. (2021). A Systematic Review on the Application of Machine Learning in Exploiting Mineralogical Data in Mining and Mineral Industry. *Minerals*, 11(8): 816–816. https://doi.org/10.3390/min11080816