Acta Geodaetica et Geophysica
https://doi.org/10.1007/s40328-024-00451-0

ORIGINAL STUDY

Machine learning assisted model based petrographic classification: a case study from Bokaro coal field

Abir Banerjee[1] · Bappa Mukherjee[2] · Kalachand Sain[2]

Received: 20 January 2024 / Accepted: 12 August 2024

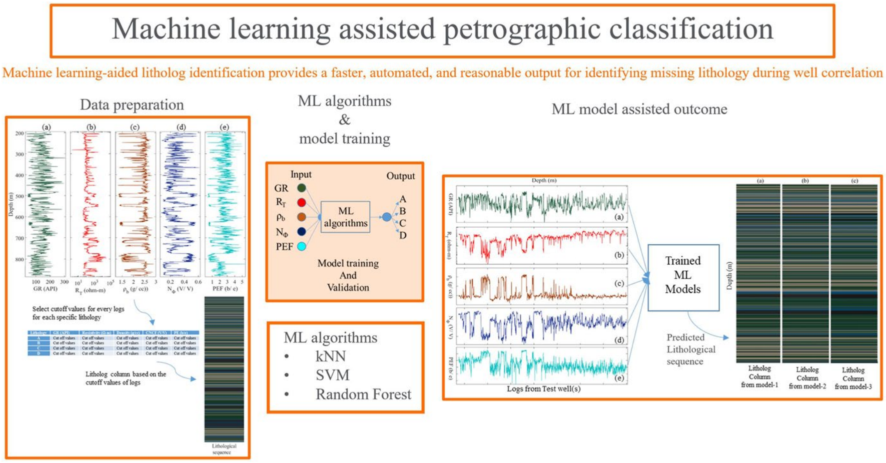
Abstract

This study applies machine learning techniques to improve petrographic classification in India's Bokaro coalfield's Barakar Formation, using conventional geophysical well logs from three wells. We analysed natural gamma ray, true resistivity, bulk density, neutron porosity, and photoelectric factor data using k-nearest neighbor (kNN), support vector machine (SVM) and random forest (RF) classifiers. A master well provided initial reference log measurement cut-off values for typical lithologies like shale, sandstone, carbonaceous shale, and coal, forming the basis of our training dataset. We assessed model accuracy using precision, recall, and F1-score metrics, finding the random forest model to be the most effective in litho-type discrimination. During the training phase, the computed overall accuracy of the predicted ML modes exceeded 89% and model accuracy hierarchy was RF>SVM>kNN. These classifiers were then applied to other well locations to predict lithological sequences, aiding in lithofacies sequence identification and potential fault extension detection. The study demonstrates the random forest model's superior precision and efficiency in lithological discrimination. Our findings enhance automated processes for identifying missing lithology during well correlation, offering valuable insights for geological interpretation in resource exploration and development. This machine learning-driven approach marks a significant advancement in subsurface geological studies.

Extended author information available on the last page of the article

Published online: 13 September 2024

Graphical abstract


Machine learning assisted petrographic classification
Machine learning-aided litholog identification provides a faster, automated, and reasonable output for identifying missing lithology during well correlation
Data preparation
ML algorithms & model training
ML model assisted outcome
(a) (b) (c) (d) (e)
Depth (m)
GR (API)
R_T (ohm-m)
ρ_b (g/cc)
N_Φ (V/V)
PEF (b/e)
Select cutoff values for every logs for each specific lithology
Litholog column based on the cutoff values of logs
Lithological sequence
Input
Output
GR
R_T
ρ_b
N_Φ
PEF
ML algorithms
A
B
C
D
Model training And Validation
ML algorithms
• kNN
• SVM
• Random Forest
Logs from Test well(s)
Trained ML Models
Predicted Lithological sequence
Litholog Column from model-1
Litholog Column from model-2
Litholog Column from model-3

Keywords Coal · Borehole logs · Lithology · Machine learning · Bokaro

1 Introduction

Understanding the inter-bedded sequence of lithologies like coal, shale, carbonaceous shale, and sandstone is important not only for well correlation but also for core sample extraction and well site decisions. Lithology prediction using machine learning (ML) eases the distinction of lithofacies where the classification depends on the geology and petrophysical properties of the specific area. Accurate description of lithological formation with highest resolution can be acquired from the resistivity image log and physical / laboratory verification of core specimen. In practice, the acquisition of advanced logs (Formation Micro-Imager Log, Borehole Televiewer log, Stratigraphic-High Resolution Dipmeter Tool) and extraction of core samples from every well is restricted due to time-taking and cost escalation. Also, manual interpretation is time-consuming process and requires expertise knowledge in the domain. Even, simultaneous interpretation of multiple logs is a tedious task, and incorporating a larger number of logs makes the interpretation more subjective (Mukherjee and Sain 2021). ML is widely used in automated interpretation of large volume of well log and seismic data and therefore integration of these datasets with ML architecture can accurately determine the lithology (e.g., Fajana et al. 2019; Dramsch 2020). ML assists energy and production (E&P) companies in evaluating reservoir parameters, tailoring drilling and completion strategies based on geological characteristics and evaluating the risk associated with individual wells.

Previous researchers have conducted numerous studies such as petrophysical quantification, lithological characterisation, geotechnical characterisation, etc., based on automated interpretation of sub-surface data from geophysical well logs (e.g., Zhou et al. 2001; Oyler et al. 2010). Previously, researchers successfully integrated statistical approaches, supervised and unsupervised classification in ML for lithofacies determination and reservoir characterisation. (e.g., Busch et al. 1987; Sun et al. 2020; Mukherjee and Sain 2021).

However, implementation of AI/ ML in coal reservoir was limited to (i) coal density estimation using Radial Basis Function neural network (RBFNN) (and Self Organized Map (SOM), (ii) coal quality estimation (Ghosh et al. 2016; Zhou and O'Brien 2016), and (iii) determination of lithological classification of coal layers using well logs (e.g., Horrocks et al. 2015; Roslin and Esterle 2016; Srinaiah et al. 2018; Maxwell et al. 2019). For example, artificial neural network (ANN), support vector machine (SVM), and Naïve Bayes classifier were applied by Horrocks et al. (2015) to identify coal and sandstone using well log data of Queensland, Australia. Another application by Maxwell et al. (2019) demonstrates the output of supervised machine learning techniques for automatic classification of thermally affected coal seams and its implication in resource estimation process. However, current advances in machine learning are stimulating interest among researchers. While the concept of using supervised classification to determine lithofacies is not new, the comparative analysis of results from various ML techniques captivates researchers' interest in achieving optimal outcomes. Recent studies by Banerjee and Chatterjee (2022) in Raniganj coal field of India deduce the application of probabilistic neural network and multi-layered neural network in mapping of pore pressure and reservoir parameters using 2D post-stack time migrated seismic data. Although previous studies demonstrated quality results, they focused solely on classifying coal and non-coal lithologies in coal fields (Horrocks et al. 2015; Zhong et al. 2020).Previous researchers have carried out lithology prediction study using several ML algorithms and come out with higher accuracy such as: Kumar et al. (2022) have carried out comparative study on lithology prediction study in Talcher coalfield India and reported all studied five classifiers (SVM, DT, RF, MLP, XGBoost) provided results with more than 88% accuracy. Prajapati et al. (2024) achieved over 85% accuracy in predicting lithology in the Cambay Basin using kNN, SVM, and ANN. Mukherjee et al. (2024) have predicted lithology with 71% accuracy in the geologically complex petroliferous Lakadong-Therria formation of the Bhogpara oil field of the Assam-Arakan Basin using various ML algorithms (kNN, SVM, decision tree (DT), RF, extreme gradient boosting (XGBoost) and artificial neural network (ANN)). Zhang et al. (2022) have reported a comparative study on lithology prediction study in Junggar Basin, China using kNN, SVM, DT, RF and XGBoost classifiers and achieved more than 88% accuracy. Thus, the estimation of lithologies in coal exploration has paramount importance, in peeping into the in-situ geological setting. Therefore, to reduce the time consumption and efficiency in processing the well log data, we have adopted ML techniques. Till date, very few detailed petrographic classification studies have been conducted in the Bokaro coalfield. This is particularly significant for large coalfields like Bokaro, where the volume of data to be analyzed is substantial. Accurately classifying the types of coal present in the Bokaro field enables the development of efficient extraction and utilization strategies. This aids in better resource management and planning for future extraction. The primary objective of the study is to apply the widely used ML architecture such as k-nearest neighbor (kNN), support vector machines (SVM), and random forest (RF) on well logs to distinguish the lithological sequences reasonably in an automated manner and to discuss the observation of fault presence in the study area.

