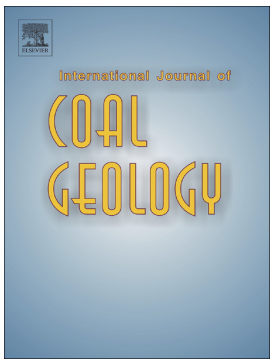

Intelligent classification of coal structure using multinomial logistic regression, random forest and fully connected neural network with multisource geophysical logging data

Zihao Wang, Yidong Cai, Dameng Liu, Feng Qiu, Fengrui Sun, Yingfang Zhou

PII: S0166-5162(23)00026-5

DOI: https://doi.org/10.1016/j.coal.2023.104208

Reference: COGEL 104208

To appear in: *International Journal of Coal Geology*

Received date: 14 June 2022

Revised date: 15 November 2022

Accepted date: 12 February 2023

Please cite this article as: Z. Wang, Y. Cai, D. Liu, et al., Intelligent classification of coal structure using multinomial logistic regression, random forest and fully connected neural network with multisource geophysical logging data, *International Journal of Coal Geology* (2023), https://doi.org/10.1016/j.coal.2023.104208

Intelligent Classification of Coal Structure Using Multinomial Logistic Regression, Random Forest and Fully Connected Neural Network with Multisource Geophysical Logging Data

Zihao Wang[a,b], Yidong Cai[a,b]*, Dameng Liu[a,b], Feng Qiu[a,b], Fengrui Sun[a,b], Yingfang Zhou[c]

[a]*School of Energy Resources, China Univesity of Geosciences, Beijing 100083, China*

[b]*Coal Reservoir Laboratory of National Engineering Research Center of CBM Development & Utilization, China University of Geosciences, Beijing 100083, China*

[c]*School of Engineering, Fraser Noble Building, King's College, University of Aberdeen, AB24 3UE Aberdeen, UK*

Abstract: The structure of coal indicates the degree of its fragmentation after tectonic movement, which affects the exploration and development of coalbed methane (CBM). Although coal core observations are the most convenient and intuitive way of identifying the coal structure, they are not applicable for use in unexplored coal seams without CBM wells, and they are also very time-consuming. In comparison, geophysical-logging interpretation of the coal structure is more efficient and economical. However, although qualitative methods, such as principal component analysis (PCA), can be used to identify the coal structure with geophysical logging, the interpretation is limited by the calculation ability, and improvements are required based on the structure of an empirical model. Multinomial logistic regression (MLR), random forest (RF), and deep fully connected neural network (DNN) are effective machine learning methods more accurate than the traditional method that with model-aided identification. In this respect, the MLR method is a classical method based on mathematical linear regression, and it has a low construction cost; RF is an ensemble learning algorithm based on a decision tree and use of a bagging algorithm; and DNN is a deep learning model based on self-built feature engineering that has high classification accuracy under a large amount of data training and provides obvious advantages in visual coal classification problems. In this work, the three machine learning methods, MLR, RF, and DNN, were used to identify the coal structure. Two sets of logging data comprising different quantities from the Anze Block of the southern Qinshui Basin, North China, were selected to quantitatively compare the accuracy of coal structure identification with partial coal core observation. The results showed that for 210 and 840 samples, respectively, the accuracy was 76% and 77% for MLR, 83% and 86% for RF, and 82% and 86% for DNN. These results show that the MLR and DNN methods are superior for use with minimal and maximum amounts of data, respectively, and the RF method provides overall accuracy. Furthermore, an algorithmic classification of the coal structure was established, and the geological factors controlling the predicted structure, such as geostress, coal seam thickness, and burial depth, were distinguished.

Keywords: coal structure identification; logging data; machine learning; random forest; neural network; regression

1. Introduction

The structure of coal reflects the coal crushing degree (Fu et al., 2009; Liu et al., 2022) and can be classified as primary coal, cataclastic coal, granulated coal, and mylonitic coal (Gao et al., 2018; Lv et al., 2019). The coal structure differs in accordance with the regional geological structure, and differences in its pore structure and permeability relate to its petrophysical properties (Qin, 2018; Huang, 2017; Wang et al., 2020). The structure is not only an indicator of the coalbed methane (CBM) reservoir properties, but it is also used to evaluate and design hydraulic fracturing operations with respect to CBM development (Mastalerz et al., 2008; Shi et al., 2020; Wang et al., 2020). The traditional method used to determine the coal structure is by direct observation of the coal seam through a drilling coal core (Hoek et al.,1997). However, due to core sampling costs and the conditions involved in direct coal seam observations, coal seam evaluations are constrained to a limited sampling profile (Shi et al., 2020; Chen et al., 2021).

Therefore, geophysical logging can be used as an alternative to enable the accurate and efficient determination of the coal structure. It is a low-cost and efficient method that enables effortless and continuous quantification of critical geophysical parameters and is widely used in reservoir classification and evaluation in the petroleum and coal industry. Several studies have used geophysical logging in this respect. For example, Raeesi (2012) used logging data to classify and identify hydrocarbon reservoir lithofacies and their heterogeneity. Hernandez-Martinez (2013) used the R/S method based on the fractal theory for facies recognition by determining the complexity of the logging signal. Additionally, Sharawy (2016) applied principal component analysis (PCA) and cluster analysis to raw and normalized logging data and used the conventional well logs of four boreholes and data from a traditional core analysis conducted in one of these wells to identify the electrofacies of the Kareem Formation. Several other studies have also been conducted to interpret the coal structure using logging data, and such methods have been directly applied to instruct the actual CBM production (Li et al., 2011; Kumar et al., 2022). However, although many coal structure prediction methods have used logging data, most of these methods have involved conducting a qualitative analysis based on empirical models or PCA (Fu et al., 2009; Teng et al., 2015; Ren et al., 2018). Statistical analysis methods, such as linear regression, have also been used, but the actual problem of predicting the structure of the coal body exhibits non-linearity (Wang et al., 2020; Cao et al., 2020; Shi et al., 2020; Chen et al., 2021).

Traditional machine learning methods, including multinomial logistic regression (MLR), support vector machine (SVM), and random forests (RF), focus on strict and visible mathematical logic, and these methods can be used to conduct a quantitative analysis using mathematical methods under computer power (Guo et al., 2021). The essence of MLR is the superposition of multiple

linear regressions, and the essence of SVM is the distance formula from point to line. Empirical formula and other methods have thus been used to conduct quantitative analyses of the coal structure, and classical machine learning algorithms have been found to have a good calculation effect when inverting the coal structure using high fitting logging data (Fu et al., 2009; Teng et al., 2015). Unlike SVM and MLR, RF is a machine learning method that has a strong ability to process tabular data using decision trees (DTs). It uses an optimized bagging algorithm for accuracy, which makes it applicable for use with classification problems in the coal field, and its dataset often provides a better classification accuracy than deep learning models (Gordon et al., 2022; Wang et al., 2022). Maxwell (2019) compared RF, gradient boosted machines (GBM) and DNN for delineating altered and non altered coal from geophysical log data, and found that RF had the highest accuracy. With the continuous development of machine learning methods, deep learning methods (such as neural networks) have attracted attention and achieved good results in many disciplines and fields (Siregar et al., 2017). For example, they have been widely applied in predicting rock types and petrographic identification (Imamverdiyev and Sukhostat, 2019; Liu et al., 2021). In the coal research area, Chatterjee (2022) compared four machine learning algorithms and used SVM to build the best REY potential model classification scheme. Wojtecki (2022) employed machine learning algorithms to assess the rockburst hazard status of underground coal mine openings, and Wei (2022) integrated the synergistic effects of coal and biomass in pyrolysis to build a model of pyrolysis under the RF algorithm. Zhang (2022) presented a method for predicting the coal self-ignition tendency using MLP and RF machine learning methods based on 204 sets of CPT experimental data. Deep learning has thus been widely applied in the field of image classification and intelligent recognition of coal (Wang et al., 2022; Xiao et al., 2022; Zhang et al., 2022).

Machine learning is best suited to classification problems, and classification models that invert the structure of the coal body from logging data meet this requirement. In most studies, due to the overflow of computing power, the results of traditional machine learning algorithms are accurate, while neural networks waste computing costs (Xu et al., 2021). Previously, the DNN was used to classify coal macerals and the contact angle for differently ranked coals (Zhao et al., 2022; Ibrahim, 2022; Tiwary, 2020). However, fewer studies have focused on using the classical classification algorithm, and they have directly promoted the use of neural networks on a large-scale. Nevertheless, Tiwary (2020) used the RF model to classify different phases of coal macerals (organic constituents) and minerals (inorganic constituents), and the results reached an accuracy of 0.9, which is higher than most non-overfitting neural network models. Ibrahim (2022) applied function networks, support vector machine (SVM), and RF to predict the contact angle in coal, and the accuracy of all was found to be higher than 0.94. Such results clearly assist in assessing and learning about the effectiveness and norms of machine learning.

A few machine learning methods have been selected for coal structure prediction (Shi et al., 2020; Chen et al., 2021). However, the best method for use in practice has not yet been identified, and

there is need to establish a comprehensive method for identifying coal structure, due to the limited amounts of coring data. Therefore, in this study, we compared the use of three algorithms in the prediction of coal structure: the MLR algorithm (representing classical machine learning), RF algorithm (representing ensemble learning) and the DNN algorithm (representing deep learning). In this paper, we first review the identification of the coal structure and then establish the MLR, RF, and DNN algorithm models for predicting coal structure using different-logging datasets from the Anze Block in the southern Qinshui Basin, North China. The performances of the methods are then compared in terms of prediction accuracy and using environment. Finally, the impacts on coal structure from geological controls, including geostress, coal thickness, and burial depth, are explored.

2. Geological Background

The Qinshui Basin is surrounded by the uplifts of the Taihang Mountains, Huo Mountains, Wutai Mountains, and Zhongtiao Mountains (Cai et al., 2011). There is a NE-SW main syncline in the Qinshui Basin that reflects the W-E extrusion. Combined with a series of normal faults in the area, the partial deflection of the syncline was first controlled by NW-SE compression in the Yanshanian orogeny. With the development of late NE-SW normal faults in the block, strike-slip shear occurred, and the faults correspond to the early Himalayan NW-SE extension. In relation to the strike-slip tectonic movements, the Qinshui Basin became a pull-apart basin (Teng et al., 2015; Ren et al., 2018). The Anze Block is located in the southern Qinshui Basin (Fig. 1), which experienced the superimposition of the Indosinian, Yanshanian, and Himalayan orogenies, and complex structural patterns resulted after coal seam formation (Cai et al., 2011; Wang et al., 2020).

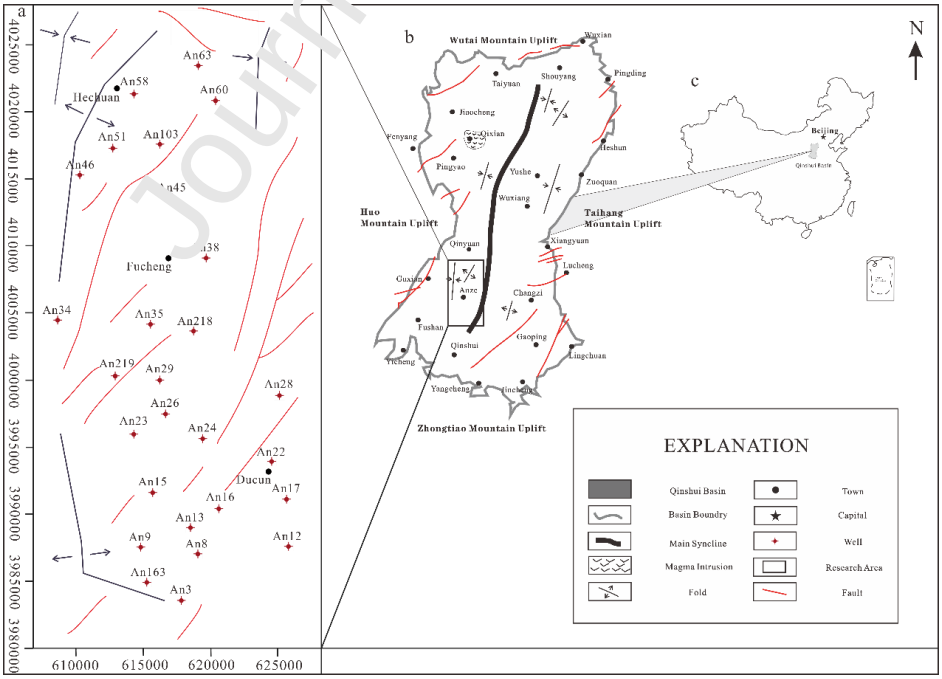

Fig. 1 Generalized map showing the location and structural outline of the Anze Block, southern Qinshui, North China

The No.3 coal seam in the Permian Shanxi Formation (P1s) and No.15 coal seam in the Carboniferous Taiyuan Formation (C3t) are the main target seams for CBM exploration and development (Cai et al., 2011; Wang et al., 2020). These seams are semi-anthracite to anthracite with $R_{o,max}$ of 1.9%–2.7%. The coal structure of samples from the drilling coal core of the No. 3 coal seam of the Shanxi Formation is shown in Fig. 2. The main burial depth of the No.3 coal seam varies from 850 m to 1250 m, and the thickness changes from 5 m to 8 m (with an average of 7 m). The coal seam gradually deepens and thickens from the northwest to southeast. The roof and floor lithology of the No. 3 coal seam is mainly dark to gray mudstone and tight sandstone with good sealing, which makes this the most continuous and stable coal seam in the region. Therefore, the No.3 seam was selected to investigate the coal structure in this study.