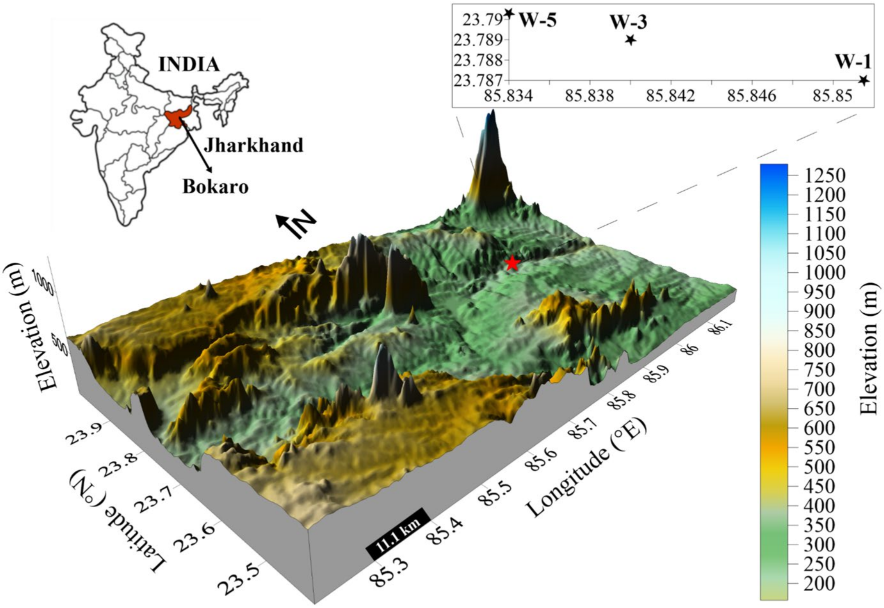
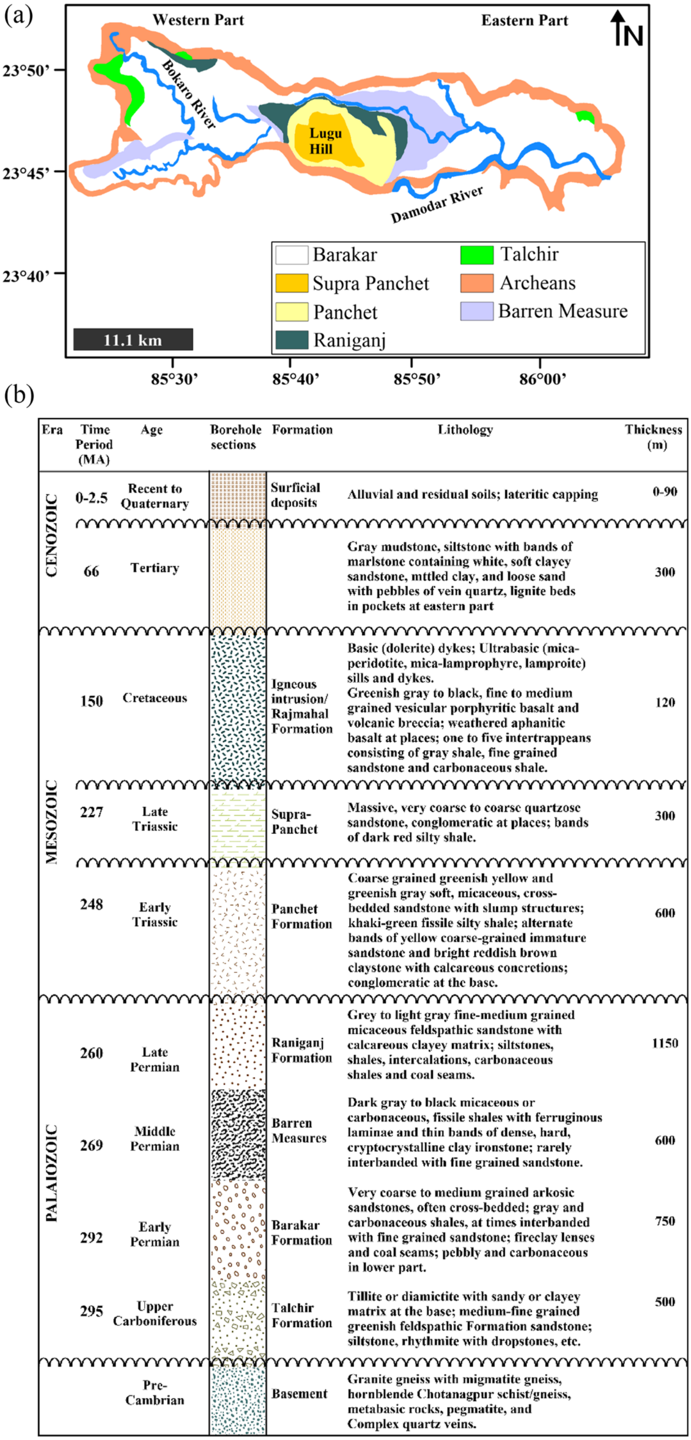


Fig. 1 Geological map of the Bokaro Coalfield and the location of wells. Where well-5 (W-5), well-3 (W-3) and well-1 (W-1) are specified by their latitude and longitude

2 Geological setting

The Bokaro Coal Field (CF) is situated within the Damodar Valley CF, which are located in the eastern region of India in the state of Jharkhand. This CF takes the form of an elongated strip of Gondwana sediments, stretching over a distance of approximately 64 kms from east to west, with a width of roughly 12 kms (Banerjee and Chatterjee 2021). The presence of Lugu Hill divides the Bokaro CF into two distinct sections: the eastern and western parts. The West Bokaro CF encompasses an area spanning 180 square kilometres, while the East Bokaro CF covers an expanse of 208 square kilometres (Banerjee and Chatterjee 2021). Figure 1 represents the elevation map of Bokaro CF and the investigated area falls in the east Bokaro CF with the respective position of the three wells. The geological formations in the Bokaro CF follow a typical sequence from the top down, including the Mahadeva, Panchet, Raniganj, Barren, Barakar, and Talchir formations. Particularly the Barakar Formation is significant within this coalfield, which dates back to the Early Permian age. Figure 2a represents the geological map depicting the surface exposure of various formations. Figure 2b shows the vertical stratigraphic succession along with age, thickness and lithology of the area. The Barakar formation predominantly consists of a variety of sediment types, including coarse to fine-grained sandstone, conglomerate, gray shale, carbonaceous shale, fine clay and coal seams. The deposition of Gondwana sediments commenced during the early stages of the Upper Carboniferous period when glacial climatic conditions prevailed. Over time, the climate transitioned to become warm and humid, a trend that persisted through the remainder of the Upper Carboniferous and the entire Permian period. With the onset of the Triassic period, a warm and dry climate took hold and continued throughout the Triassic era. Moving into the Jurassic period,



Era	Time Period (MA)	Age	Borehole sections	Formation	Lithology	Thickness (m)
CENOZOIC	0-2.5	Recent to Quaternary		Surficial deposits	Alluvial and residual soils; lateritic capping	0-90
	66	Tertiary			Gray mudstone, siltstone with bands of marlstone containing white, soft clayey sandstone, mttled clay, and loose sand with pebbles of vein quartz, lignite beds in pockets at eastern part	300
MESOZOIC	150	Cretaceous		Igneous intrusion/ Rajmahal Formation	Basic (dolerite) dykes; Ultrabasic (mica-peridotite, mica-lamprophyre, lamproite) sills and dykes. Greenish gray to black, fine to medium grained vesicular porphyritic basalt and volcanic breccia; weathered aphanitic basalt at places; one to five intertrappeans consisting of gray shale, fine grained sandstone and carbonaceous shale.	120
	227	Late Triassic		Supra-Panchet	Massive, very coarse to coarse quartzose sandstone, conglomeratic at places; bands of dark red silty shale.	300
	248	Early Triassic		Panchet Formation	Coarse grained greenish yellow and greenish gray soft, micaceous, cross-bedded sandstone with slump structures; khaki-green fissile silty shale; alternate bands of yellow coarse-grained immature sandstone and bright reddish brown claystone with calcareous concretions; conglomeratic at the base.	600
PALAIOZOIC	260	Late Permian		Raniganj Formation	Grey to light gray fine-medium grained micaceous feldspathic sandstone with calcareous clayey matrix; siltstones, shales, intercalations, carbonaceous shales and coal seams.	1150
	269	Middle Permian		Barren Measures	Dark gray to black micaceous or carbonaceous, fissile shales with ferruginous laminae and thin bands of dense, hard, cryptocrystalline clay ironstone; rarely interbanded with fine grained sandstone.	600
	292	Early Permian		Barakar Formation	Very coarse to medium grained arkosic sandstones, often cross-bedded; gray and carbonaceous shales, at times interbanded with fine grained sandstone; fireclay lenses and coal seams; pebbly and carbonaceous in lower part.	750
	295	Upper Carboniferous		Talchir Formation	Tillite or diamictite with sandy or clayey matrix at the base; medium-fine grained greenish feldspathic Formation sandstone; siltstone, rhythmite with dropstones, etc.	500
		Pre-Cambrian		Basement	Granite gneiss with migmatite gneiss, hornblende Chotanagpur schist/gneiss, metabasic rocks, pegmatite, and Complex quartz veins.	

Fig. 2 **a** Topographic formation distribution map of Bokaro coalfield (after, Paul et al. 2018), **b** Vertical stratigraphic sequence of Bokaro coalfield

the climate remained largely warm and humid. Remarkably, the geological history of the Gondwana sediment saw a period of approximately 700 million years characterized by minimal geological activity (Banerjee and Chatterjee 2021). During the Upper Carboniferous period, a geosyncline with an east–west orientation began to take shape. Within this narrow basin, the basin floors started to subside concurrently with the accumulation of an increasing amount of sediments. Notably, a rhythmic cyclic deposition pattern is prevalent in the context of fluvial sedimentation. In the Early Permian Barakar Formation, each of these cycles initiated with the deposition of coarse sand, followed by the gradual buildup of clay. Ultimately, vegetation took root in the clay, effectively completing the cycle. Consequently, each cycle is characterized by a sequence that evolves from coarse conglomerate or pebbly sandstone, progressing through coarse to medium-grained sandstone, fine-grained sandstone, shale, and culminating in coal deposition. This lithological sequence has been comprehensively examined through field outcrop studies and geophysical well log analyses of subsurface formations. The thickness of these fining upward cyclic depositions typically falls within the range of 10 to 20 m (Banerjee and Chatterjee 2021). At the top of the cyclic sedimentation sequence within the Barakar Formation lies the coal deposits. These coal deposits were influenced by an auto-cyclic process driven by the lateral migration of streams, which, in turn, was triggered by variations in subsidence within the basin. The presence of clastic deposits from rivers flowing laterally often disrupted sedimentation, leading to thinner cycles in areas where the number of cycles was higher. The target area is the East Bokaro CF, where geological data and detailed research have substantiated the five prospective coal seams (from bottom to top): "Karo-VIII," "Bermo," "Kargali Bottom," and "Kargali Top."

3 Methodology

Machine learning is a approach of data learning and simultaneously understanding the existing prevailing patterns in datasets by automated means without human interference. The new datasets used as input for training require gradual and constant learning and adapting capability to handle the unknown existing in the unused datasets. The adopted flowchart presented in Fig. 3, illustrates the log data as input. Log data of three wells were utilized for training and validation using three widely used ML architectures of supervised learning: k-nearest neighbor (kNN), support vector machines (SVM), and random forest (RF). The results of the classified lithology between these architectures were compared to determine the superior result. In methodology sub-section, we will discuss about the (i) data sets and the quality check, (ii) ML architectures: kNN, SVM, and RF and (iii) machine learning evaluation of model performance. A detail of these techniques is described in sub-section.

3.1 Data sets and preparation

The well logs data were acquired by Oil and Natural Gas Corporation Limited from the drilled CBM wells in east Bokaro CF, having 0.1524 m as sampling interval in W-1 and W-5 whereas W-3 is having a sampling interval of 0.0762 m. Well logs acquired from three boreholes (Well IDs: W-1, W-3, and W-5) drilled in the exposed Barakar formation undercrossing all lithology from 200 to 925 m consisting of caliper, natural

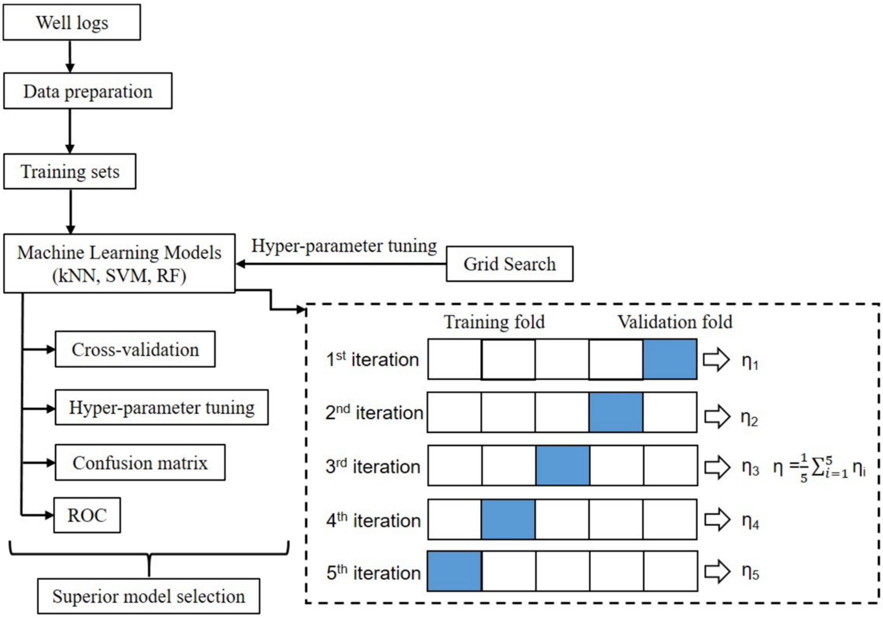


Fig. 3 The adopted flowchart of the study illustrating the steps involved

Table 1 Typical lithologies abbreviated with truncated codes and their limited range of the petrophysical information

Lithology	Code	GR (API)	Resistivity (Ω-m)	Bulk Density (g/cc)	Φ_N (V/V)	PE (barns/e)
Coal	A	38–76	500–100000	1.30–1.80	0.42–0.68	0.3–1.43
Sandstone	B	30–140	20–80	2.40–2.55	0.08–0.18	1.3–4.73
Shale	C	130–250	100–400	2.55–2.75	0.21–0.36	1.5–5.45
Carb Shale	D	150–200	500–1000	1.80–2.20	0.36–0.52	0.7–3.84

gamma-ray (GR), resistivity (R_T), bulk density (ρ_b), neutron porosity (Φ_N) and photoelectric factor (PEF). In the Barakar formation, we distinguished the lithology into four types: coal, shale, sandstone, and shaley coal. Previous study by Banerjee and Chatterjee (2021) have distinguished the cut-off range of well log responses to distinguish the litho type as coal, shale and sand stone in west Bokaro coalfield. Evaluating the previous study, we have obtained some modified cut-off log parameters in east Bokaro CF and have also classified carbonaceous shale as additional lithology along with the previous three litho type. Table 1 represents typical lithologies abbreviated with truncated codes and their limited range of petrophysical information. To train the ML models, only one well W-5 consisting of complete geology and data spectrum was used as master well for training purpose and the other two wells (W-1 and W-3) were applied with the finalized ML models for testing and validation. Figures 4, 5, 6 represents the available logs Caliper (inch), GR (API), R_T (Ohm-m), ρ_b (g/ cm^3), ϕ_N (V/ V) and PEF (barns/e) with respect to the measured depth (m) of W-1, W-3 and W-5, while Fig. 6 additionally displays model-based lithology in master well W-5. The wells W-1, W-3 and

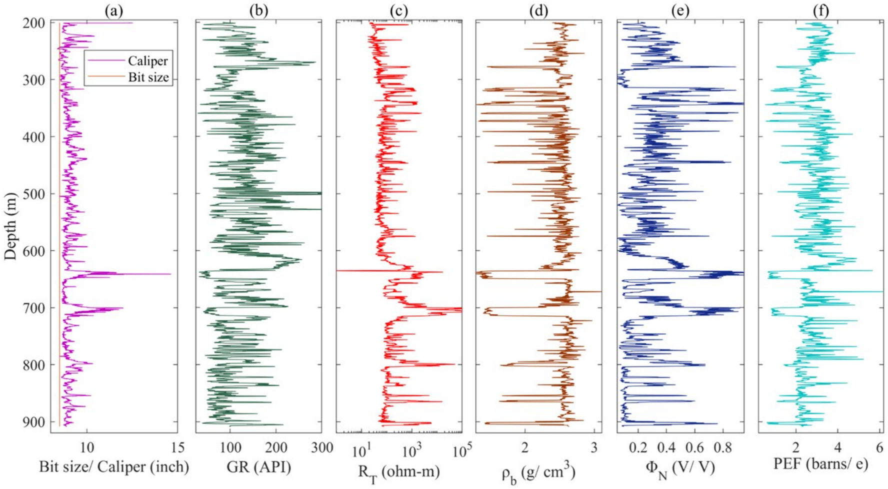


Fig. 4 Wireline log responses at W-1: **a** overlay caliper and bit size, **b** natural gamma ray (GR), **c** resistivity (R_T), **d** bulk density (ρ_b), **e** neutron porosity (Φ_N), **f** photoelectric factor (PEF)

W-5 contains respective data spanning from 200 to 908 m, 201–925 m and 200–880 m with 4650, 9503, and 4460 data points. Further, we performed well correlation studies among the studies wells to map out the continuation of the coal seam among the studies wells. Against all seam, a few of them are the objective coal seams also designated as target coal seam that has been prospective for the CBM extraction. During correlation, the objective seams named A, B, C, D and E are observed only in W-5 while W-1 contains A and B while W-3 contains C, D and E. Both the wells W-1 and W-3 have some of these seam missing. Figure 7 represents the well correlation among the studied wells (a) in W-5 we have five coal seams, namely A, B, C, D and E; (b) in W-3 we have three coal seams, namely C, D and E; (c) in W-1 we have two coal seams, namely A and B; (d) simplest coal seam correlation among the wells. Based on the well correlation study, we have seen that seam C, D and E are continuous among the W-5 and W-3, whereas, seam A and B are continuous among the W-5 and W-1, which infers the presence of possible fault in between the wells.

3.2 k-Nearest neighbor (kNN)

K-nearest neighbor (kNN) is a machine learning algorithm that is widely used for both supervised and unsupervised learning tasks. It is renowned for its simplicity and versatility in various ML applications. kNN is mostly used in classification problems and the method classifies new cases based on a similarity measure but it can also be used for regression. kNN is a non-parametric algorithm, meaning it doesn’t rely on any specific assumptions about the underlying data distribution (Fix and Hodges, 1952). The basic functioning of kNN is finding the similarity between the new and the available data by correlating it and henceforth categorizing the new data in the most similar category within the available dataset. In this method, the available data are stored and a new data point is classified based on similarity. This assures the appearance of the new data into

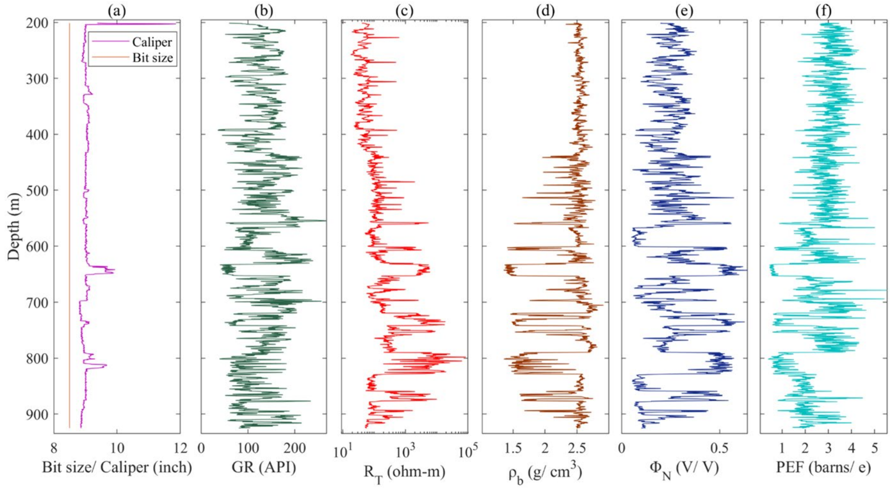


Fig. 5 Wireline log responses at W-3: **a** overlay caliper and bit size, **b** natural gamma ray (GR), **c** resistivity (R_T), **d** bulk density (ρ_b), **e** neutron porosity (Φ_N), **f** photoelectric factor (PEF)

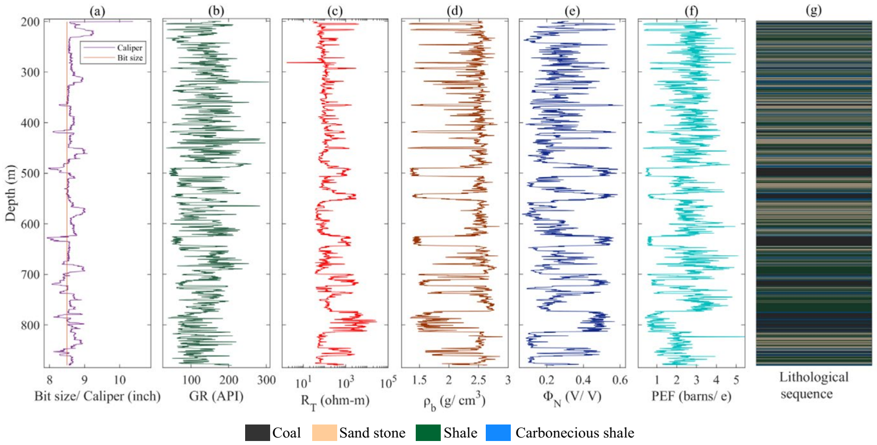


Fig. 6 Wireline log responses and model based lithological sequence at W-5: **a** overlay caliper and bit size, **b** natural gamma ray (GR), **c** resistivity (RT), **d** bulk density (ρb), **e** neutron porosity (ΦN), **f** photoelectric factor (PEF) and **g** core derived lithological sequence. Where, each lithology is represented through a specific color code

easily classifiable categories. (Cover and Hart, 1967; Azuaje, 2006) kNN is also known as lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset by classifying the data into a category much similar to the new data. In kNN, initially, the number of neighbor (k) is selected, then, the distance function of k number of neighbors was calculated. Thereafter, kNN relies on a distance function to compute and select the k nearest neighbouring data points. Among these k neighbors, it counts the number of data points in each category or class. The algorithm then assigns the new

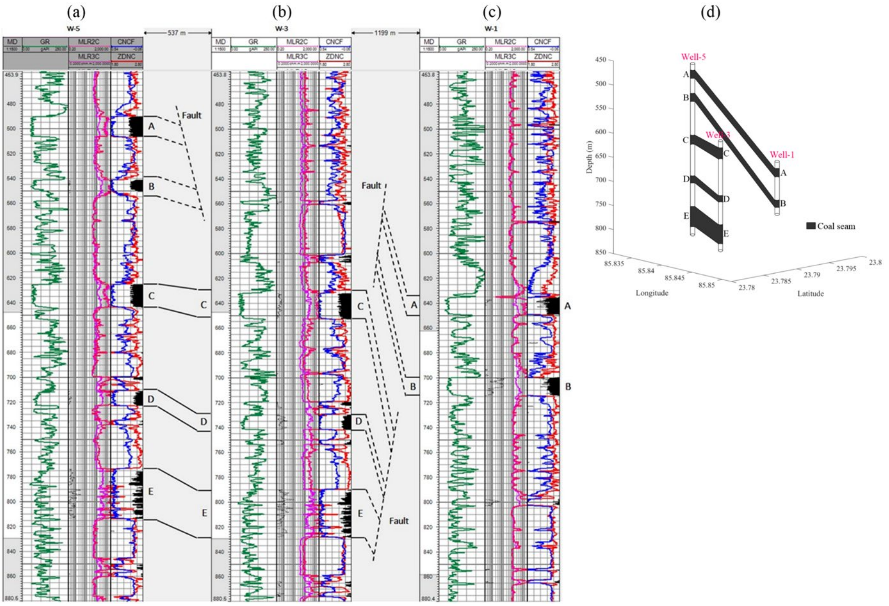


Fig. 7 Well correlation among the studied wells **a** in W-5 we have five coal seams, namely A, B, C, D and E; **b** in W-3 we have three coal seams, namely C, D, E; **c** in W-1 we have two coal seams, namely A and B; **d** simplest coal seam correlation among the wells. Based on the well correlation study, we have seen that seam C, D and E are continuous among the W-5 and W-3, whereas, seam A and B are continuous among the W-5 and W-1, which infers the presence of fault in between the wells

data point to the category with the maximum count among its neighbors, thus preparing the model for classification. The distance function can be Euclidean (ED), Manhatten (MD) and Minkowski (MI) and are represented as:

$$ED = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{1}$$

$$MD = \sum_{i=1}^{k} |x_i - y_i| \tag{2}$$

$$MI = \left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q} \tag{3}$$

where, x_i and y_i denotes are the values of the i-th feature for points *x* and *y*, respectively. x and *y* represent two points (or data instances) in an n-dimensional feature space. Each point is described by its coordinates in this space, which correspond to the values of the features.