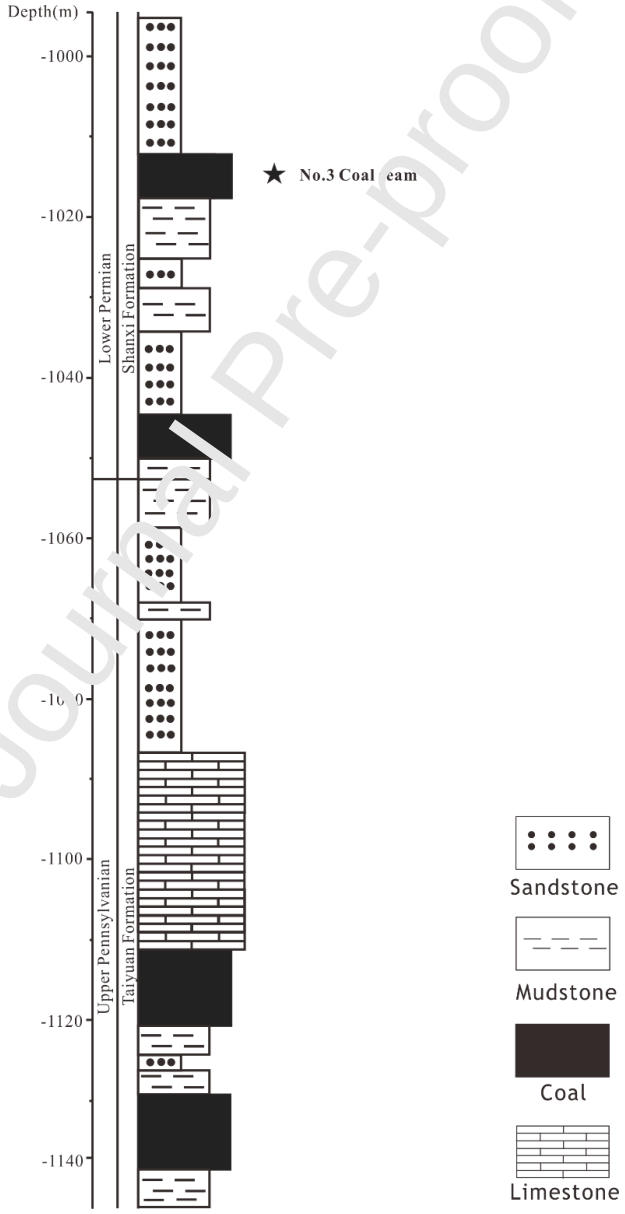

Fig. 2 Stratigraphic section including targeted the No.3 coal seam of the Anze Block in the southern Qinshui Basin

3. Methodology

3.1. Identifying coal structure by core observation

The structure of coal can be classified into primary coal, cataclastic coal, granulated coal, and mylonitic coal (Gao et al., 2018). Primary coal is characterized by the development of endogenous butt and face cleats, the tectonic fissures are weakly developed, and bedding fissures are more common (Yang et al., 2021). Cataclastic (structured) coal normally has a macroscopic fragmented structure that is mainly characterized by a hard and mostly complete coal structure. The bedding and endogenous fractures clearly show slight deformation of the coal structure and multiple sets of structural fractures. Granulated (structured) coal generally has a granulated structure with loose granular powder particles and comparatively better overall sorting. The particle size is relatively uniform (with generally less than 5-mm crumb and crushed powder) and with sub-angular to sub-rounded larger blocks (Wang et al., 2022). From observations, the granulated block has roundness and wear, and the primary coal structure has almost disappeared. Mylonitic (structured) coal is generally characterized by its fine compact powder particles and fine fissures. The primary coal structure body is severely broken, maintaining the coal core integrity is difficult, and stratification and endogenous fissures are difficult to identify.

In this work, 830 coal core samples with corresponding logging data were collected from the Shanxi Formation No.3 coal seam. According to the above classification of coal core observations, 208 primary coal samples, 504 cataclastic coal samples, and 118 granulated coal samples were identified. No mylonitic coals were observed. If we directly adopted this classification scheme, it would not have been possible to establish an adequate machine learning model, due to the imbalanced selection of sample types. Therefore, to ensure the credibility of the inversion of geophysical logging through machine learning, the granular structure and the mylonitic structure were classified into one category. The coal structure of the Anze Block can thus be roughly classified into three categories: Type 1 is primary coal and comprises 208 samples, Type II is cataclastic coal with 504 samples, and Type III coal includes both granulated coal and mylonitic coal with118 samples.

3.2. Correlation analysis and optimization of logging parameters

3.2.1. Identifying coal structure using empirical methods with logging data

To identify the coal structure using empirical methods and logging data (Fu et al., 2009; Teng et al., 2015), the well-logging curves of the natural gamma ray (GR), compensated neutron (CNL), density (RHOB), sonic (DT), caliper (CAL), shallow resistivity (RLLS), deep resistivity (RLLD), and spontaneous potential (SP) can usually be employed, as shown in Fig. 3. Due to the density difference between the coal seam and the clastic rock layer, the coal density curve shape shows a sudden decrease. With the increased coal fragmentation degree from primary coal to structured coal, pores and fractures developed, and the content of radioactive substances per unit volume was reduced. Therefore, low anomalies appear on the natural gamma curve. However, the ions in the

fluid enhance the conductivity and decrease the resistivity, and the resistivity curve thus shows a decline. When drilling coal seams, the strength of the structured coal is low, the structure is loose, and the well diameter must be increased because the wall of the CBM well is prone to collapse.

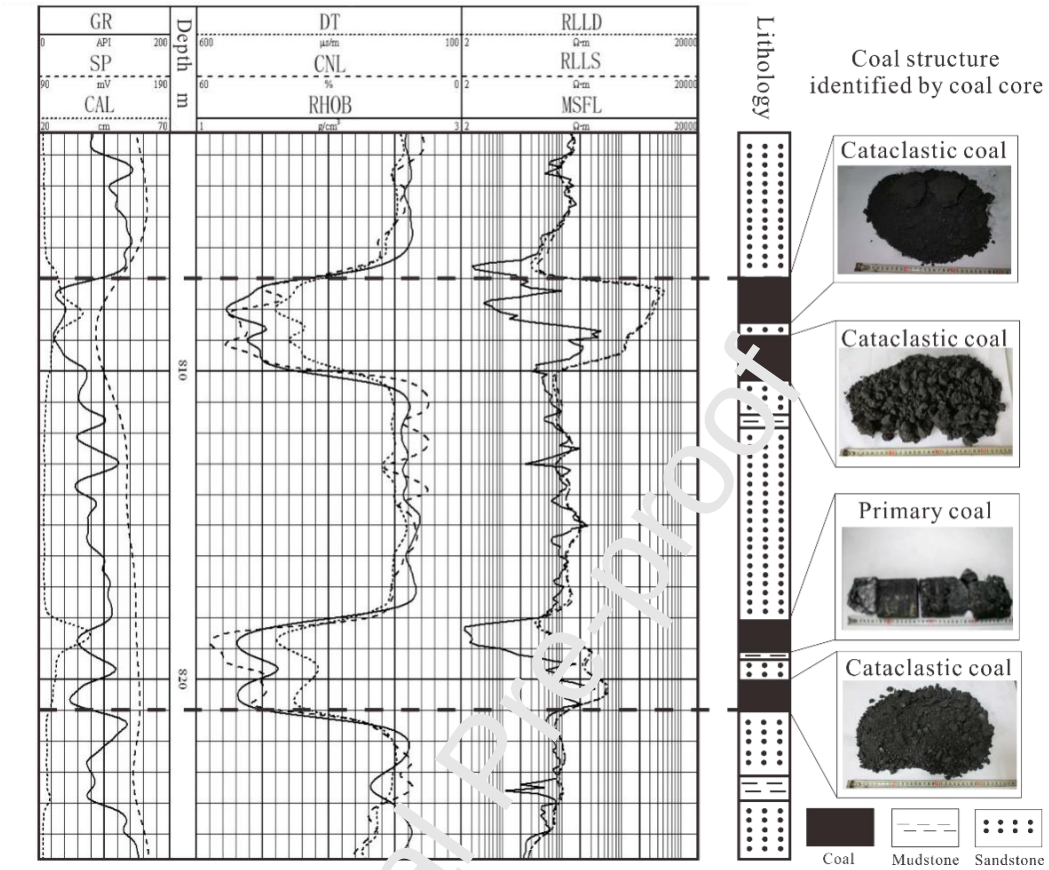

Fig. 3 Coal structure by core observation and its logging curve characteristics of Well An 26

The deep resistivity of the logging response is more obvious for the coal seam and microspherical focus logging (MSFL) is weakly sensitive to changes in the coal structure. Furthermore, after comparing the logging data of coals in all CBM wells with a structure type identified through a data visualization analysis (Fig. 4), the DT, CNL, CAL, GR, SP and RHOB were found to have correlations with the increasing fragmentation of the coal structure. Therefore, they were also used as input parameters to identify the coal structure.