3.3 Support vector machine (SVM)

The architecture of Support Vector Machine (SVM) falls under the supervised category and is popular and widely used for classification and regression solutions (Vapnik 1995). In this method, hyperplanes are constructed that classify datasets into separate classes. The distance between the data points and hyperplanes is measured; while the closed point lying of the dataset to the hyperplane is known as the support vector and the inter-distance amongst these vectors is called margin or street. During classification, a larger margin with respect to hyperplane is considered good whereas a smaller margin is weak for classification and therefore requires more parameters for fine-tuning.

Let's consider an occurrence of linearly separable data, y = sign (m^T x + c) (in short (m,c)).

In the datasets of n points, the training for n points is required, hence, P = {(x_1,y_1), (x_2,y_2),……. (x_n, y_n)} where yiε(1, − 1).

Hence, the Euclidean distance between the hyperplane and x_i is expressed as (Vapnik 1995):

$$r = \frac{|m^T x + c|}{||m||} \tag{4}$$

The aim of SVM as observed in Eq. (1) is to maximize the factor $||m||^{-1}$ and to provide the best optimum solution to the problem as (Vapnik 1995):

$$(m^*, c^*) = \underset{m,c,\eta_i}{\arg\min} \frac{||m||^2}{2} + C\sum_{i=1}^{n} \eta_i \tag{5}$$

In the context of hyperplane-based classification, 'm' and 'c' symbolize the normal vector and intercept of the hyperplane, respectively. Meanwhile, 'C' and ' η_i' are used to represent the penalty and slack parameters. These parameters play a crucial role in balancing the smoothness of decision boundaries while ensuring accurate classification of data points.

In cases where data points cannot be linearly separated, SVM employs a kernel technique to transform the non-linearly separable data into a higher-dimensional space, where they become linearly separable. One of the commonly used kernels for this purpose is the radial basis function (RBF) kernel. This equation can be expressed as follows:

$$K(x_i, x_j) = e^{(-\delta||x_i - x_j||^2)} \tag{6}$$

where δ = 1/2σ^2 is a controlling parameter for adjusting the hierarchy level of curvature that is required for the decision boundary.

3.4 Random forest (RF)

The Random Forest (RF) algorithm, a supervised ensemble method, was originally introduced by Ho (1995). It relies on a technique known as the random subspace method for learning. Afterward, RF method was modified by Breiman (1996) incorporating the bagging approach. In this approach, the fundamental strategy entails selecting a random subset from the original dataset. For each of these subsamples, decision trees are built to perform pattern classification. The final output is determined through a majority vote from the ensemble of decision trees within the forest. The benefit of employing the bagging method within the RF algorithm is twofold. Firstly, it leads to an improvement in overall accuracy. Secondly, it helps mitigate overfitting by leveraging the average predictions obtained from multiple decision trees, as highlighted by Breiman (2001).

The RF classification algorithm with model input dataset (D) is expressed as (Breiman 2001),

$$D = \{(x_1, y_1), \ldots.., (x_i, y_i), \ldots., (x_n, y_n)\}, y_i \in \{-1, 1\}, \quad x_i \in R^M \tag{7}$$

where ϵ(x) is the base learner, n is the number of the samples in each data subset, M is the number of the features allowed in each split inside a base learner. The model output G(x) as final classifier using majority vote is expressed as (Breiman 2001),

$$G(x) = \underset{y \in \{-1,1\}}{\arg\min} \sum_{t=1}^{T} (\in_t (x) = y) \tag{8}$$

The above discussed three ML methods, were implemented in wells. Well-5 is considered as master well which is used for the training samples, and well-1 and well-3 were considered as validation wells.

3.5 Machine learning evaluation of model performance

To evaluate the performance of the predictive machine learning models, cross-validation and hyper-parameter tuning, confusion matrix and receiver operating characteristic (ROC) were employed as assessment tools. The details of the basic performance are discussed:

3.5.1 Cross-validation and hyper-parameter tuning

The task of finding a suitable ML algorithm is conducted from a significant cross-validation tool and for achieving optimum model performance, hyper-parameters are used (Haykin 2009). In this method, a mother well is selected and the data are partitioned into training and testing sets. Subsequently, each dataset was evaluated k^{th} times using K-Fold cross-validation algorithm. In this sequence, all data are studied comprehensively for a particular ML architecture and its respective accuracy is calculated based on the score magnitude that indicates the capability of a particular ML algorithm to grasp the design against the target for a given dataset (Haykin 2009).

Another crucial step in the process of constructing machine learning models is hyper-parameter tuning, which is responsible for determining the optimal and final hyperparameters of a given ML algorithm. The GridSearchCV method accomplishes this by systematically exploring a range of parameter values on a grid and assessing the performance using K-Fold cross-validation (Hall 2016; Meshalkin et al. 2020).

3.5.2 Confusion matrix

A tabular representation in the form of confusion matrix presenting actual and predicted values. A ML classifier generates the actual and predicted values based on the statistical approach. These statistical measures can be classified as (i) true positives (TP), (ii) true negatives (TN), (iii) false positive (FP), and (iv) false negative (FN) (Navin and Pankaja 2016). The coordination between the actual and the predicted values is reflected in the diagonal component of the confusion matrix (Navin and Pankaja 2016). The confusion matrix can be better represented from normalization of the confusion matrix thereby representing the magnitude in the observed range between 0 and 1. There are two important metrics; (i) precision, and (ii) recall, which are used to evaluate the model performance. The precision and recall metrics can be expressed as (Navin and Pankaja 2016):

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

The precision or recall equilibrium of the result is provided by the F1 score also known as F measure. The equation for measuring F1 score is given as:

$$F1 \text{ score} = \frac{2 * (precision * recall)}{(precision + recall)} \tag{11}$$

3.5.3 Receiver operating characteristics (ROC)

The concept of ROC curve was used for the first time during the Second World War with an application to differentiate the noise from the radar signal and to recognize the actual data. Later, ROC was implemented to understand the performance of the predicted models (Bressan et al. 2020). ROC curve represents false-positive rate (FPR) versus true positive rate (TPR) plot. Where TPR and FPR are expressed as:

$$FPR = \frac{FP}{FP + TN} \tag{12}$$

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

4 Results

Various geophysical logs show numerous range of values. Presently, GR log varies from 29–412 API; R_T varies from 0.98 to 100,000 Ohm-m; Φ_N varies from 0.03 to 1.09 v/ v; ρ_b varies from 1.29 to 3.11 g/cm^3 and PEF varies from 0.30 to 6.19 barns/ e. Statistical analysis of the log curve was carried out by determining mean, variance, standard deviation, coefficient of skewness and kurtosis. Table 2 tabulates the statistical analysis of the training sample (prepared from the data at W-5), and data of the test wells W-3 and W-1. The target variable in this study is different lithologies derived from cut-off criteria from well log parameters, which has been validated from the previous study of geophysical logs in the same field. The Barakar formation comprises four different types of lithology namely coal, carbonaceous shale, shale and sand stone. Here each of the lithology was abbreviated with a categorical code, such as litholog class: CL represents coal, CSH represents carbonaceous shale, SH represents shale and SST represents sand stone. Input and target features from the training wells are affirmed, wherein input features were scaled to eradicate the influence of any single input feature. The selection and implementation of machine

Table 2 Statistical analysis of the training sample (prepared from the data of W-5), and data of the test wells W-3 and W-1

Training sample/ Well ID	Log responses and saturations	Minimum	Maximum	Mean	Variance	Standard deviation	Coefficient of skewness	Coefficient of kurtosis
W-5	GR	33.19	307.13	130.41	2383.7	48.82	0.22	−0.55
	R_T	1.62	30,111.5	535.61	2,326,275	1525.21	9.062	121.496
	ρ_b	1.30	2.90	2.33	0.16	0.39	−1.36	0.38
	Φ_N	0.06	0.62	0.27	0.02	0.13	0.60	−0.75
	P_e	0.30	5.44	2.34	0.92	0.96	−0.29	−0.51
W-3	GR	35.70	268.49	128.91	1680.63	40.99	0.09	−0.58
	R_T	17.71	85,510.1	932.99	$1.27*10^7$	3416.91	10.87	182.16
	ρ_b	1.35	2.91	2.41	0.11	0.34	−1.89	2.31
	Φ_N	0.05	0.64	2.26	0.02	0.13	0.58	−0.18
	P_e	0.39	5.54	2.58	0.82	0.90	−0.55	−0.26
W-1	GR	29.06	412.2	121.37	2569.62	50.69	0.71	0.59
	R_T	0.98	100,000	1198.51	$6.9*10^7$	8320.94	10.48	115.50
	ρ_b	1.29	3.11	2.45	0.10	0.32	−2.27	4.33
	Φ_N	0.03	1.09	0.28	0.03	0.18	1.17	1.19
	P_e	0.52	6.19	2.72	0.59	0.77	−0.24	0.70

learning algorithms according to the nonlinearity of the data set is imperative. Here, we presented a comparative study, by implementing three ML algorithms namely kNN, SVM, and RF to chalk up the dependency of log responses with individual lithologies form the training data set and to predict lithology from independent input features at test wells. In the present study, we have chosen leave-one-out cross-validation scheme (Zhou et al. 2020) with fivefold, beforehand model training to examine the predictive accuracy of the fitted models and able to compare all the models using the same validation scheme. In cross-validation scheme no data is reserved for testing by default setting. Presently we have chosen k = 5, which implicates entire training data (4460 data point in a single input feature) was segregated into 5 folds (divisions) of equal size (892 data points in a single feature of each fold). Now one subset was used in model validation and the remaining subsets were used to trained models. Thus, during training phase, at a time 80% is used for training and rest of the 20% validation-fold data was used to assesses model performance (validation of the models). This process was repeated for 5 times (as k = 5), so that each subset was used just once for model validation and the average error was computed for all folds. Further, by measuring the consistent accuracy from each folds we have tuned the hyperparameters using the GridSearchCV algorithm. Table 3 shows the optimum hyperparameter and the value of the parameter used in each of the models.