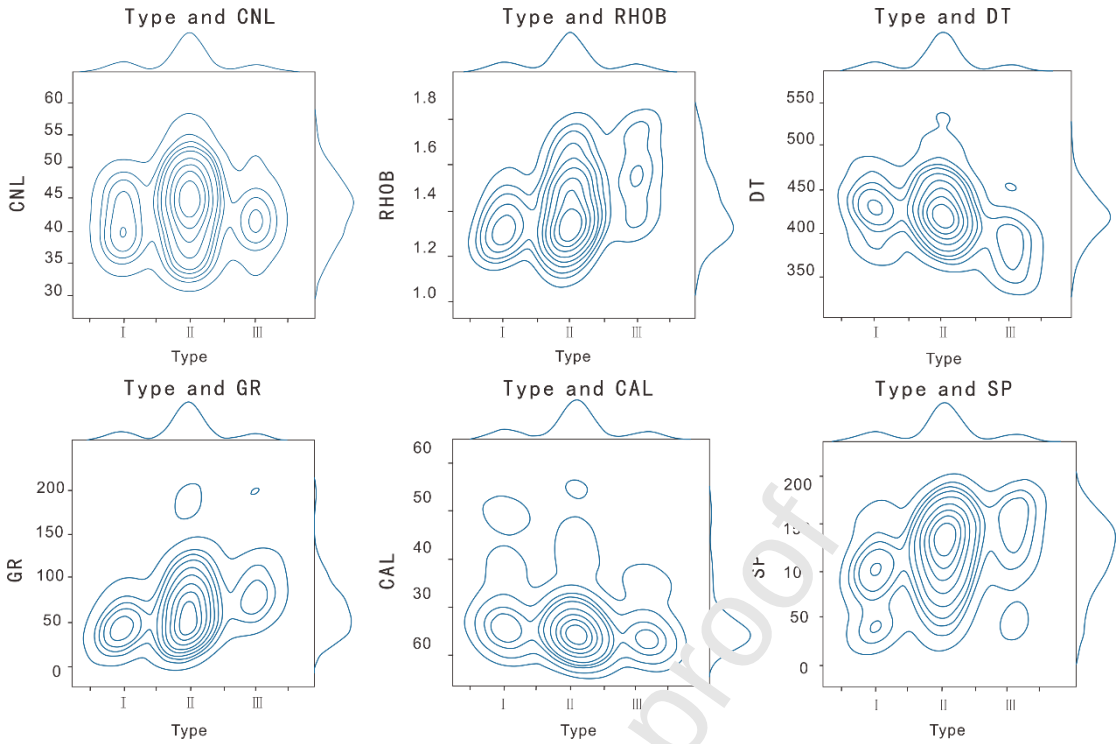

Fig.4 Data visualization of CNL, RHOB, DT, GR, CAL, SP logging data and coal structure TPYE values

3.2.2. Logging parameter correlation analysis and preferred parameters

Based on the above analysis, CNL, CAL, DT, RLLD, GR, and RHOB were selected to identify the coal structure. Of the statistical classification schemes (the Pearson, Kendall, and Spearman (Rock., 1987)), the Pearson scheme is more commonly used in reservoir studies. Additionally, when two continuous variables are linearly correlated, it is preferable to use the Pearson product difference correlation coefficient for analysis, and the attribute is then selected after the linear correlation parameters are determined without the zero value or optimization of the low-correlation value. The Pearson correlation coefficient varies from −1 to 1. A coefficient with a value of 1 means that X and Y can be well described by the linear equation. All data points fall well on a straight line, and Y increases with increasing X. A coefficient with a value of −1 means that all data points fall on the straight line, and Y decreases with increasing in X. If the value of a coefficient is 0, it means that there is no linear relationship between the two variables (Pearson., 1895). The calculation formula of the correlation coefficient can be expressed by Eq. (1),

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}. \quad (1)$$

The Python programming language was used to create a correlation matrix, as shown in Fig. 5.

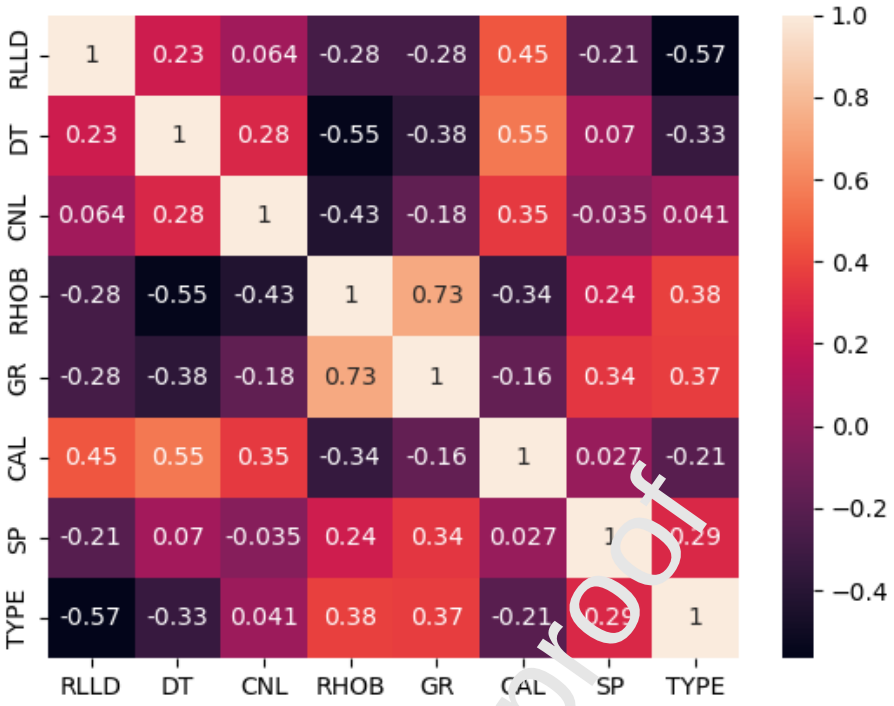

Fig. 5 Correlation matrix heatmap for logging data and coal structure TYPE

The correlation matrix and data visualization indicated that the values of the GR, RHOB, DT, and RLLD logging parameters were correlated with the coal structure type (TYPE), while the CNL values were poorly correlated. The correlation coefficient between CNL and TYPE had a low value of 0.041 from the correlation matrix, which means that there was no correlation between CNL and TYPE.

3.3. Multinomial logistic Regression

The essence of logistic regression is based on the classification under generalized linear regression, and it is a statistical method that takes the linear regression prediction boundary as the boundary, maps the data to be classified and the boundary relationship to the probability distribution by a function, and provides a classification according to the distribution. The steps involved include obtaining a generalized linear regression solution and designing and mapping the classification data function (Abrougui et al., 2019). The model used in linear regression solving can be expressed by Eq. (2),

$$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b, \quad (2)$$

where x_i is an eigenvalue of the input, w_i is the model parameter under regression, and inputs of multiple x_i will train all w_i values. The logic function is based on a successful linear regression solution, and a formula applied to the decision surface. The logic model under the binary classification of logic regression can be expressed by Eq. (3),

$$p(y|x(i),\theta) = (h_\theta(x(i)))_y(1 - h_\theta(x(i)))_{1-y}. \quad (3)$$

For a multi-classification problem, the main processing methods used are the multiple logistic regression binary classification method and the SoftMax processing method (Nazmi et al., 2020)

based on one-to-one, one-to-many, or multi-to-many. The SoftMax method is essentially an extension of the binary classification method, and its effect is better when there is more to classify. In contrast, the multiple logic binary classification method is adopted owing to its superior accuracy in multi-classification when there are fewer categories. The corresponding logic model can be expressed by Eq. (4),

$$p(y = k|X, \theta) = h_\theta(k)(X). \quad (4)$$

To obtain a multiline logistic regression model parameter estimation with a given training data set, $T = \{(x_1, y_1), (x_2, y_2),..., (x_n, y_n)\}$, where $x_j \in R_n$, $y_j \in \{1, 2,...,k\}$, the maximum likelihood estimation method can be used to estimate the model parameters and thus obtain a multiline logistic regression model.

3.4. Random Forest

Random Forest (RF) is based on a decision tree (DT) and is a type of ensemble analysis (Breiman, 1996). It requires minimal input parameters and little hyperparameter tuning. The DT has a tree structure that describes the classification of instances, as shown in Fig. 6.

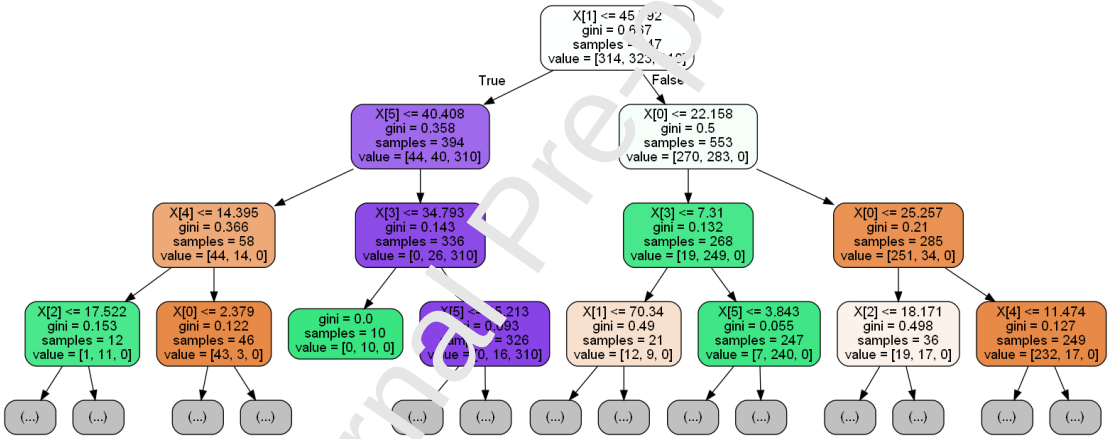

Fig. 6 Decision tree visualization for coal structure prediction

The DT consists of nodes and directs edges. Internal nodes represent a feature in each, and leaf nodes represent a class in each; therefore, there are two types of nodes in the tree (Hashemizadeh et al., 2021). The essence of a DT is to solve a set of conditional statements. Each internal node represents the condition associated with the tree model logic, and each leaf node provides the conclusion associated with tree model logic. After each sample enters the DT, there are and only one path can pass.

Based on a set of multiple DTs, the RF first extracts training samples from the training set with replacement randomly. By using the new training set, the sub-models are trained. For the classification problem, a voting method is used, and the classification category of the submodel with the most votes is the final category (Breiman, 1996). Approximately 63.2% of the samples in the final initial training set appear in the sampled set. The advantage is that each learner uses only 63.2% of the samples, and the remaining 36.8% can be used for conducting an out-of-bag estimation (Breiman, 1996). The formula for the out-of-wrap estimation is expressed by Eq. (5),

$$H^{oob}(x) = \text{argmax}_{y\in\gamma} \sum_{t=1}^{T} I\!I\,(h_t(x) = y) \cdot I\!I(x \neq D_t), \quad (5)$$

where D_t is the actual sample set used by h_t, and $H^{oob}(x)$ is the out-of-wrap error on sample set *x*. The out-of-bag estimation formula for the generalization error of bagging is expressed by Eq. (6),

$$\in^{oob} = \frac{1}{|D|} \sum_{(x,y)\in D} I\!I(H^{oob}(x) \neq y). \quad (6)$$

The training set of each tree is different, and it contains repeated training samples. Another feature is that compared with the DT, each split process of the tree in the RF does not use all the features to be selected. However, it randomly selects certain features from all the features to be selected, and then selects the optimal feature from the randomly selected features. Therefore, RFs do not easily result in overfitting and they have a good anti-noise ability. Additionally, the main control parameters of RF include correlations between any two trees in the forest, the ability to classify each tree in the forest, and the number of feature selection.

3.5. Deep fully connected neural network

3.5.1. Algorithm preprocessing

Neural networks essentially comprise a combination of single neurons. The trained neural network usually triggers one-to-many and many-to-many signals between neurons to simulate the trigger form of a biological nerve. The trigger form (trigger, non-trigger) can be directly indicated by the machine signal 0/1, and its logical significance directly corresponds to the step function under the ideal (Abdul-Majeed et al., 2021; Xu et al., 2021). Neurons are combined according to a certain layered structure, and a neural network is thus created. In practice, as an activation function is required to satisfy the firing from neurons, the traditional sigmoid gradient function is prone to gradient disappearance (Welper, 2020), which makes it difficult to optimize important gradients. The Relu function is often selected as an alternative. In this respect, there is a direct 0 indication on activation, and the gradient function has only simple 0 and 1 values, which are beneficial for the gradient.

3.5.2. Neural network structure

Based on the Graphviz library under Python for visualizing the output neural network framework graph, the neural network framework diagram was redrawn as shown (Fig. 7).