The performance of the ML predictive models was evaluated through the Overall Accuracy, Precision, Recall and F1-score obtained from the confusion matrix of each ML model. The confusion matrix of the ML models are depicted in Fig. 8, which illustrates the confusion matrix at training phase for studied ML algorithms: (a) kNN, (b) SVM and (c) RF. Where, class CL represents coal, class CSH represents carbonaceous shale, class SH represents shale and class SST represents sandstone. For each ML models we have computed the Precision, Recall and F1-score parameters from their confusion matrix for a lithology type (class) given in Tables 4, 5, 6**.** In Tables 4, 5, 6, the precision, recall and F1-Score of the kNN, SVM and RF models at the training phase for a particular lithology are tabulated. The overall accuracy of the predictive ML models are varying from ~89.1 to 96.4%. Thus, for each lithology obtained from the confusion matrix of all three studied ML classifiers, values of Precision vary from 85.74 to 98.51, values of Recall vary from 81.72 to 98.91 and values of F1-Score vary from 78.11 to 190.34.

The accuracy of the predictive models for each lithology type (class) are obtained from Receiver Operating Characteristic (ROC) plot. These plots are derived from of true positive rate against false positive rate of the ML models for each class, wherein the area under the receiver operating characteristic curve (AUC) is one of the user defined parameter to measure accuracy in the present analysis AUC value. The performance of the ROC spectrum can be better or worse depending on the points that lie in the ROC graph. The points lying in the graph at the top left corner indicate a better performer compared to the points that lie near the diagonal and hence are represented as poor performers of the classifier. For achieving an ideal classification of the required class using ML algorithm, the ROC curve originating from the origin gradually increases to (0, 1) and then attains a stable state. ROC curves of classifier model for lithology type (a) coal (class CL), (b) carbonaceous shale (class CSH), (c) shale (class SH), and (d) sandstone (class SST) for kNN, SVM, and RF is shown in Figs. 9, 10, 11. AUC values for lithological classes obtained from: (i) kNN model are varies from 0.96 to 1.0; (ii) SVM model are varies from 0.97 to 1.0; and (iii) RF model are varies from 0.98 to 1.0. Considering each of the ROC characteristics, Fig. 12 shows the bar graph of the accuracy of the machine learning models during the training phase. It can be seen that the RF shows better result with an accuracy of 97% compared to

Table 3 Optimum hyperparameter used in the studied models

Model	Hyperparameter (Symbol)	Parameter value
kNN	Number of nearest neighbors (n-neighbor)	3
SVM	Penalty parameter of the error term (c)	1000
	Kernel coefficient for Gaussian function (γ)	0.001
RF	Minimum number of sample at leaf node (min sample leaf)	1
	Minimum number of samples required to split an internal node (min sample split)	2
	Number of trees in the forest (n-estimators)	100
	Max depth	6

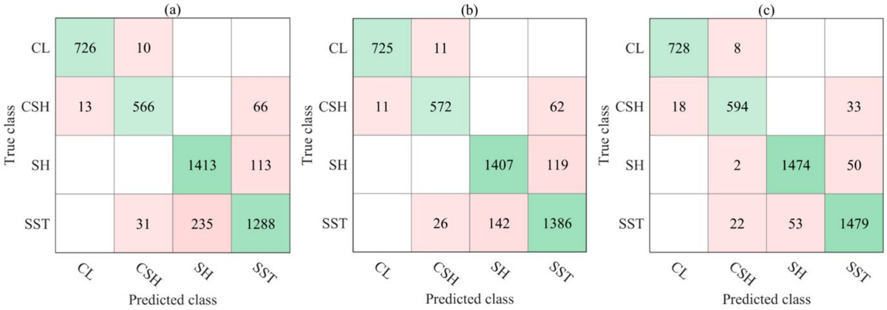


Fig. 8 Confusion matrix at training phase for studied ML algorithms: **a** kNN, **b** SVM and **c** RF. Where, class CL represents coal, class CSH represents carbonaceous shale, class SH represents shale and class SST represents sandstone

Table 4 Precision, Recall and F1-Score of the kNN models at training phase for a particular lithology

Class (Lithology type)	Precision (%)	Recall (%)	F1-Score (%)
Cl	98.24	98.64	107.695
CSH	93.24	81.72	82.00584
SH	85.74	92.59	94.16044
SST	87.79	82.88	78.11095

Table 5 Precision, Recall and F1-Score of the SVM models at training phase for a particular lithology

Class (Lithology type)	Precision (%)	Recall (%)	F1-Score (%)
Cl	98.51	98.51	103.7163
CSH	93.92	88.62	89.43897
SH	90.83	92.20	93.03995
SST	88.85	89.19	84.41129

Table 6 Precision, Recall and F1-Score of the RF models at training phase for a particular lithology

Class (Lithology type)	Precision (%)	Recall (%)	F1-Score (%)
Cl	97.59	98.91	101.778
CSH	94.89	92.09	91.27241
SH	96.53	96.59	97.27525
SST	94.69	95.17	190.34

91% and 87% accuracy for SVM and kNN. In the training phase all the ML predictive modes show accuracy in the range of 89.1–96.4%. Thus, the models are now ready to yield reasonable prediction of lithologs from unseen wireline logs of test wells.

We satisfactorily apprehended the patterns of the petrophysical parameters inherited into the studied log responses associated with numerous lithologies though the ML models. Subsequently, we fed the input features from the training well (W-5) and predicted the lithologies at test wells (W-1 and W-3). Since, we do not have core derived lithologies at the test wells, therefore, the accuracy of ML models in lithology prediction was computed using unseen log data of the test wells. It is seen that accuracy of ML models for lithology prediction varies from ~89 to 97%. Through the visual analysis of the ML predicted lithologies it is clearly observed that all models provide more or less similar prediction at the same depth interval, which itself act as a validation. Figure 13 represents the comparison between model-based lithological sequence and machine learning-assisted lithology sequences at W-5, (a) model-based lithological sequence (b) lithological sequence predicted using kNN, (c) lithological sequence predicted using SVM and (d) lithological sequence predicted using RF. Figures 14 and 15 shows the comparison between machine learning-assisted lithology sequence prediction of W-3 and W-1, using (a) kNN, (b) SVM and (c) RF. As RF model gives the best result, therefore, RF model was correlated among the wells. During correlation with ML generated lithological model (Fig. 15), it was observed that the lithological sequence in wells were not matching with each other and some of the coal seams were missing as the same was understood from Fig. 7. Hence, this was verified from the resistivity image log that fault was present in between these wells.

5 Discussion

ML algorithm is helpful in E&P industry where we can emphasize a procedure for selecting a model out of numerous models, based on their performance. In the present ML algorithm, we strive to do our best to highlight the computation power and capability to handle the data in lesser time and higher accuracy. In the comparative analysis of each ML lithological models (kNN, SVM, and RF), it has been established that for determining the lithological sequence RF model is yielding the superior result compared to the other two (SVM and kNN) (96.4% for RF; 91.2% for SVM; and 89.1 for kNN). Once a particular model is fixed, it is convenient to propagate to the other wells. As illustrated in Fig. 13, the demonstrated approach efficiently detects lithological units as thin as 0.3048 m, with machine learning models properly projecting the presence of the thin beds generated from core data. The application of ML increases the accuracy in resolving the lithology with greater confidence than manual interpretation.

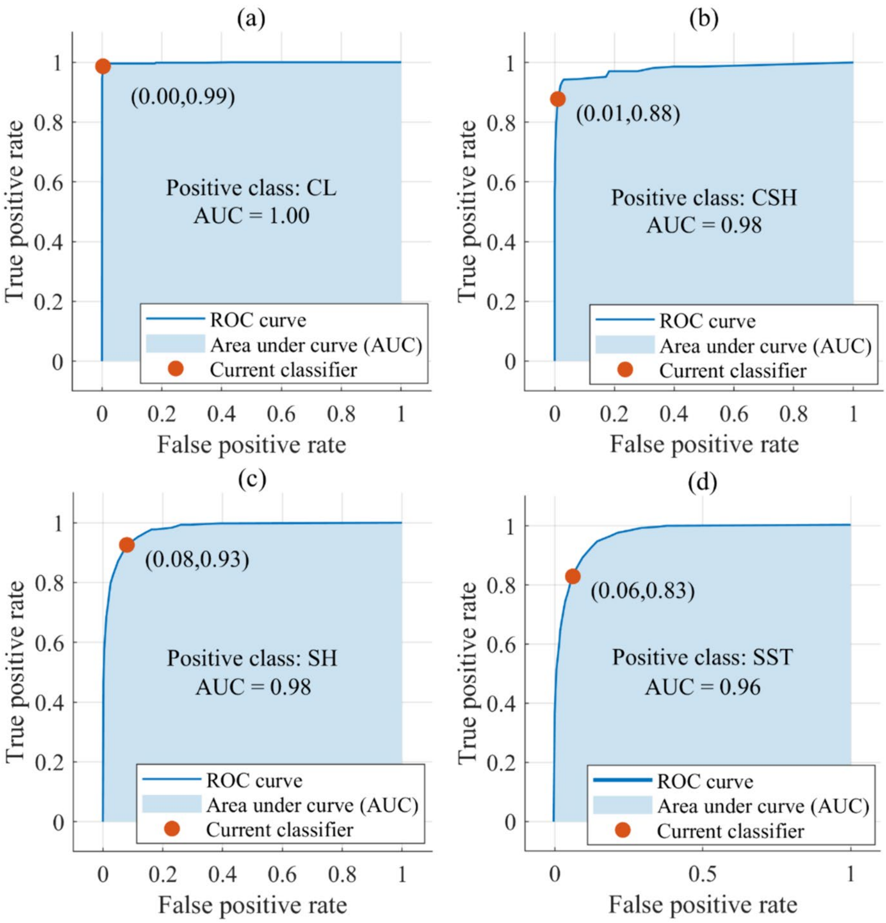


Fig. 9 Receiver Operating Characteristic (ROC) curves of k-nearest neighbor (kNN) classifier model **a** for lithology type coal (class CL), **b** for lithology type carbonaceous shale (class CSH), **c** for lithology type shale (class SH), **d** for lithology type sand stone (class SST)

In the present study we have performed with the k-fold cross-validation scheme for training and validating our predictive models to protect models against overfitting / underfitting. Bressan et al. (2020) and Kumar et al. (2022) have used k-fold cross-validation scheme for model training purpose in lithology prediction. However, other available cross-validation techniques are namely holdout, repeated random sub-sampling, stratify and resubstitution have been used by previous researchers (e.g., Refaeilzadeh et al. 2009). For hyperparameter tuning, we used Grid Search Cross-Validation (GridSearchCV) for finding optimal hyperparameters for the ML models to increase the performance of the ML models. There are many algorithms exists for hyperparameter tuning such as Random Search, Grid Search, Manual Search, Bayesian Optimizations, etc. Researchers have suggested that selection of better hyperparameter tuning method is imperative and it is depending on understating of priorities, constraints and objectives of the problem statement and data. However, we have implemented GridSearchCV method in hyperparameter tuning based on its previous applications in lithology prediction study (e.g., Kumar et al. 2022). Many researchers used