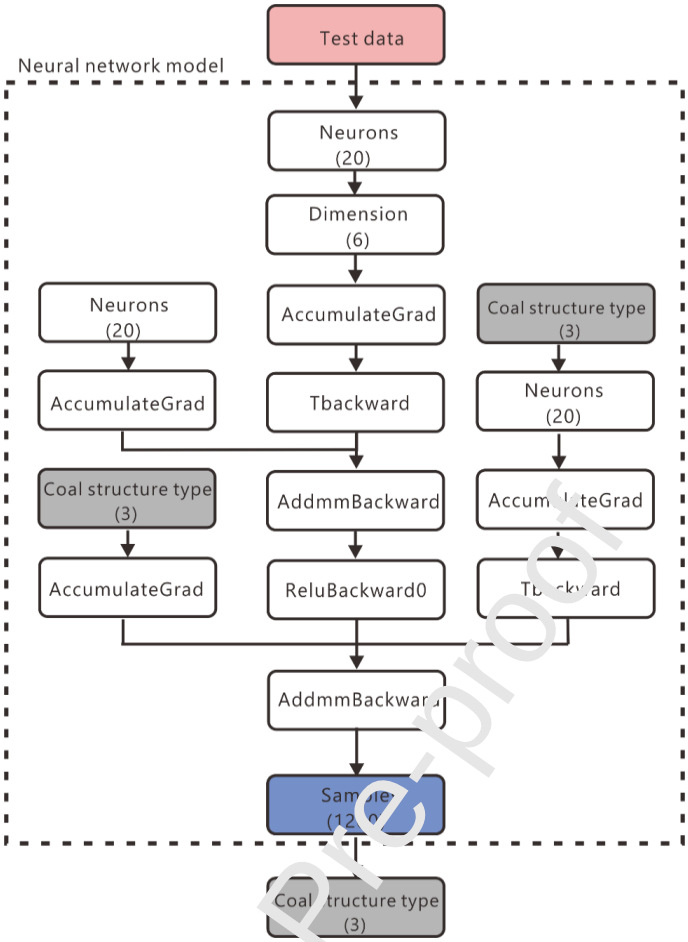

Fig. 7 The logic structure of deep fully connected neural network for coal structure identification

The number of hidden layer nodes affects the neural network performance, and it can thus be determined using an empirical formula. An intermediate calculation is automatically released at the end of a gradient, and the specific formula is expressed by Eq. (7),

$$h = \sqrt{m+n} + a, \tag{7}$$

where h is the number of hidden layer nodes, m is the number of input layer nodes, n is the number of output layer nodes, and a is the adjustment constant between 1 and 10. In this work, we established a scale (6-30-3) of neural networks under 20 neurons. For the intrinsic calculation of neural networks, back propagation is required to calculate the gradient correlation. However, due to the dependence of back propagation on memory and the consideration of overfitting, a direct error report is avoided (Welper, 2022), and an intermediate calculation is automatically released at the end of a gradient. The classification results for the coal structure can be obtained by inputting logging data through this training model.

4. Results and Discussion

4.1. Comparison of methods used to identify coal structure (MLR, RF, and DNN)

Shi (2020) discussed the importance of overfitting and solved it by using set training data and testing data at a ratio of 7:3. However, overfitting also occurs in relation to factors such as the number of training sessions used. Therefore, as the data sets were divided using the ratio of 7:3, it

was important to determine the optimum number of training steps in neural network model to avoid overfitting. In comparing the results of different methods, overfitting should be avoided by setting an aborted training accuracy or by reflecting the accuracy that varies with the amount of data. Due to the influence of difference in samples number, the accuracy of machine learning algorithms is different (Cao et al., 2020). Compared with an artificial weighting method, such as PCA, different sample numbers and combinations are conducive to comparing inversions and evaluating the coal structure (Chen et al., 2021). In this work, 107 data and 830 data of two orders of magnitude were selected for multi-attribute and multi-label classification. Each classification corresponded to six labels, and the sample training under permutation and the combination had a sufficient magnitude difference. Of these, 107 data sets for training included 18 samples of Type I, 75 samples of Type II, and 13 samples of Type III, and the 830 data sets included 208 samples of Type I, 504 samples of Type II, and 118 samples of Type III.

4.1.1 Data preprocessing and model initialization

The selected data samples were obtained from the same logging data of the same company in the block. However, as data noise can have a large impact on the training results, and random samples have a large data imbalance, the data and parameters still need to be preprocessed. The specific steps used include data cleaning, missing value supplement, and hyperparameter optimization (Wang et al., 2022).

The logging data were preliminarily cleaned using a triple standard deviation as the boundary of abnormal values. According to the condition of a normal distribution, the probability of taking values in the interval ($u \pm 3\sigma$) was 99.73%, while the probability of taking values outside the interval was less than 0.3% and could be deemed as a minimum probability event (Froncisz et al., 2020). A function of screening out small probability event data was realized through a subset, which was used as the low probability value boundary for further data cleaning. Before adding missing values, the logging data were normalized to 0–100. Due to the unevenness of the three label samples in the dataset, the SMOTE interpolation method was used to supplement the labels with less data (Sinha et al., 2019). The visualization model after data preprocessing is shown in Fig. 8.

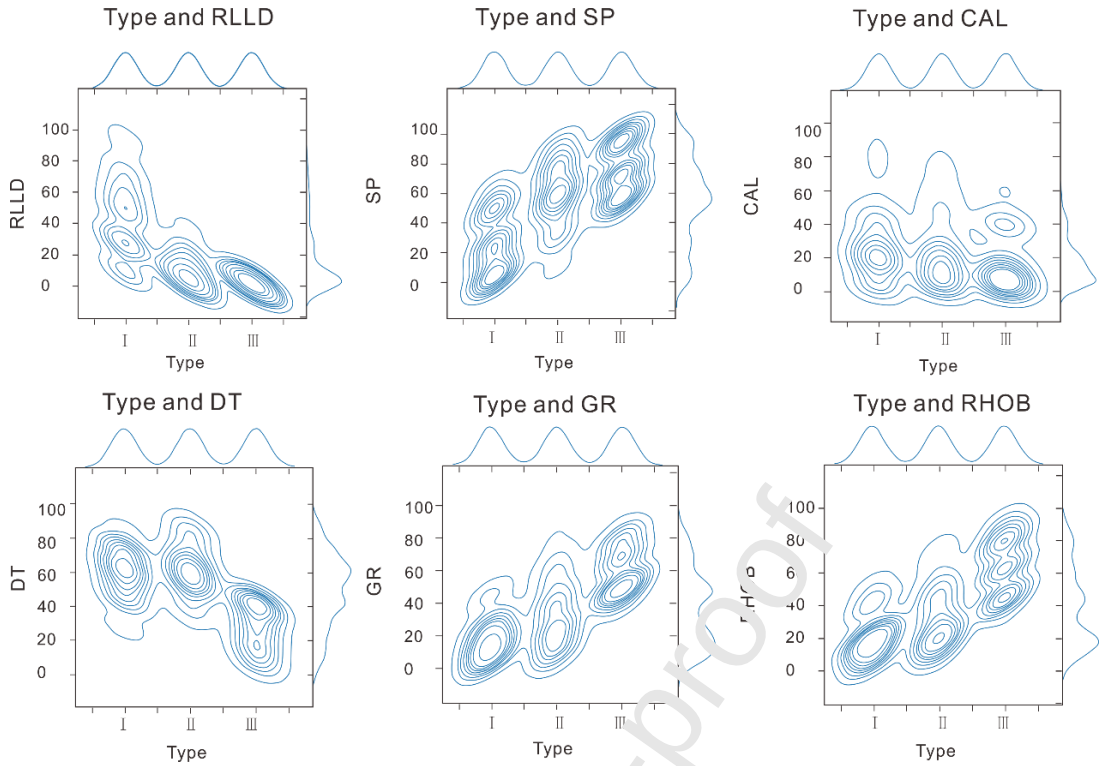

Fig.8 Data visualization of CAL, RHOB, DT, GR, CAL, SP logging data and coal structure TPYE values

For the MLR algorithm under the Sklearn framework, the function is not usually expanded for hyperparameter optimization but only for manual tuning of the random number seeds. Therefore, the function Gridsearchcv in the Sklearn library was used for RF, and empirical formula adjustment parameters and manual adjustment parameters were combined for DNN.

4.1.2 Multinomial logistic Regression

Python was used to establish and test the multi-classification logic model. For the selected 300 data set, 70% of the data were extracted as training data to establish the model, and 30% was used to test the model accuracy. The specific processing was conducted using the Sklearn framework of Python (Bizhani and Kuru, 2022). After splitting the 300 and 1200 sample data sets, there were 210 and 840 training sets and 90 and 360 test set samples, respectively. After normalization and mapping of the classification data, the accuracy of the MLR training set reached 93%, and the accuracy of the MLR test set was 91%. The same method processes were used for the 1200 data set. The final accuracy of the MLR training set was 84%, and the accuracy of the MLR test set was 76%. Two trained models were used to predict the coal structure for the 150 data set, and a confusion matrix of each was created. The classification result is shown in Table 1.

Table 1 Multiline logistic regression (MLR) model test confusion matrix

		210 training data			840 training data		
	Prediction	Type I	Type II	Type III	Type I	Type II	Type III
Actual	Type I	15	3	0	15	3	0
	Type II	22	86	8	20	86	10
	Type III	2	1	13	2	0	14
	Accuracy		0.76			0.77	
	Precision		0.65			0.65	
	Recall		0.80			0.82	
	F1		0.69			0.69	

4.1.3 Random Forest

Hyperparameter optimization is effective for the RF algorithm (Gordon et al., 2022), and the GridSearchCV function was used in RF hyperparameter optimization to obtain a maximum tree depth of 150, a maximum number of separated features of 12, a minimum number of separated samples of 2 and a total number of trees of 34. RF and MLR are distinguished from DNNs as a machine learning algorithm. However, unlike MLR and the basic building block of a RF (the DT), RFs are an ensemble learning algorithm. RFs and DT models were established and compared in Python's Sklearn framework, which is based on tuned parameters, and a cross-validation method was used to produce the results shown in Fig. 9. The classification results are shown in Table 2.

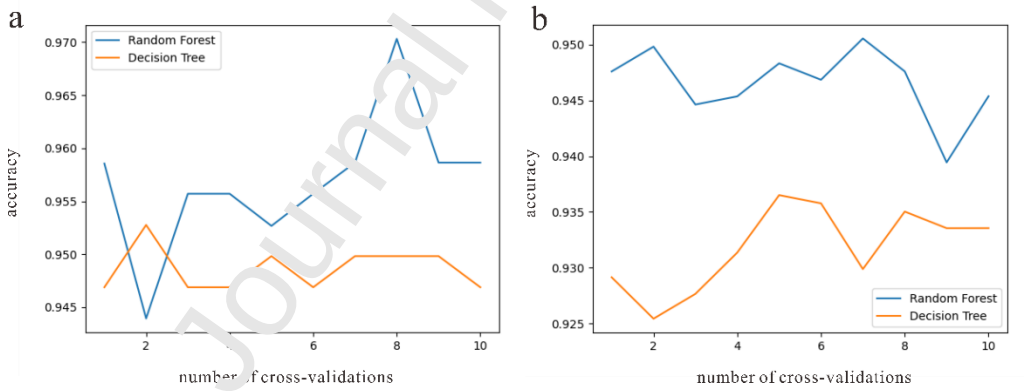

Fig. 9 Test accuracy of RF and DT under cross-validation. a. The relationship between the accuracy of RF and DT in cross-validation for 300 sample data set; b. The relationship between the accuracy of RF and DT in cross-validation for 1200sample data set.

Table 2 Radom Forest (RF) model test confusion matrix

		210 training data			840 training data		
	Prediction	Type I	Type II	Type III	Type I	Type II	Type III
Actual	Type I	8	10	0	7	11	0
	Type II	5	105	6	0	113	3
	Type III	2	1	13	1	6	9
	Accuracy		0.83			0.86	
	Precision		0.72			0.83	
	Recall		0.70			0.64	
	F1		0.71			0.70	

4.1.4 Deep fully connected neural network (DNN)

The neural network model is established under the Pytorch framework in Python (Hussain et al., 2021). Based on the neural network model parameters determined by an empirical formula, the model was trained in a loop using the abort accuracy rate, which was preset using the conditional statements in the model training function written by the author of this paper, and an optimal learning rate parameter of lr = 0.01 was finally obtained. The same data set of 300 samples was selected, and 70% of the data were extracted as training labels and the remaining 30% were used as test data. In the DNN training process, the loss function value converged after 300 training steps, and the test accuracy of the verification set was 0.97. To test the accuracy of the DNN model, the same method was used to process the 1200 data set, where 70% of the data were extracted as training data and 30% of the data was used as test data. The classification result is shown in Table 3. After 1200 training steps, the overall accuracy of both trained models in the test set reached 97%, as shown in Fig. 10.