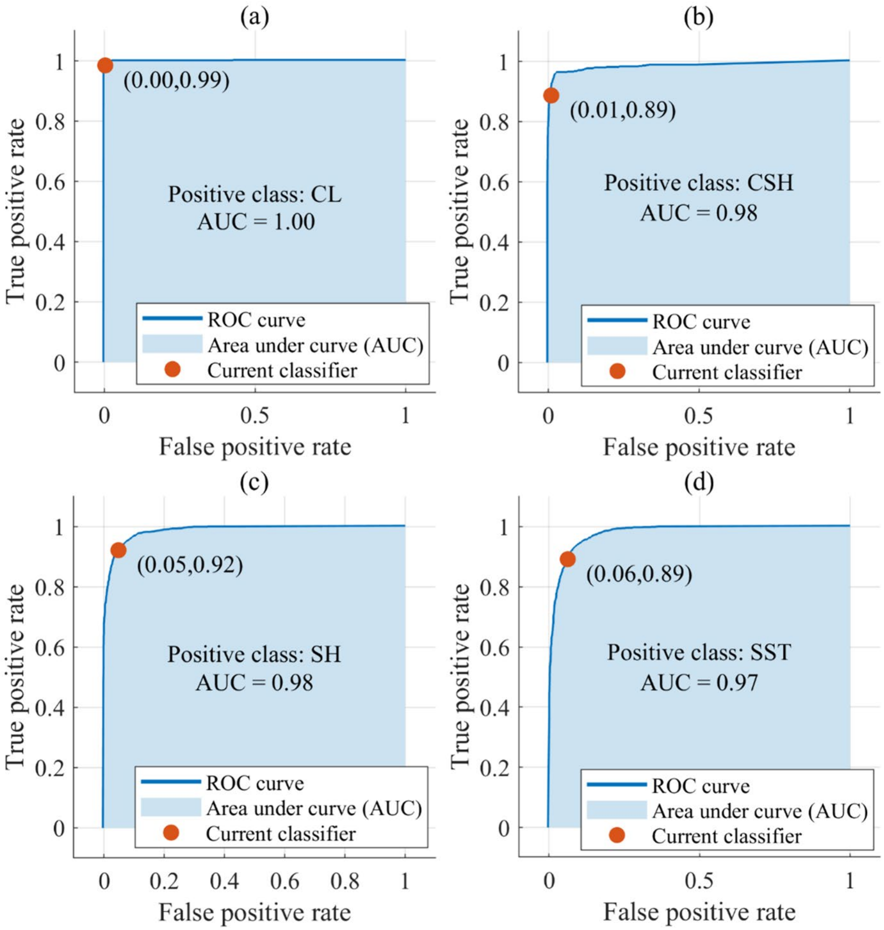


Fig. 10 Receiver Operating Characteristic (ROC) curves of support vector machine (SVM) classifier model **a** for lithology type coal (class CL), **b** for lithology type carbonaceous shale (class CSH), **c** for lithology type shale (class SH), **d** for lithology type sand stone (class SST)

standardization and normalization of data sets beforehand to use ML techniques to achieve higher accuracy of the predictive models (Jobe et al. 2018). However, many have reported the advantages and disadvantages of the use of normalized and standardized data in ML works (Xu et al. 2019). In our present case as we achieved more than 94.1% accuracy of ML models during the training stage. Hence, normalized and standardized data were not used to train ML models.

The RF can classify the lithology and help to identify the missing coal seam based on correlation. The possible fault has been anticipated which helps to build the sub-surface model. ML methodology applied in this study is beneficial in the automated determination of the lithological sequence. A model established from ML is best proven when the model matches with the subsurface litholog. Hence, the input data used for model generalization needs to be ensured that the model is trained to generalize the validation dataset for a specific area and find patterns from the training datasets. The coal seams identified in ML

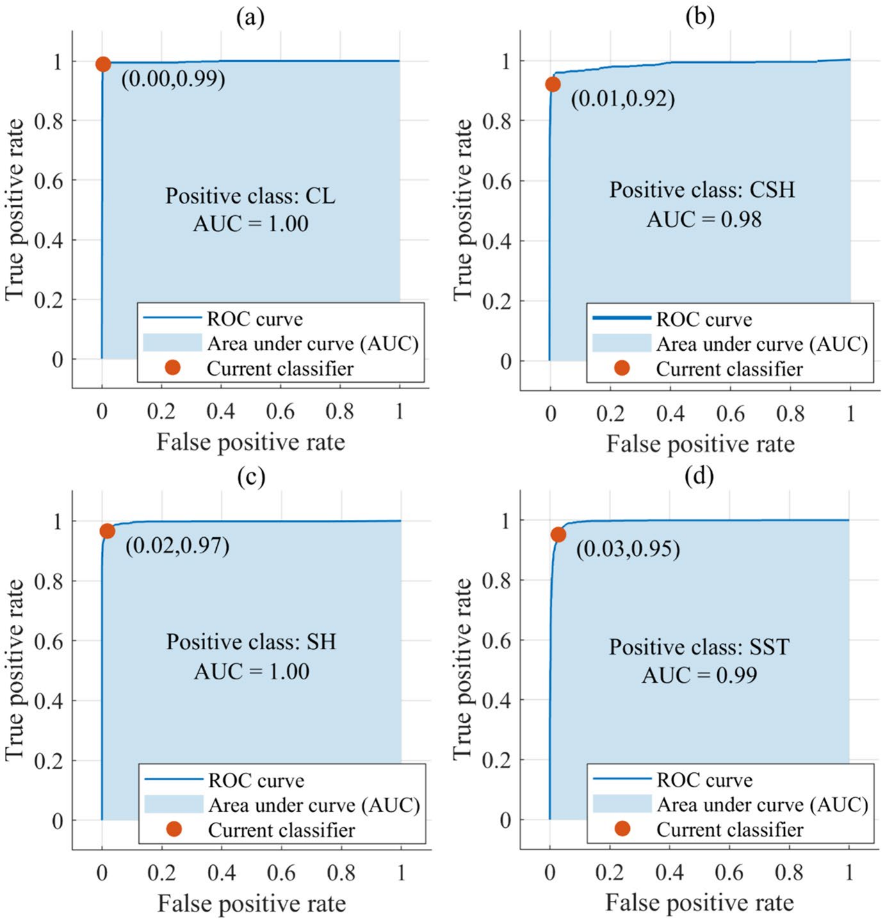


Fig. 11 Receiver Operating Characteristic (ROC) curves of random forest (RF) classifier model **a** for lithology type coal (class CL), **b** for lithology type carbonaceous shale (class CSH), **c** for lithology type shale (class SH), **d** for lithology type sand stone (class SST)

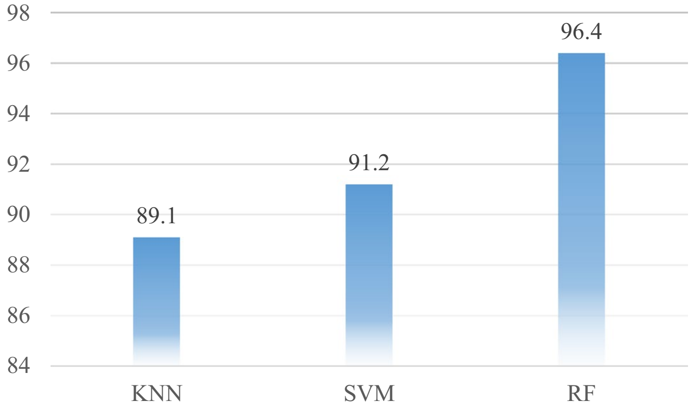


Fig. 12 Accuracy of the machine learning models during training phase

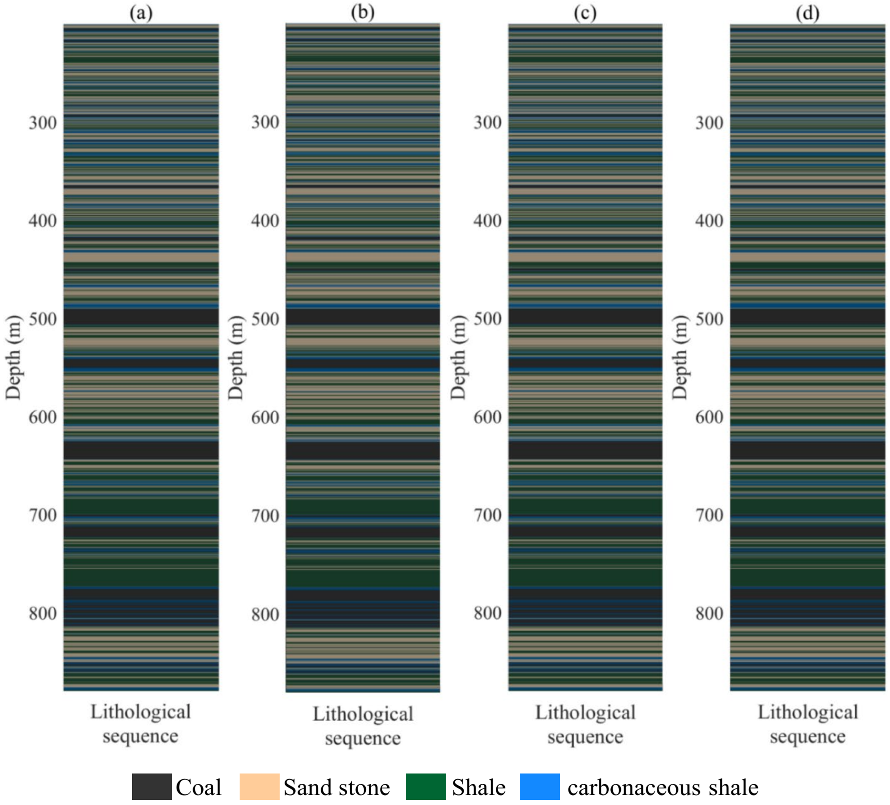


Fig. 13 Comparison between model based lithological sequence and machine learning assisted lithology sequences at W-5, **a** model based lithological sequence, **b** lithological sequence predicted using kNN, **c** lithological sequence predicted using SVM and **d** lithological sequence predicted using RF

models have been compared. Figure 16 represents the correlation of important coal seam among the RF-predicted lithologs at the studied wells (a) W-5, (b) W-3, and (c) W-1. We have also seen a few local coal seams and missing target coal seams in these three nearby wells. This could be possibly due to the existence of fault. A correlation between the wells helps to identify the possible fault presence between the wells. Subsequently, the missing coal sequence in these wells indicates the presence of fault between these wells. The result interpreted with the dip pattern in the resistivity image log in well W-3 corroborates the presence of fault. Resistivity image log records the micro-resistivity magnitude in image form of the entire borehole cross-section in clockwise direction (North-East-South-West-North). The image is presented in static and dynamic form, where static image represents the color code in constant color band throughout the borehole depth interval, whereas in dynamic image, the code changes with lithological depth interval say 5 m each, to enhance the geological features. Figure 17 represents the resistivity image log of the entire borehole cross section in clockwise direction in the form of static and dynamic image with dip track showing the dip magnitude and azimuth of the bedding. In Fig. 17a, above 567 m depth, a consistent dip of magnitude 30°–40° with azimuth of N220°–230°E is observed, however when the wellbore trajectory moves beyond this depth, an abrupt change in the dip