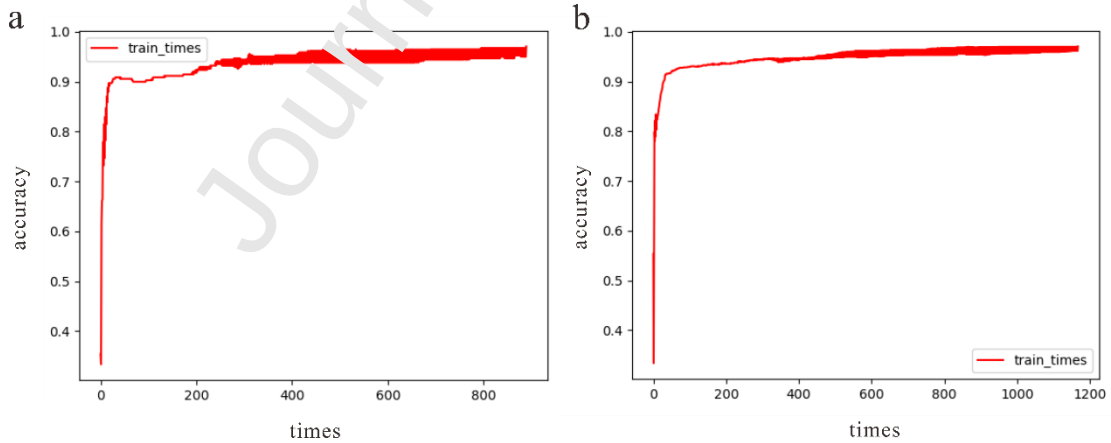

Fig. 10 Loss function value and test accuracy. a. The relationship between training times and accuracy for 300 sample data set; b. The relationship between training times and accuracy for 1200 sample data set.

Table 3 Neural network model (DNN) test confusion matrix

		210 training data			840 training data		
	Prediction	Type I	Type II	Type III	Type I	Type II	Type III

Actual	Type I	9	9	0	11	7	0
	Type II	4	109	3	5	108	3
	Type III	0	11	5	0	6	10
Accuracy			0.82			0.86	
Precision			0.72			0.78	
Recall			0.58			0.72	
F1			0.62			0.74	

4.2. Effectiveness and evaluation of MLR, RF, and DNN

Accuracy is based on the proportion of correct values among all samples, precision is based on the proportion of predicted correct values among all predicted values, recall is based on the proportion of predicted correct values among all correct values, and F1 is based on the average value of precision and recall (Ye et al., 2021). In the evaluation, metrics such as precision and recall can be used to assist the classification under conditions where the predicted values are close, and the classification sample is less balanced. When used in practical classification, precision focuses on how correct the prediction results are for the prediction set, and recall relates to how correct the prediction results are for the sample set. In practice, different indicators are used for data sets with relatively uneven samples, and comparisons can be made between different sample size models within a single algorithm. However, no cross comparisons should be made between different predictors. The accuracy, precision, recall, and F1 of MLR, RF, and the neural network algorithms using different samples are shown in Fig. 11.

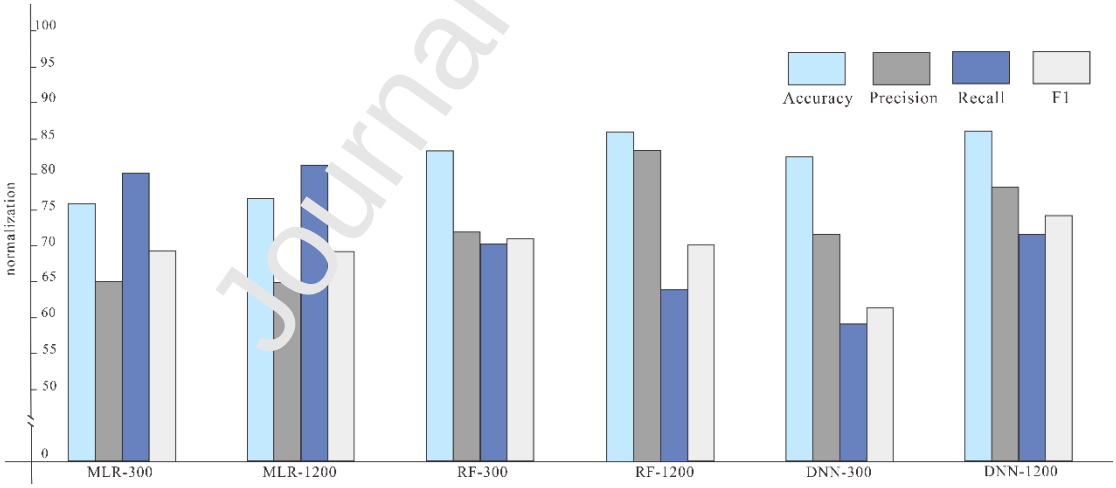

Fig. 11 Accuracy, precision, recall and F1 values of the classifiers at 300 and 1200 samples.

The overall accuracy of the MLR algorithms was low: 76% for 300 samples and 77% for 1200 samples. Unlike the other two classification methods used in coal structure logging data inversion, a maximum sample size is beneficial for the inversion process and ensuring the correctness of the data (Pino-Mejías et al., 2017). Both the DNN and RF algorithms showed higher prediction accuracies of 82% for 1200 samples. As both algorithms follow the effect of information gain, the prediction accuracy of both methods improved with an increasing number of samples. However, a

comparison of the accuracy, recall, and F1 of the models showed that the RF algorithm was superior when applied to 300 samples. The MLR algorithm does not provide the effect of information gain, but it is advantageous with respect to its computational cost. At the same random extraction scale, the difference arises from the underlying operation logic (Onifade et al., 2021). Its accuracy of over 76% makes it largely adequate for coal structure prediction in small sample sets, and it can be used in a combination of preprocessing algorithms with the other algorithms in large sample set predictions, to reduce the subsequent workload and allow for comparative validation. In cases where there are more information points and the available test set and training set are larger, the RF algorithm and deep learning approach are superior, and they are also beneficial for building a combination of deep learning algorithms, such as those involved in image recognition.

4.3. Single and multi-well discrimination results and evaluation

For the classified data, the scaling method was used to normalize the multi-attribute and different data sets, and a parallel coordinate system diagram was then compiled based on the classification (Type I, II, and III) under the group of 1200 sample data, as shown in Fig. 12.

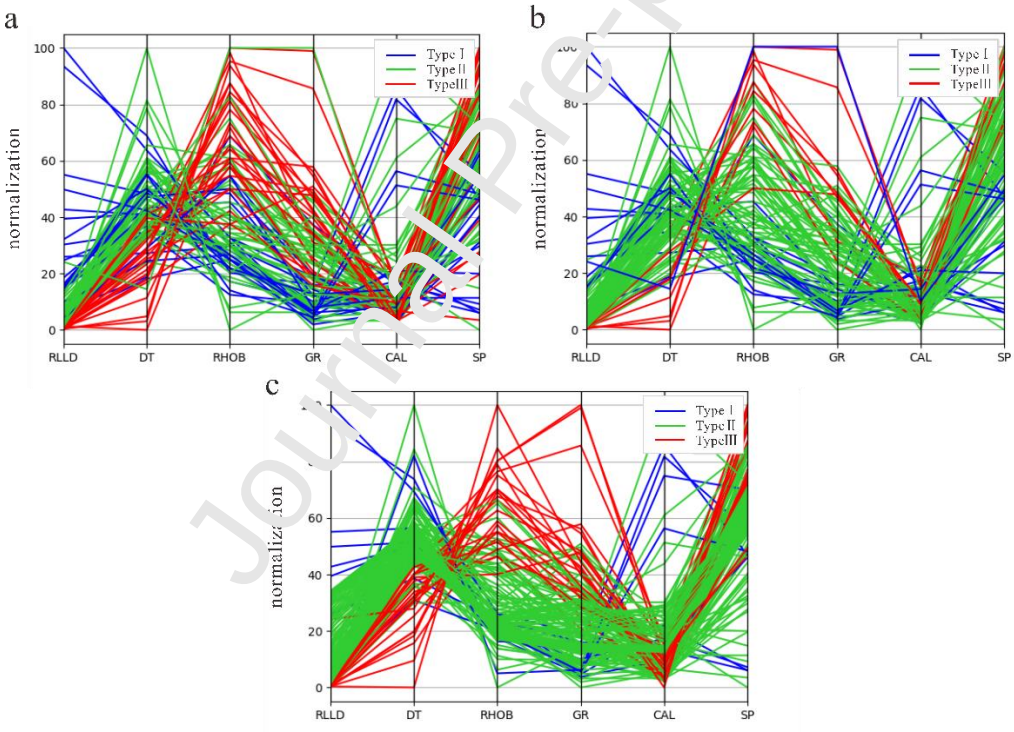

Fig. 12 Parallel coordinate system diagram of relationship between coal structure and logging data. A. Parallel Coordinate System under multinomial logistic regression; B. Parallel coordinate system under neural network; C. Parallel coordinate system under random forest.

For MLR classification, there were peaks with good relative discrimination for Types I, II, and III. However, the peaks for CAL and GR were obviously interlaced for different coal structure types and the correlation between them was not high. For RHOB, Type III had relatively high peak values, low-value areas, and a general classification effect. The distribution of the parallel

coordinate system of the DNN was relatively good. RLLD, GR, DT, RHOB were associated with different Type I, II, and II distributions, and the CAL peak under obvious scaling was relatively well distinguished between the Types I, II, and III categories; therefore, it had an overall good classification effect. The RF model provided the best classification and many large striped distributions were visible in the parallel coordinate system. The banded distribution reflected the consistency and high accuracy of the classification.

All of the models were used to identify the coal structure against the logging data from Well An22, as shown in Fig. 13. The classification accuracy of MLR was greater than 90%; although the classification error was very low, it can be further improved through the use of more samples. For the coal structure classification of Well A22, MLR may provide a small probability error, but it still determined the coal structure in different coal seams. The results of RF and DNN were 100% accurate in this test. However, compared with MLR, neural network classification has a relatively high training cost; however, high accuracy is assumed when data numbers are high enough. The RF algorithm not only showed the highest accuracy with 1200 samples, but it was also 100% accurate when predicting the coal structure of the An22 well using multiple 300 sample predictions.

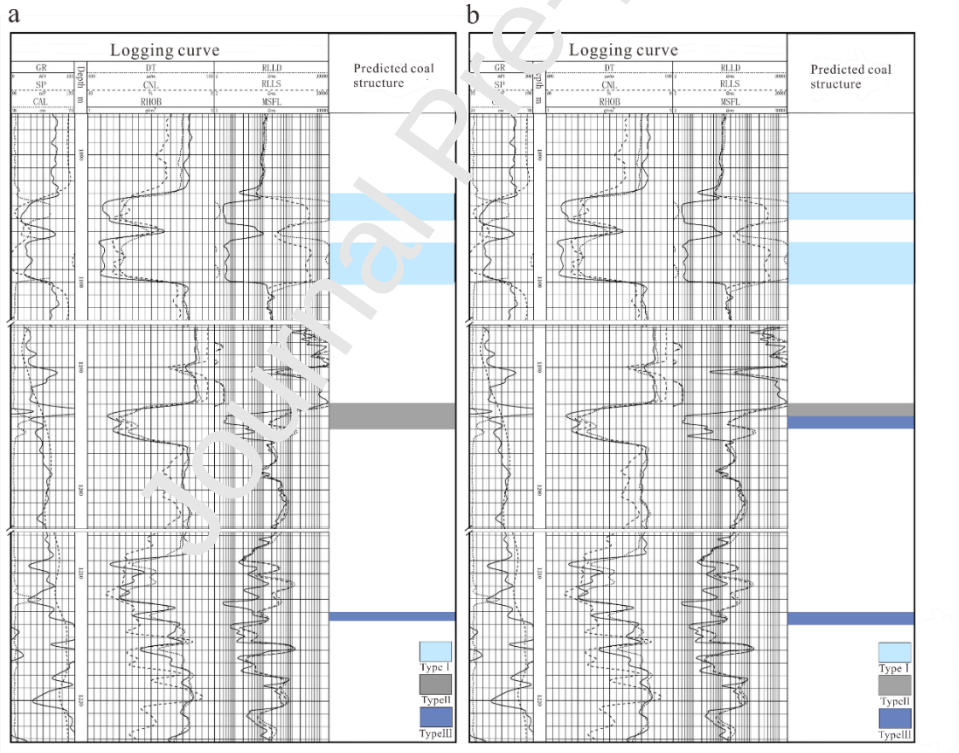

Fig. 13 Prediction chart of single well coal structure in Well An22. A. Prediction diagram of coal structure under neural network and random forest; B. Prediction diagram of coal structure under multiple logistic regression

The results of the comparison between coal structure predictions in multiple wells are shown in Fig. 14. The MLR algorithm results varied the most, and classification was biased toward a coal structure with greater fragmentation. There were only small differences between the results when using the DNN and RF algorithms. A single classifier can have varying degrees of error in the

classification predictions depending on the algorithm. The specific selection of classifiers for different situations can meet the needs of multi-well coal structure prediction.