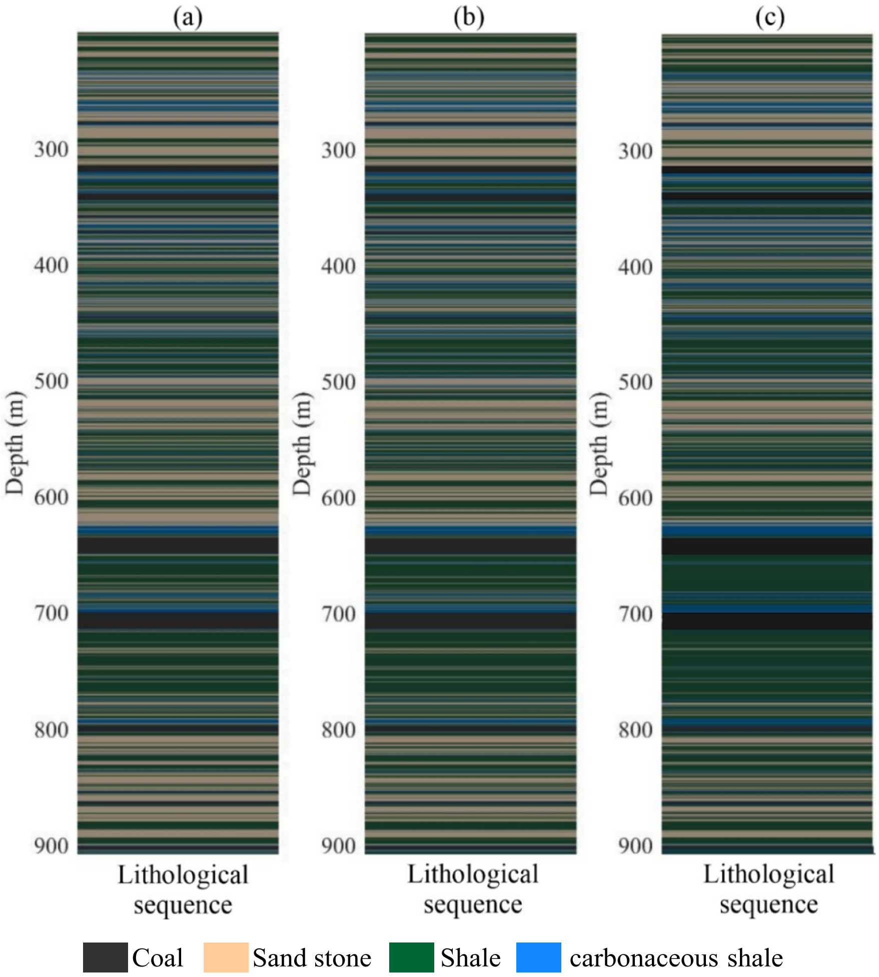


Fig. 14 Comparison between machine learning assisted lithology prediction at W-3, **a** lithological sequence predicted using kNN, **b** lithological sequence predicted using SVM and **c** lithological sequence predicted using RF

magnitude (10°–40°) and azimuth (N135°–250°E) in a scattered pattern is observed, indicating the well path through fault plane, a representation of the image from depth interval 582–585 m is shown in Fig. 17b.

The implemented ML classifiers have several merits and demerits in data analysis and prediction. The kNN classifier has the simplest mathematical background and is easy to implement. However, kNN is sensitive to the noise content of the data and requires optimal “k” values for reasonable performance. kNN is also not recommended for large and highly nonlinear data sets. SVM can effectively handle high-dimensional data set having smaller sample numbers. SVM is robust against overfitting as its diagnostic steps involve the use

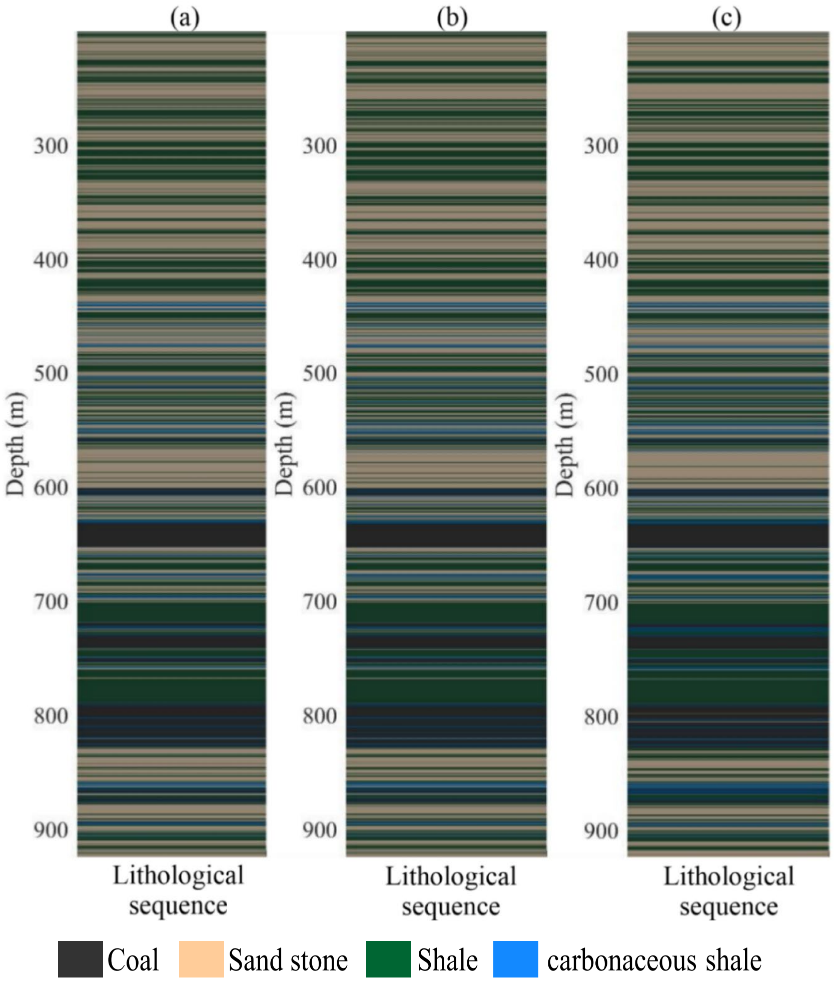


Fig. 15 Comparison between machine learning assisted lithology prediction at W-1, **a** lithological sequence predicted using kNN, **b** lithological sequence predicted using SVM and **c** lithological sequence predicted using RF

of versatile kernel. The interpreter has to carefully choose the kernel function and hyperparameters to get the best result functions (Cawley and Talbot 2004).

6 Conclusions

In this manuscript, we explore the use of machine learning (ML) techniques, specifically kNN, SVM, and RF algorithms, for deducing lithologies within subsurface geological structures using geophysical log data. These logs, which show a broad spectrum of values, underwent statistical analysis to discern the responses corresponding to different

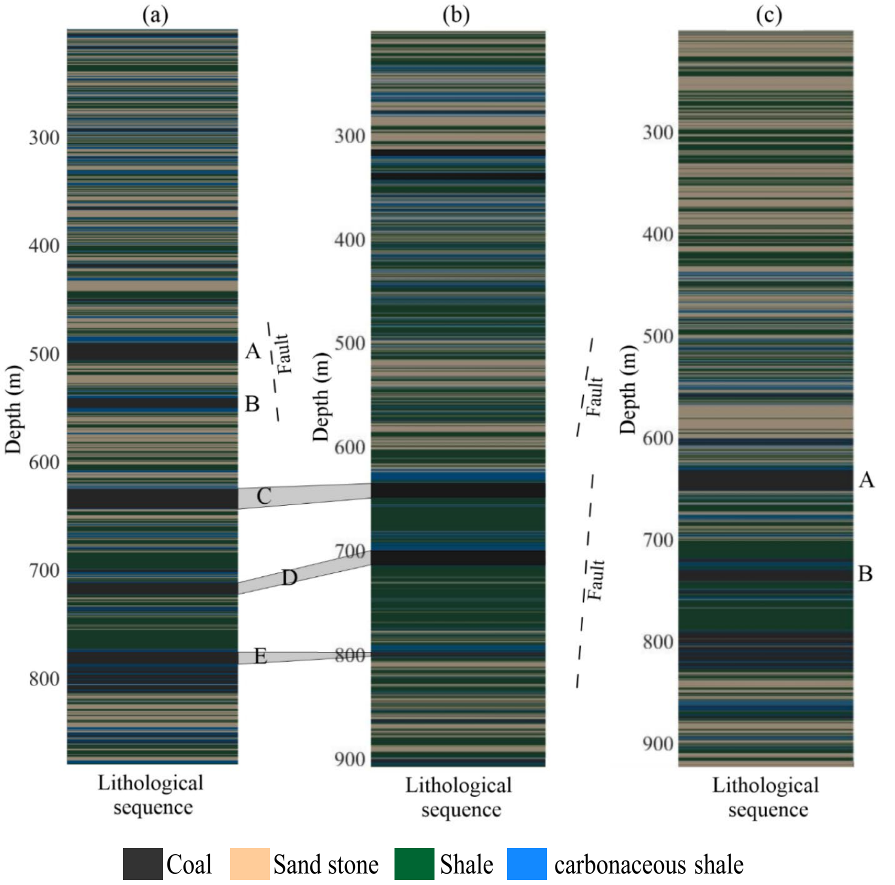


Fig. 16 Correlation of important coal seam among the RF predicted lithologs at the studied wells **a** RF predicted litholog at W-5, **b** RF predicted litholog at W-3, **c** RF predicted litholog at W-1. We have also seen a few local coal seam present in the studied wells

lithologies. Notably, the RF algorithm outperformed kNN and SVM in predicting lithology, boasting an impressive accuracy of ~ 96%. The training and validation of these ML models were rigorously conducted through k-fold cross-validation, complemented by hyperparameter optimization using GridSearchCV. We evaluated the models' effectiveness using various measures like precision, recall, F1-score, and ROC analysis, where the AUC values varied between 0.96 and 1.0 across different lithological categories.

Furthermore, the ML models reliably predicted lithologies in both the training well and in test wells (W-1 and W-3), with accuracies ranging from 89 to 96%. A comparison of the lithological sequences derived from the models and the ML-assisted predictions underscored the consistency of the ML approach. The study also shed light on potential faults and missing coal seams through a correlation analysis between wells, demonstrating ML's extensive utility in subsurface characterization. Overall, the ML methodology was found to be highly efficient, yielding more accurate and rapid results than traditional manual methods, thus establishing itself as a crucial tool for predicting lithological sequences in the exploration and production sector.