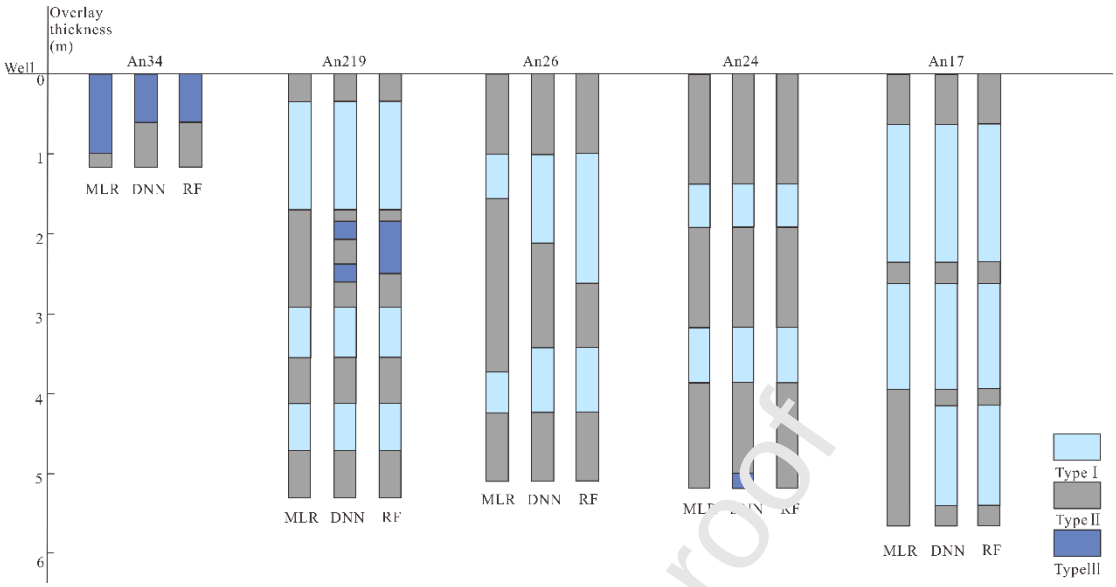

Fig. 14 Predicted charts of multiple well coal structure for wells An34, An219, An26, An24 and An17 under multiple classifiers.

4.4 Coal structure distribution and its geological controls

4.4.1 Coal structure distribution

Based on the scalability and the high accuracy of using the DNN algorithm, it was chosen as the algorithm to predict the coal structure in this region. Fig.15 shows that the Type III coal structure mainly occurs in areas with high structural strength, including folds and normal faults. Type I and Type II coal structures are distributed throughout the Anze area. Overall, the degree of coal fragmentation is higher as it approaches the area where the structural strength is the highest near the faults and folds (Danesh et al., 2022). In the Anze Block, 80% of the predicted Type III coal structure is developed within the near-fault zone. There are both rupture zones at faults and gentle zones inside graben barriers, which are mixed zones of high stress and low stress (Oliveira et al., 2022), and the accompanying coal structure thus differs. There are over 40 faults in the Anze Block, and only three of these are reverse faults. The fault distances range from 35 m to 130 m, which indicates that a strike-slip process occurs in this region. Other high angle normal faults are related to the graben barriers (Qiu et al., 2022). A comparison between two wells shows that the Type I coal structure in Well An219 has a thickness of 3 m, and that of Well An38 is less than 0.5 m, which are located at the lower plates of the fault. Two models have been established for coal structure development in these regional graben barriers. The main structural model of the uplift barrier is a strike-slip pull-apart basin that developed in the northern part of the Anze Block, and the other is the regional early widespread development of normal faults. The latter model is known as the restricted graben-barrier model, and this model development is consistent with normal fault formation where the graben is the active plate. Normally, the relative active plate in the graben

barrier is greatly affected by geostress, and this causes a higher broken plate. Therefore, for Well An219, much of the Type III coal structure developed due to the presence of the graben as an active plate.

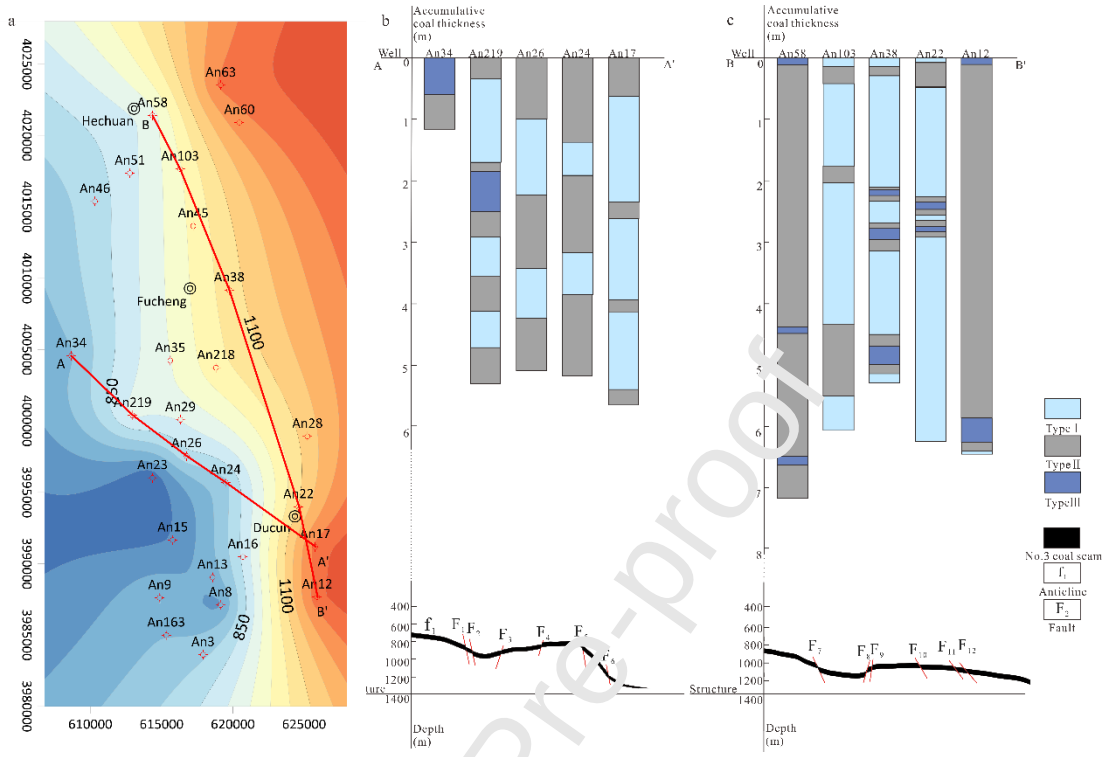

Fig.15 Profile of geological structure and coal structure distribution by DNN algorithm. A. No.3 coal burial depth contour map; B. Well An34 – Well An17 geological structure and burial depth; C. Well An58 – Well An12 geological structure and burial depth

4.4.2 Geostress

A series of tectonic models associated with the graben-barrier model have been developed to reflect the combined stresses in the deep and shallow parts (Liu et al., 2009; Li et al., 2021; Oliveira et al., 2022). However, fewer studies have focused on combining specific geological structural units, such as the graben barrier, with an evaluation of coal or CBM. Previous work (Kayseri-Özer et al., 2022) demonstrated that coal-bearing sediments can be used to interpret the paleogeography and paleoclimate under the graben unit. However, the graben only provides the structural features, and there has been less in-depth analysis of the relationship. Recently, Danesh et al., (2022) conducted an experimental investigation, established a relationship between the geological structure unit and CBM production under the graben barrier, and focused on the fault structure unit. In this study, we attempt to understand the geological controls on coal structure development in the Anze Block in relation to the graben-horst model and local faults and folds. The relationship between the regional geostress distribution and the distribution and location of wells is shown in Table 4.

Table 4 Maximum and minimum principal stresses in wells of different burial depths

Well	Depth(m	Max. stress (MPa)	Min. stress	Vertical	Lateral pressure

	)		(MPa)	stress (MPa)	coefficient
An34	739	18.8	11.78	14.78	1.04
An24	780	19.82	12.51	15.6	1.04
An26	826	20.92	13.34	16.52	1.04
An219	910	21.23	13.57	18.2	0.96
An58	915	23.06	14.94	18.3	1.04
An103	980	24.62	16.11	19.6	1.04
An38	1070	26.78	17.73	21.4	1.04
An22	1098	27.45	18.24	21.96	1.04
An17	1265	31.46	21.24	25.3	1.04
An12	1280	31.82	21.51	25.6	1.04

As shown in Table 4, there is a clear positive correlation between the magnitude of the geostress (measured in both directions of the wells) and the depth. However, as shown in Fig.15, highly fragmented coal is noted to occur at the shallowest depths and in areas such as in the core of the fold where there was low geostress in the late period. This shows that development of the coal structure originates from the destruction of a high fragmentation zone via stresses relating to fold formation. The later geostress was less developed and modified throughout the region. The area of coal fragmentation is located near the fault fragmentation zone, which was also the zone of high fault fragmentation when early stresses formed the fault. The control of the coal structure shows that the early geostress formed the structure and directly controlled the development of coal structure in the area, while the later geostress had a weaker influence on the development of highly fractured coal.

Coals in the strong geostress area generally have a severely fragmented structure. Of the logging data points, 14 in Well An24 were predicted to be Type III coal developed in the axial zones of the fold, and 15 in Well An219 were predicted to be Type III coal situated near the fold and fault. Due to the existence of two graben-barrier models, differences occur in coal structure. However, in areas of high geostress, such as in the axial of the fold, the fold slip zone of the near fault contains over 95% of the predicted Type III coal structure, which shows a high correlation with geostress. As shown in Fig. 15, the cumulative thickness of the coal structure reveals the graben and barrier models and the relationship between the coal structure and the regional structure. The local stress-strain characteristics show that folds and faults are areas of strong structural stress, are related to the Type III coal structure, and their deformation also has strong impacts on the coal structure type (Fig. 16). For example, the coal structure is severely broken at the fold axial, fault fracture zone, and in relation to the graben structure, where Type III coal is mostly developed. Additionally, the coal structure is more stable at the fold wing, fault plate, and barrier structure, where thicker Types I and II coal structures exist.

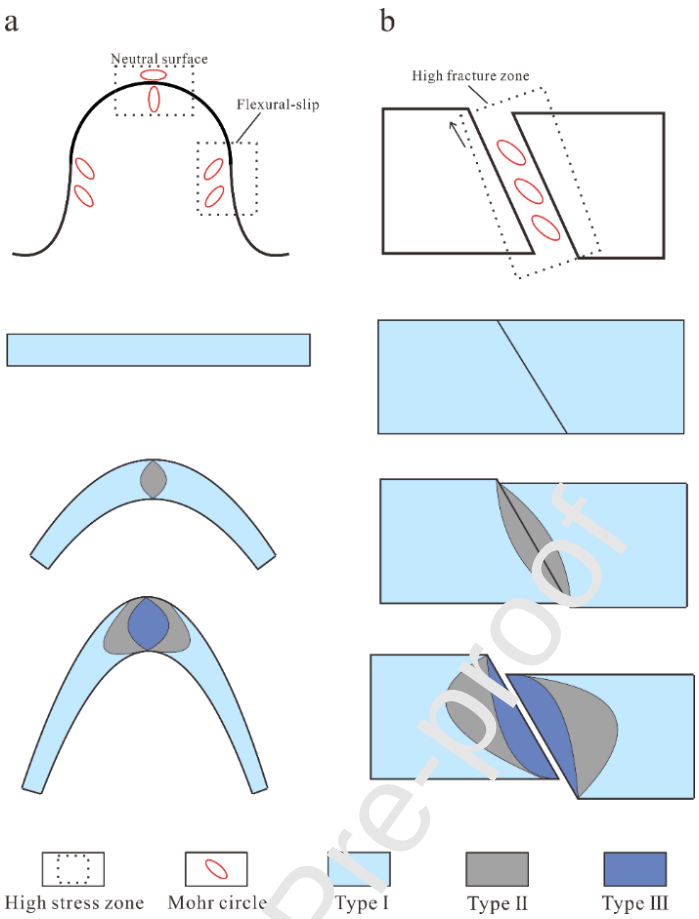

Fig. 16 Development pattern diagram of high strain zone and coal structure under structural control. A. Development pattern diagram of fold high strain zone and coal structure; B. Development pattern of fault high strain zone and coal structure

4.4.3 Coal thickness and burial depth

Several studies (Huang et al., 2017; Zhang et al., 2017) have investigated the relationship between coal thickness, coal burial depth, coal structure, and thermal movement. Fig. 17 shows that the Type II coal structure mostly appears in thick coal seams, and Type I structure coal occurs in seams found at larger burial depths.