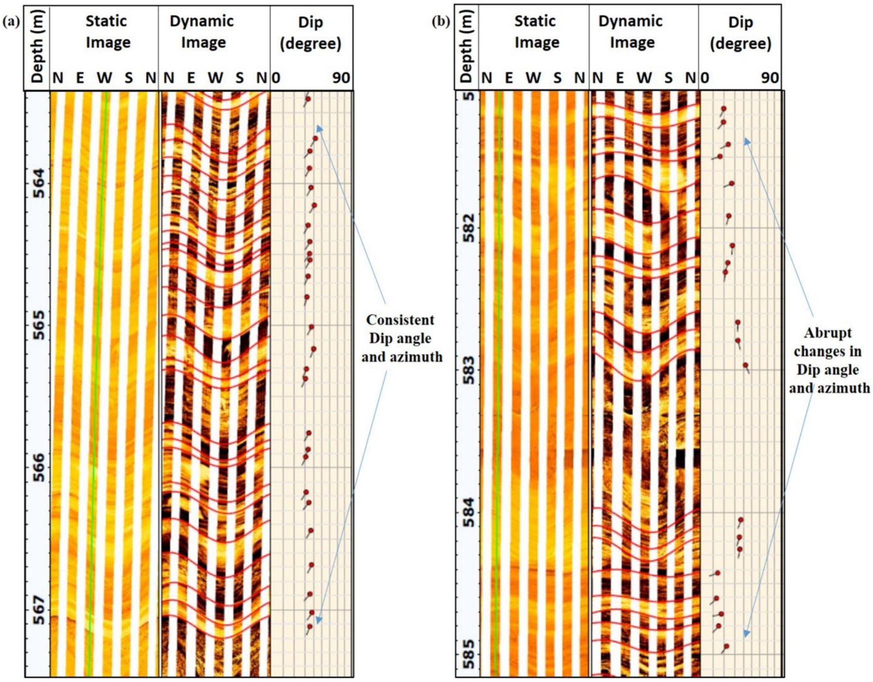


Fig. 17 Resistivity image log of well W-3 illustrating the variation of structural pattern of static and dynamic image with dip magnitude and azimuth **a** Depth interval 563.5–567.5 m, where dip magnitude and azimuth is consistent, **b** Depth interval 581–585 m, where abrupt change in dip angle and azimuth is observed

This study aimed at identifying lithological sequences using a predefined geophysical well log, analysed through ML algorithms such as kNN, SVM, and RF, which underwent thorough training and testing. Key performance indicators for the ML models, including the confusion matrix and ROC curve, were carefully examined. The top-performing ML algorithm was then selected for optimal lithological sequence determination. When applied to well log datasets, the ML algorithm exhibited an overall accuracy exceeding 89%, affirming its effectiveness in differentiating between coal and non-coal lithologies (shale, carbonaceous shale, and sandstone). This ML approach is particularly adept at offering a more detailed and precise classification of coal in the Bokaro coal field, considering the intricate interplay between coal characteristics and geological features like faults. These models are also instrumental in predicting possible faults by identifying missing coal seams during the correlation of wells.

Acknowledgements The authors express their gratitude to ONGC Ltd. for granting permission to present this paper. They also extend their thanks for the invaluable support and encouragement received from Mr. A.K. Dwivedi (Ex-Director of Exploration at ONGC) throughout the research endeavour. The authors would like to acknowledge the Director of the Wadia Institute of Himalayan Geology in Dehradun for granting permission to publish this work. Additionally, KS wishes to express appreciation to SERB-DST for the JC Bose National Fellowship provided. We thankful to the Editor Prof. György Hetényi and anonymous reviewers for their constructive suggestions to improve the manuscript within necessary time and efforts. BM gratefully acknowledges the Science and Engineering Research Board (SERB), Government of India (Project Number:

CRG/2023/001267) for partly sponsoring this work. This work was carried out with WIHG's contribution number WHG/ 0379.

Author Contributions Abir Banerjee: Writing (Original draft), Resources (Data & Softwares), Validation; Bappa Mukherjee: Oversight, Conceptualisation, Code Development, Methodology, All Computation and data analysis, Investigation, Software, Figures, Writing (review and editing manuscript), Manuscript revised and correspondence; Kalachand Sain: Resources, Validation.

Funding This research did not receive any specific grant from funding agencies in the public, commercial or non-profitable sectors.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Azuaje F (2006) Review of "Data Mining: Practical Machine Learning Tools and Techniques" by Witten and Frank. BioMed Eng OnLine 5:51. https://doi.org/10.1186/1475-925X-5-51

Banerjee A, Chatterjee R (2021) A methodology to estimate proximate and gas content saturation with lithological classification in Coalbed methane reservoir, Bokaro field, India. Nat Resour Res 30:2413–2429. https://doi.org/10.1007/s11053-021-09828-2

Banerjee A, Chatterjee R (2022) Pore pressure modeling and in-situ stress determination in Raniganj basin, India. Bull Eng Geol Env 81:49. https://doi.org/10.1007/s10064-021-02502-0

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Bressan TS, de Souza MK, Girelli TJ, Junior FC (2020) Evaluation of machine learning methods for lithology classification using geophysical data. Comput Geosci 139:104475

Busch JM, Fortney WG, Berry LN (1987) Determination of lithology from well logs by statistical analysis. SPE Form Eval 2(4):412–418

Cawley GC, Talbot NLC (2004). Fast leave-one-out cross-validation of sparse least-squares support vector machines. Neural Netw 17(10):1467–1475

Cover TM, Hart PE (1967) Nearest-neighbor pattern classification. IEEE Trans Inf Theory 13: 21–27. https://doi.org/10.1109/TIT.1967.1053964

Dramsch JS (2020) 70 years of machine learning in geoscience in review. Adv Geophys 61:1–55

Fajana AO, Ayuk MA, Enikanselu PA (2019) Application of multilayer perceptron neural network and seismic multiattribute transforms in reservoir characterization of Pennay field, Niger Delta. J Pet Explor Prod Technol 9:31–49

Fix E, Hodges JL (1952). Discriminatory Analysis - Nonparametric Discrimination: Small Sample Perform. Mathematics

Ghosh S, Chatterjee R, Shanker P (2016) Estimation of ash, moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling. Fuel 177:279–287

Hall B (2016) Facies classification using machine learning. Lead Edge 35(10):906–909

Haykin S (2009) Neural networks and learning machines, 3rd edn. Prentice Hall, New York

Ho TK (1995) Random decision forest. Proceedings of the 3rd international conference on document analysis and recognition, Montreal, 14–16 August 1995, 278–282.

Horrocks T, Holden EJ, Wedge D (2015) Evaluation of automated lithology classification architectures using highly-sampled wireline logs for coal exploration. Comput Geosci 83:209–218

Jobe TD, Vital-Brazil E, Khait M (2018) Geological feature prediction using image-based machine learning. Petrophysics 59(6):750–760

Kumar T, Seelam NK, Rao GS (2022) Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India. J Appl Geophys 199:104605. https://doi.org/10.1016/j.jappgeo.2022.104605

Maxwell K, Rajabi M, Esterle J (2019) Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. Int J Coal Geol 214:103284

Meshalkin Y, Shakirov A, Orlov D (2020) Koroteev D (2020) Well-Logging based lithology prediction using machine learning. Eur Assoc Geosci Eng Conf Proc Data Sci Oil Gas 1:1–5

Mukherjee B, Sain K (2021) Vertical lithological proxy using statistical and artificial intelligence approach: a case study from Krishna-Godavari Basin, offshore India. Mar Geophys Res 42:3. https://doi.org/10.1007/s11001-020-09424-8

Mukherjee B, Kar S, Sain K (2024) Machine Learning Assisted State-of-the-Art-of Petrographic Classification From Geophysical Logs. Pure Appl. Geophys. https://doi.org/10.1007/s00024-024-03563-4

Navin JRM, Pankaja R (2016) Performance analysis of text classification algorithms using confusion matrix. Int J Eng Tech Res 6(4):2321–2869

Oyler DC, Mark C, Molinda GM (2010) In situ estimation of roof rock strength using sonic logging. Int J Coal Geol 83:484–490

Paul S, Ali M, Chatterjee R (2018) Prediction of compressional wave velocity using regression and neural network modeling and estimation of stress orientation in Bokaro coalfield, India. Pure Appl Geophys 175:375–388. https://doi.org/10.1007/s00024-017-1672-1

Prajapati R, Mukherjee B, Singh UK et al. (2024) Machine learning assisted lithology prediction using geophysical logs: A case study from Cambay basin. J Earth Syst Sci 133:108. https://doi.org/10.1007/s12040-024-02326-y

Refaeilzadeh P, Tang L, and Liu H (2007) On comparison of feature selection algorithms. In Proc. AAAI-07 Workshop on Evaluation Methods in Machine Learing II. pp. 34–39

Roslin A, Esterle JS (2016) Electrofacies analysis for coal lithotype profiling based on high resolution wireline log data. Comput Geosci 91:1–10

Srinaiah J, Raju D, Udayalaxmi G, Ramadass G (2018) Application of well logging techniques for identification of coal seams: a case study of Auranga Coalfield, Latehar district, Jharkhand state. India J Geol Geophys 7(1):1–11

Sun Z, Jiang B, Li X, Li J, Xiao K (2020) A data driven approach for lithology identification based on parameter-optimized ensemble learning. Energies 13(15):3903

Vapnik VN (1995) The nature of statistical learning theory, New York.

Xu C, Misra S, Srinivasan P, Ma S (2019) When petrophysics meets big data: What can machine do? In: *SPE Middle East Oil and Gas Show and Conference*. OnePetro.

Zhang J, He Y, Zhang Y, Li W and Zhang J (2022) Well-logging-based lithology classification using machine learning methods for high-quality reservoir identification: A case study of Baikouquan Formation in Mahu area of Junggar Basin, NW China; Energies 15:3675

Zhong R, Johnson JR, Chen Z (2020) Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). Int J Coal Geol 220:103416

Zhou B, O'Brien G (2016) Improving coal quality estimation through multiple geophysical log analysis. Int J Coal Geol 167:75–92

Zhou K, Zhang J, Ren Y, Huang Z and Zhao L (2020) A gradient boosting decision tree algorithm combining synthetic minority oversampling technique for lithology identification; Geophysics 85 WA147–WA158.

Zhou B, Hatherly P, Guo H, Poulsen B (2001) Automated geotechnical characterisation from geophysical logs: examples from southern Colliery, central Queensland. Explor Geophys 32:336–339

Authors and Affiliations

Abir Banerjee[1] · **Bappa Mukherjee**[2] · **Kalachand Sain**[2]

✉ Bappa Mukherjee
bappa.ism@gmail.com

[1] Deparment of Sub-Surface Team, Oil and Natural Gas Corporation Limited, Bokaro, Jharkhand 827001, India

[2] Seismic Interpretation Laboratory, Wadia Institute of Himalayan Geology, Dehradun, Uttrakhand 248001, India