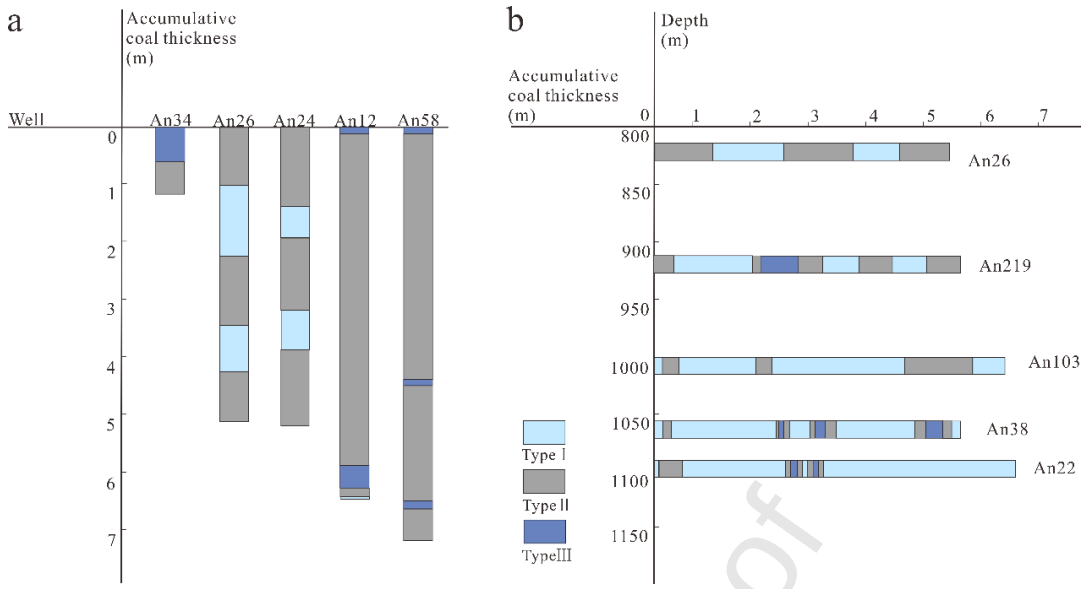

Fig. 17 Coal structure distribution with burial depth and coal seam thickness by DNN algorithm. A. Comparison chart of coal thickness and coal structure; B. Comparison chart of burial depth and coal structure

The thickness and burial depth of coal seams are closely related to geological structures. The thickest Type III coal structure is developed in the targeted No.3 coal seam of Well An34, which has the lowest structural coal thickness, and 60% of the coal has a Type III structure. However, the second thickest Type III coal structure is predicted in Well An12, where the second coal seam thickness is 6.5 m. From the perspective of burial depth, the Type I coal structure predominates from 800 m to 1100 m. However, changes in the Type III coal structure in Well An 103 do not follow the law of coal seam thickness and burial depth. Therefore, it is necessary to further explore the relationship between the coal structure and the structural unit.

In the Anze Block, the coal structure distribution is segmented into the three categories in accordance with the geological structure unit, coal seam thickness, and burial depth, as shown in Fig.18. The distribution mainly relates to the continuous deformation behavior from the fold hinge zone and the graben barrier fault zone to the tensile stress strong action zone. In a simple geological structure unit, this geological structure effect is not obvious. From Well An26 to Well An58, the thickness changes from 6 m to 8 m. The thickness of Type II coal structure changes from 70% to 90% in total, and Type III coal structure occurs in Well An58, reflecting the transition from stable zone to fracture zone. The same in Well An17 and An12, which Type II coal structure changes from 20% to 90%.

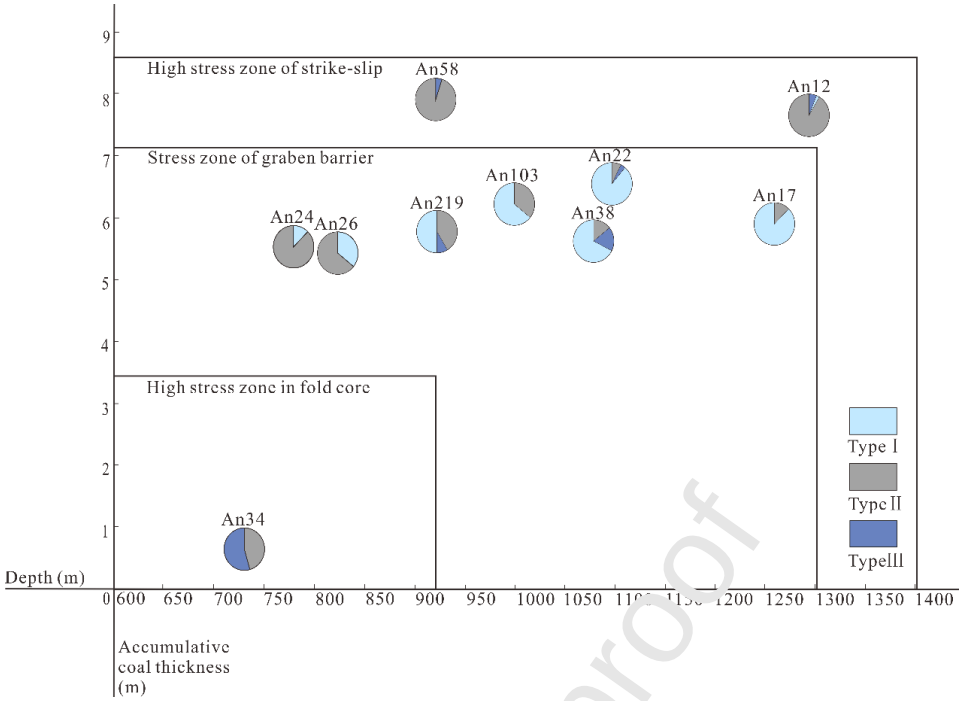

Fig. 18 Coal structure segment of CBM well with burial depth, coal seam thickness, and structure unit by DNN algorithm

The coal thickness is highly correlated with the geological structure in the Anze Block, and the effects of coal thickness on the coal structure can be directly evaluated. Additionally, the burial depth has a good correlation with the coal structure. For a simple geological structure, such as that of the graben barrier, the proportion of coal with a different structure has a good relationship with the burial depth. From a depth of 750 m to 1300 m, the thickness of the predicted Type II coal structure changes from 20% to 30% continuously in the stress zone of the graben barrier. Besides, it is also because the main fracture zone of the graben is the boundary fault zone, which causes a good development of coal seam thickness in the interior. That is, the structure unit can be simply deemed as the fracture development along the fault. The relationship between burial depth, coal thickness and coal structure is closely related to the geological structure. The combination of the above two parameters can determine the coal structure distribution. The higher the burial depth is, the higher the proportion of primary coal is. It is anticipated that the results of this study will assist in guiding the exploration and exploitation of CBM in deep coal seams.

5 Conclusions

In this study, the machine learning algorithms of MLR, RF, and DNN were used to interpret the coal structure based on multisource-logging data, with the aim of evaluating their use and optimizing the coal structure interpretation method when using different data set conditions. The effects of geological controls on the predicted coal structure were also explored. The following main conclusions can be drawn:

1) The MLR, RF, and DNN algorithms all provided good coal structure predictions, but the sensitivity of the three algorithms differed when applied to different datasets. The MLR algorithm

had weaker overall accuracy and was least sensitive to information gain and least resistant to noise; the DNN algorithm had better accuracy and a strong information gain; the RF had the best overall accuracy and strong noise resistance. For a sample group of 300, the accuracies of the MLR, RF, and DNN methods were 76%, 83%, and 82%, respectively, but for a group of 1200 samples, their accuracies were 77%, 86%, and 86%, respectively.

2) There is generally a high degree of coal fragmentation in the western part of the Anze Block, due to the existence of folds. Over 60% of the coal in Well A34 in the western part is Type III coal, and nearly 80% of coal in Well An17 in the eastern part is Type I. From north to south, the coal structure is mainly controlled by the active plate in the graben barrier. Although it is out of the control of folds, only a small amount of Type III coal exists in Wells An38 and An22, due to the high geostress near the active plate.

3) The main geological controlling factors of coal structure in the Anze Block are geostress, coal thickness, and buried depth. The control of geostress and coal seam thickness on the coal structure is reflected by the effect of tectonics on the coal structure. Type III coal is mainly developed near the strong geostress unit (the active plate of the graben barrier), and coals buried at different depths have generally the same structure. The higher the burial depth of targeted coal, the better the coal structure is preserved, and proportion of primary coal is larger.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant nos. 42130806, 41922016 and 41830427).

References

Abdul-Majeed, G. H., Y. Al Dunainawi, G. Soto-Cortes & J. A. Al-Sudani (2021) cc. SPE Journal, 26, 1290-1301.

Abrougui, K., K. Gabsi, B. Mercatoris, C. Khemis, R. Amami & S. Chehaibi (2019) Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). Soil & Tillage Research, 190, 202-208.

Bizhani, M. & E. Kuru (2022) Towards drilling rate of penetration prediction: Bayesian neural networks for uncertainty quantification. Journal of Petroleum Science and Engineering, 219, 111068.

Breiman, L. (1996) Bagging predictors. Machine Learning 24(2): 123-140.

Cai, Y. D., D. M. Liu, Y. B. Yao, J. G. Li & Y. K. Qiu (2011) Geological controls on prediction of coalbed methane of No. 3 coal seam in Southern Qinshui Basin, North China. International Journal of Coal Geology, 88, 101-112.

Cao, L. T., Y. B. Yao, D. M. Liu, Y. H. Yang, Y. J. Wang & Y. D. Cai (2020) Application of seismic curvature attributes in the delineation of coal texture and deformation in Zhengzhuang field,

southern Qinshui Basin. AAPG Bulletin, 104, 1143-1166.

Chatterjee, R. & S. Paul (2013) Classification of coal seams for coal bed methane exploitation in central part of Jharia coalfield, India - A statistical approach. Fuel, 111, 20-29.

Chen, S. D., P. C. Liu, D. Z. Tang, S. Tao & T. Y. Zhang (2021) Identification of thin-layer coal texture using geophysical logging data: Investigation by Wavelet Transform and Linear Discrimination Analysis. International Journal of Coal Geology, 239, 103727.

Danesh, N. N., Y. X. Zhao, T. Teng & M. S. Masoudian (2022) Prediction of interactive effects of CBM production, faulting stress regime, and fault in coal reservoir: Numerical simulation. Journal of Natural Gas Science and Engineering, 99, 104419.

El Sharawy, M. S. & B. S. Nabawy (2016). "Determination of electrofacies using wireline logs based on multivariate statistical analysis for the Kareem Formation, Gulf of Suez, Egypt." Environmental Earth Sciences 75(21): 1394.

Froncisz, M., Brown, P., & Weryk, R. J. (2020). Possible interstellar meteoroids detected by the Canadian Meteor Orbit Radar. Planetary and Space Science, 190, 104980.

Fu, X. H., Y. Qin, G. G. X. Wang & V. Rudolph (2009) Evaluation of coal structure and permeability with the aid of geophysical logging technology. Fuel, 88, 2278-2285.

Gao, X., Wang, Y., Ni, X., Li, Y., Wu, X., Zhao, S., & Yu, Y. (2018). Recovery of tectonic traces and its influence on coalbed methane reservoirs: A case study in the Linxing area, eastern Ordos Basin, China. Journal of Natural Gas Science and Engineering, 56, 414-427.

Gordon, J. B., Sanei, H., & Pedersen, P. K. (2022). Predicting hydrogen and oxygen indices (HI, OI) from conventional well logs using a Random Forest machine learning algorithm. International Journal of Coal Geology, 249, 103903.

Guo, Y. N., Z. R. Zhang & F. Z. Tang (2021) Feature selection with kernelized multi-class support vector machine. Pattern Recognition, 117, 107988.

Hashemizadeh, A., A. Maaref, M. Shateri, A. Larestani & A. Hemmati-Sarapardeh (2021) Experimental measurement and modeling of water-based drilling mud density using adaptive boosting decision tree, support vector machine, and K-nearest neighbors: A case study from the South Pars gas field. Journal of Petroleum Science and Engineering, 207, 109132.

Hernandez-Martinez, E., et al. (2013). "Facies Recognition Using Multifractal Hurst Analysis: Applications to Well-Log Data." Mathematical Geosciences 45(4): 471-48

Hoek, E. & E. T. Brown (1997) Practical estimates of rock mass strength. International Journal of Rock Mechanics and Mining Sciences, 34, 1165-1186.

Huang, S. P., D. M. Liu, Y. B. Yao, Q. Gan, Y. D. Cai & L. L. Xu (2017) Natural fractures initiation and fracture type prediction in coal reservoir under different in-situ stresses during hydraulic fracturing. Journal of Natural Gas Science and Engineering, 43, 69-80.

Hussain, N., A. A. Farooque, A. W. Schumann, F. Abbas, B. Acharya, A. McKenzie-Gopsill, R. Barrett, H. Afzaal, Q. U. Zaman & M. J. M. Cheema (2021) Application of deep learning to detect Lamb's quarters (Chenopodium album L.) in potato fields of Atlantic Canada.

Computers and Electronics in Agriculture, 182, 106040.

Ibrahim, A. F. (2022) Application of various machine learning techniques in predicting coal wettability for CO2 sequestration purpose. International Journal of Coal Geology, 252, 103951.

Imamverdiyev, Y., & Sukhostat, L. (2019). Lithological facies classification using deep convolutional neural network. Journal of Petroleum Science and Engineering, 174, 216-228.

Kayseri-Ozer, M. S. & T. Emre (2022) Palaeovegetation and paleoclimate in the SW Turkey - a study based on the early-middle Miocene coal-bearing sediments from the Buyuk Menderes Graben. Review of Palaeobotany and Palynology, 297, 104560.

Kumar, T., N. K. Seelam & G. S. Rao (2022) Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India. Journal of Applied Geophysics, 199, 104605.

Li, J. Q., D. M. Liu, Y. B. Yao, Y. D. Cai & Y. K. Qiu (2011) Evaluation of the reservoir permeability of anthracite coals by geophysical logging data. International Journal of Coal Geology, 87, 121-127.

Li, Z. T., D. M. Liu, Y. J. Wang, G. Y. Si, Y. D. Cai & Y. B. Wang (2021) Evaluation of multistage characteristics for coalbed methane desorption-diffusion and their geological controls: A case study of the northern Gujiao Block of Qinshui Basin, China. Journal of Petroleum Science and Engineering, 204, 108704.

Liu, D. M., Y. B. Yao, D. Z. Tang, S. H. Tang, Y. Che & W. H. Huang (2009) Coal reservoir characteristics and coalbed methane resource assessment in Huainan and Huaibei coalfields, Southern North China. International Journal of Coal Geology, 79, 97-112.

Liu, Y., Z. L. Zhang, X. Liu, L. Wang & X. H. Xia (2021) Deep learning-based image classification for online multi-coal and multi-class sorting. Computers & Geosciences, 157, 104922

Liu, Z., D. Zhu, H. Yang, W. Wang & W. Yang (2022) Experimental research on different metamorphic grades of coal bodies with macro–mesoscopic structure fractal characteristics. Geomechanics for Energy and the Environment, 100337.

Mastalerz, M., A. Drobniak, D. Strapoc, W. S. Acosta & J. Rupp (2008) Variations in pore characteristics in high volatile bituminous coals: Implications for coal bed gas content. International Journal of Coal Geology, 76, 205-216.

Maxwell, K., M. Rajabi & J. Esterle (2019) Automated classification of metamorphosed coal from geophysical log data 10 using supervised machine learning techniques. International Journal of Coal Geology, 214, 103284

Nazmi, S., X. Y. Yan, A. Homaifar & E. Doucette (2020) Evolving multi-label classification rules by exploiting high-order label correlations. Neurocomputing, 417, 176-186.

Oliveira, M. E., A. S. Gomes, F. M. Rosas, J. C. Duarte, G. S. Franca, J. C. Almeida & R. A. Fuck (2022) Impact of crustal rheology and inherited mechanical weaknesses on early continental

rifting and initial evolution of double graben structural configurations: Insights from 2D numerical models. Tectonophysics, 831, 229281.

Onifade, M., A. I. Lawal, J. Abdulsalam, B. Genc, S. Bada, K. O. Said & A. R. Gbadamosi (2021) Development of multiple soft computing models for estimating organic and inorganic constituents in coal. International Journal of Mining Science and Technology, 31, 483-494.

Pearson, K. (1895) Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58(347-352), 240-242.

Pino-Mejias, R., A. Perez-Fargallo, C. Rubio-Bellido & J. A. Pulido-Arcas (2017) Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO2 emissions. Energy, 118, 24-36.

Qin, Z. H. (2018) New advances in coal structure model. International Journal of Mining Science and Technology, 28, 541-559.

Qiu, H., S. Deng, J. Zhang, H. Lin, C. Huang, J. Han, W. Lin & Y. Zhu (2022) The evolution of a strike-slip fault network in the Guchengxu High, Tarim Basin (NW China). Marine and Petroleum Geology, 140, 105655.

Raeesi, M., A. Moradzadeh, F. D. Ardejani & M. Rahimi (2012) Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data and artificial neural networks. Journal of Petroleum Science and Engineering, 82-83, 151-165.

Ren, P. F., H. Xu, D. Z. Tang, Y. K. Li, C. H. Sun, S. Tao, S. Li, F. D. Xin & L. K. Cao (2018) The identification of coal texture in different rank coal reservoirs by using geophysical logging data in northwest Guizhou, China: Investigation by principal component analysis. Fuel, 230, 258-265.

Rock, N. M. S. (1987) Corank - A Fortran-77 Program To Calculate And Test Matrices Of Pearson, Spearman, And Kendall Correlation-Coefficients With Pairwise Treatment Of Missing Values. Computers & Geosciences, 13, 659-662.

Shi, J. X., L. B. Zeng, S. Q. Dong, J. P. Wang & Y. Z. Zhang (2020) Identification of coal structures using geophysical logging data in Qinshui Basin, China: Investigation by kernel Fisher discriminant analysis. International Journal of Coal Geology, 217, 103314.

Sinha, S., Kiran, R., Tellez, J., & Marfurt, K. (2019). Identification and Quantification of Parasequences Using Expectation Maximization Filter: Defining Well Log Attributes for Reservoir Characterization. Paper presented at the URTEC 2019.

Siregar, I., Y. F. Niu, P. Mostaghimi & R. T. Armstrong (2017) Coal ash content estimation using fuzzy curves and ensemble neural networks for well log analysis. International Journal of Coal Geology, 181, 11-22.

Teng, J., Y. B. Yao, D. M. Liu & Y. D. Cai (2015) Evaluation of coal texture distributions in the southern Qinshui basin, North China: Investigation by a multiple geophysical logging method. International Journal of Coal Geology, 140, 9-22.

Tiwary, A. K., S. Ghosh, R. Singh, D. P. Mukherjee, B. U. Shankar & P. S. Dash (2020) Automated coal petrography using random forest. International Journal of Coal Geology, 232, 103629.

Wang, G., D. Y. Han, X. J. Qin, Z. Liu & J. F. Liu (2020) A comprehensive method for studying pore structure and seepage characteristics of coal mass based on 3D CT reconstruction and NMR. Fuel, 281, 118735.

Wang, H., Lu, S., Qiao, L., Chen, F., He, X., Gao, Y., & Mei, J. (2022). Unsupervised contrastive learning for few-shot TOC prediction and application. International Journal of Coal Geology, 259, 104046.

Wang, L. L., Z. J. Long, Y. Song & Z. H. Qu (2022) Supercritical CO2 adsorption and desorption characteristics and pore structure controlling mechanism of tectonically deformed coals. Fuel, 317, 123485.

Wang, Y., Bai, X., Wu, L., Zhang, Y., & Qu, S. (2022). Identification of maceral groups in Chinese bituminous coals based on semantic segmentation models. Fuel, 308, 121844.

Wang, Z. Z., J. N. Pan, Q. L. Hou, B. S. Yu, M. Li & Q. Y. Niu (2018) Anisotropic characteristics of low-rank coal fractures in the Fukang mining area, China. Fuel, 211, 182-193.

Wei, H., Luo, K., Xing, J., & Fan, J. (2022). Predicting co-pyrolysis of coal and biomass using machine learning approaches. Fuel, 310, 122248.

Welper, G. (2022) Universality of gradient descent neural network training. Neural Networks, 150, 259-273.

Wojtecki, Å. u., Iwaszenko, S., Apel, D. B., Bukowska, M. a., & MakÃwka, J. (2022). Use of machine learning algorithms to assess the state of rockburst hazard in underground coal mine openings. Journal of Rock Mechanics and Geotechnical Engineering, 14(3), 703-713.

Xiao, D., Le, T. T. G., Doan, T. T., & Le, B. T. (2022). Coal identification based on a deep network and reflectance spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 270, 120859.

Xu, A. K., H. M. Chang, Y. J. Xu, R. Li, X. Li & Y. Zhao (2021) Applying artificial neural networks (ANNs) to solve solid waste-related issues: A critical review. Waste Management, 124, 385-402.

Yang, H., W. Y. Bi, Y. G. Zhang, J. K. Yu, J. W. Yan, D. J. Lei & Z. N. Ma (2021) Effect of tectonic coal structure on methane adsorption. Journal of Environmental Chemical Engineering, 9.

Ye, Z., Guo, S., Chen, D., Wang, H., & Li, S. (2021). Drilling formation perception by supervised learning: Model evaluation and parameter analysis. Journal of Natural Gas Science and Engineering, 90, 103923.

Zhang, J. Y., D. M. Liu, Y. D. Cai, Z. J. Pan, Y. B. Yao & Y. J. Wang (2017) Geological and hydrological controls on the accumulation of coalbed methane within the No. 3 coal seam of the southern Qinshui Basin. International Journal of Coal Geology, 182, 94-111.

Zhang, L., Song, Z., Wu, D., Luo, Z., Zhao, S., Wang, Y., & Deng, J. (2022). Prediction of coal self-ignition tendency using machine learning. Fuel, 325, 124832.

Zhang, Y., Wang, J., Yu, Z., Zhao, S., & Bei, G. (2022). Research on intelligent detection of coal gangue based on deep learning. Measurement, 198, 111415.

Zhao, B. L., S. N. Hu, X. M. Zhao, B. N. Zhou, Li, W. Huang, G. H. Chen, C. N. Wu & K. Liu (2022) The application of machine learning models based on particles characteristics during coal slime flotation. Advanced Powder Technology, 33, 103363.

Author Statement

Manuscript Title: Intelligent Classification of Coal Structure Using Multinomial Logistic Regression, Random Forest and Fully Connected Neural Network with Multisource Geophysical Logging Data

Under supervision by Dr. Yidong Cai and Dr. Dameng Liu, Zihao Wang performed sample preparation, data, modelling analysis. Fengrui Sun and Feng Qiu performed sample preparation and structure fabrication. Yingfang Zhou performed language revisions. All authors read and contributed to the manuscript.

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: