Advances in Geo-Energy Research

Vol. 9, No. 1, p. 25-37, 2023

Original article

Multi-label prediction method for lithology, lithofacies and fluid classes based on data augmentation by cascade forest

Ruiyi Han[1], Zhuwen Wang[1], Yuhang Guo[1]*, Xinru Wang[1], Ruhan A[1], Gaoming Zhong[2]

[1]*College of Geo-Exploration Science and Technology, Jilin University, Changchun 130021, P. R. China*
[2]*Northeast Oil and Gas Branch of Sinopec, Changchun 130000, P. R. China*

Keywords:
Data augmentation
deep learning
igneous rock
multi-label learning

Cited as:
Han, R., Wang, Z., Guo, Y., Wang, X., A, R., Zhong, G. Multi-label prediction method for lithology, lithofacies and fluid classes based on data augmentation by cascade forest. Advances in Geo-Energy Research, 2023, 9(1): 25-37.
https://doi.org/10.46690/ager.2023.07.04

Abstract:
Predicting the lithology, lithofacies and reservoir fluid classes of igneous rocks holds significant value in the domains of CO_2 storage and reservoir evaluation. However, no precedent exists for research on the multi-label identification of igneous rocks. This study proposes a multi-label data augmented cascade forest method for the prediction of multi-label lithology, lithofacies and fluid using 9 conventional logging data features of cores collected from the eastern depression of the Liaohe Basin in northeastern China. Data augmentation is performed on an unbalanced multi-label training set using the multi-label synthetic minority over-sampling technique. Sample training is achieved by a multi-label cascade forest consisting of predictive clustering trees. These cascade structures possess adaptive feature selection and layer growth mechanisms. Given the necessity to focus on all possible outcomes and the generalization ability of the method, a simulated well model is built and then compared with 6 typical multi-label learning methods. The outperformance of this method in the evaluation metrics validates its superiority in terms of accuracy and generalization ability. The consistency of the predicted results and geological data of actual wells verifies the reliability of our method. Furthermore, the results show that it can be used as a reliable means of multi-label prediction of igneous lithology, lithofacies and reservoir fluids.

1. Introduction

Igneous reservoirs are widely distributed in Mesozoic-Cenozoic terrestrial and marine basins, featuring global development. The highly inhomogeneous nature of igneous reservoirs leads to large variations in reservoir recovery rates, which poses a challenge to reservoir evaluation and CO_2 storage (Cai et al., 2022; Xiao, 2022; Zhang et al., 2022). The mode of volcanism, eruption type, magma composition, geological structure, and paleogeography all exert influences on reservoir recovery. Thus, the accurate determination of lithology, lithofacies and reservoir fluid classes is crucial for the quantitative assessment of igneous reservoirs. In the Eastern Depression of the Liaohe Basin in northeastern China, igneous rock reservoirs of different scales are being developed across various Cenozoic strata. These reservoirs are representative in terms of their development environments, rock types and production distributions.

Intelligent algorithms, which have been popular in igneous lithology identification, mainly use conventional logging data. For example, Xiang et al. (2020) employed a depth confidence network combined with conventional logging curves to identify igneous rocks in the eastern part of the Junggar Basin, and accurately identified thin layers of dense basalt and dense trachyte that conventional logging interpretations could not distinguish due to data resolution problems. Kuhn et al. (2020) utilized the random forest method to identify intrusive rocks in the volcanic terrain of British Columbia. Duan et al. (2020) proposed a combination of conventional logging, imaging logging and the decision tree method for volcanic lithology identification.

In terms of igneous lithofacies classification, Huang et al. (2014) proposed an igneous lithofacies delineation method

Yandy Scientific Press

*Corresponding author.
E-mail address: hanry20@mails.jlu.edu.cn (R. Han); wangzw@jlu.edu.cn (Z. Wang); guoyuhang100@jlu.edu.cn (Y. Guo); wxr22@mails.jlu.edu.cn (X. Wang); ruhan19@mails.jlu.edu.cn (R. A); zhonggm.dbsj@sinopec.com (G. Zhong).

Received May 30, 2023; revised June 19, 2023; accepted July 5, 2023; available online July 9, 2023.

based on drilling rock chips, logging and seismic data by combining the geological characteristics of igneous rocks in the Liaohe Basin. Giordano and Cas (2021) derived an igneous lithofacies classification scheme based on volume and diffuse area correlation. Various intelligent algorithms have been applied to lithofacies classification. For instance, Liu et al. (2020) devised a lithofacies identification method based on multi-resolution image clustering. Ehsan and Gu (2020) proposed a lithofacies identification scheme for Takhar shale by combining neuro-fuzzy system, cross-plot and statistical analysis. Chang et al. (2021) proposed a geophysical logging segmentation network method for lithofacies identification, which has shown excellent application value in the Bohai Bay Basin. Falivene et al. (2022) developed a convolutional neural network based on semantic partitioning architecture to identify lithofacies in cores.

Fluid identification relies heavily on pore structure, energy spectrum analysis and cross-plot. Yue and Tao (2006) proposed a reservoir fluid identification method based on wavelet transform energy spectrum analysis. Zhang et al. (2008) used cross-plot and the three-porosity curve overlap method to identify igneous gas formations. However, the above studies have been limited to considering the fluid properties under a certain lithology, while ignoring the influence of lithology and lithofacies on fluids in volcanic systems.

Furthermore, the previous research only considered single-label predictions for a single property, and there is currently no precedent for using multi-label methods to simultaneously predict the lithology, lithofacies, and reservoir fluid classes of igneous rocks. Multi-label learning is a learning method where a sample has multiple labels at the same time. Compared to multiple single-label classification scenarios, multi-label classification can make better use of information in the dataset, thereby improving the performance of the model. In geoscience, multi-label learning has been applied to airborne remote sensing (Li et al., 2022), formation delineation (Ho et al., 2023) and mineral identification (Wu et al., 2022).

In reservoir evaluation, the lithology, lithofacies and fluid class labels of each core together form a multi-label instance, which comprises a typical multi-label unbalanced dataset. such unbalanced datasets are commonly processed by methods such as oversampling, undersampling, hybrid sampling, threshold shifting, and machine learning. Zhou et al. (2020) used the synthetic minority over-sampling technique method to balance the dataset and combined it with gradient boosting decision tree for lithology identification. He et al. (2020) performed oversampling based on the theory of inheritance and the Mahalanobis distance to process unbalanced dense sandstone reservoir logging data for logging phase identification. Zheng et al. (2023) used K-means combined with the synthetic minority oversampling technique method to reconstruct the dataset and combined it with the Markov chain to improve the Bayesian inversion method to identify lithofacies of extremely inhomogeneous reservoirs.

This paper proposes a multi-label data augmented cascade forest (MLDACF) to accurately identify the lithology, lithofacies and fluid classes of volcanic reservoirs in the Eastern Depression. In this method, multi-label synthetic minority oversampling technique (MLSMOTE) is used to achieve data augmentation on an unbalanced multi-label dataset, and a cascade forest integrated with predictive clustering trees (PCT) is used as a learner to achieve classification. After validating the effectiveness of the model, it is applied to the igneous strata in the eastern depression of the Liaohe basin, and the identification results are discussed in relation to each other. The data show that the MLDACF outperforms alternative methods and provides a reliable solution for the multi-label identification of igneous reservoir lithofacies, lithology and fluid classes.

2. Study area and data

The Liaohe Basin is a Cenozoic terrestrial rift basin in northeastern China, which belongs to the northern branch of the Bohai Bay Rift system. Its terrestrial part consists of three uplifts and four depressions in a total of seven secondary tectonic units (Fig. 1(a)). This basin sits on basement rocks, which are Tertiary, Metasedimentary, Paleozoic and Mesozoic from bottom to top. The Eastern Depression is an active rift depression formed under the action of Tanlu fault and uplift of the upper mantle, which also has the most complicated geological conditions in the Liaohe Basin, covering an area of 2,300 square kilometers (Fig. 1(b)). In the early period, due to the northward movement of the Pacific plate and the leftward activity of the Tanlu fault, volcanic activity in the eastern depression increased toward the north (Busby and Bassett, 2007). Igneous rocks there were mainly formed in the first, second and third sections of the Shahejie Formation and the Dongying Formation. Over time, repeated volcanic eruptions occurred in the eastern depression, which were mainly of the spillover type, and the eruption center shifted to the north. The strong volcanic activity developed a large number of igneous rocks, resulting in a complex tectonic pattern in the eastern depression (Liu et al., 2022).

Igneous rocks of the Eastern Depression can be divided into three main categories: Volcanic lava, volcanic clastic rocks and sub-volcanic rocks. Among these, volcanic lavas are widespread, which mainly include medium-basic volcanic rocks, such as trachyte, trachyandesite and basalt. Volcanic clastic rocks are less common, with volcanic clastic lava and volcanic conglomerate being the most widespread. The sub-volcanic rocks are mainly distributed in the Shahejie Formation and are mainly diabase.

Since the Mesozoic, the eastern depression has undergone several phases of magmatism. The volcanic activity of the Eastern Depression was stronger during the period of the Fangshenbao Formation; the volcanic activity increased toward the north, and the stratigraphic lithology was characterized by thick-layered basalt with thin-layered clastic rocks. The central part of the Eastern Depression was the center of volcanic eruption during the Eocene Shahejie Formation, and its eruption was controlled by the NE-oriented deep fracture, forming three superimposed oval volcanic rock bodies in the Rehetai-Oulituozi-Huangshatuo area. This covers an area of about 90 square kilometers. In this period, the formation of the oil-bearing layer and volcanic reservoir occurred in the

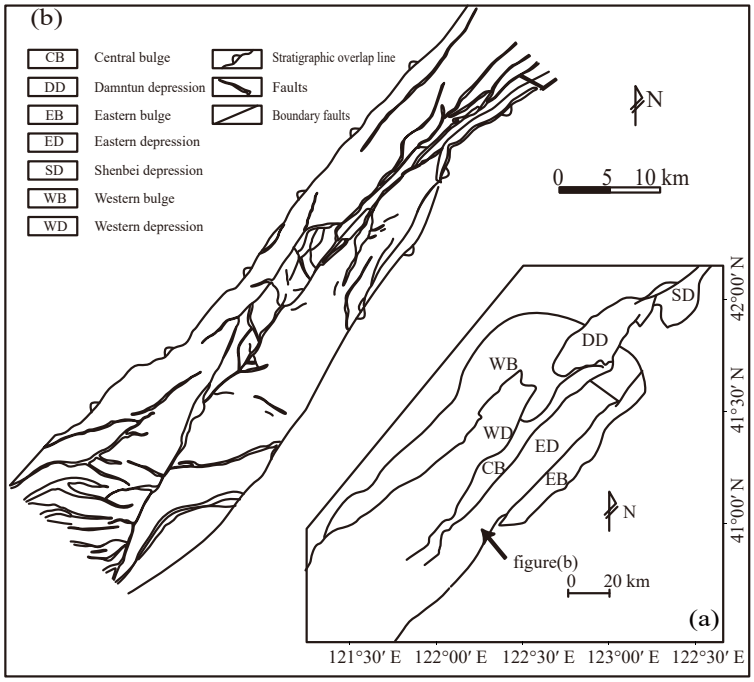

Fig. 1. Location of the study area. (a) Tectonic zoning of the Liaohe Basin and (b) distribution of Eastern depressional faults.

area. The Paleogene Shahejie Formation can be divided into three stages: The main igneous rock in the first stage is tuff, the trachyte widely developed in the second stage, and the basalt mainly formed in the third stage. The basin extension and the slide-slip effect of the Tanlu Fracture Zone from the Eocene Shahejie Formation to the Paleogene Shahejie Formation are weakened, and the volcanic activity of the eastern depression is low, but there are still some basaltic magma eruptions. The Tanlu fracture zone of the Dongying Formation entered another period of strong activity, and the strong right-slip movement led to active volcanism again in the southern part of the Eastern Depression, forming a huge thick volcanic rock system-mainly basaltic rocks-with a local development of trachyte (Liu et al., 2022).

Fig. 2 illustrates the common igneous cores and core thin sections in the Eastern Depression, with andesite shown in Fig. 2(a). The core in the figure is glassy in structure and with diamond-shaped fractures. The thin section identification reveals that the form of porphyry is pyroxene, the matrix is dominated by brown volcanic glass, a few plagioclases, pyroxene microcrystals, glass crystal interwoven structure, and with opaque dark minerals. Trachyandesite is presented in Fig. 2(b). In the pore amygdaloidal structure, the pores filled by calcite also show a slight amount of zeolite filling, with the development of irregular fissures. According to the thin section identification, the type of porphyry is mainly plagioclase, with a minor amount of orthoclase. The matrix is mainly a glassy interwoven structure, and the plagioclase is partially altered to clay minerals and calcite. Trachyte is shown in Fig. 2(c). The core is porphyritic; the porphyritic crystals are feldspar and contain a modest amount of plagioclase. Plagioclase can be seen as obvious ring bands, the rock is severely broken in the horizontal direction, a large amount of calcite is distributed on the broken surface, and some silica-filled fractures are seen. Fig. 2(d) depicts diabase. The thin section identification shows that its main minerals are plagioclase and pyroxene, with coarser crystallization, and pyroxene mostly contains altered chlorite. The plagioclase feldspar is pockmarked and filled with clay minerals, and a slight amount of black opaque magnetite is seen. Volcanic breccia, lava and basalt are shown in Figs. 2(e)-2(g), respectively. In basalt, the core is glassy, massive, with quenched breccia structure, developing blast fractures, irregular reticulated veins, contraction joints, filled with calcareous, and later oblique intersections mainly filled with silica and zeolite. The thin section features a porphyritic texture with a pore structure containing mainly olivine (15%) porphyritic crystals, as well as pyroxene porphyritic crystals, interstitial matrix with microcrystalline plagioclase, and characterized by a glassy matrix.

Considering that there are four modes of magmatism in the Liaohe Basin: Eruption, overflow, extrusion, and intrusion, there are also four in situ environments: Closed, semi-open, open, and aquatic. Based on the magmatic mode of action and environmental differences, Huang et al. (2014) classified the intermediate basal volcanic rocks in the region into explosive facies, effusion facies, volcanic sedimentary facies, volcanic conduit facies, intrusive facies, and extrusive facies. Feng et al. (2016) completed a volcanic lithofacies mapping of the eastern depression using drilling data to constrain seismic data (Fig. 3). The results showed that two main sedimentary sequences exist in the area: The first set of sequences includes volcanic conduit facies, extrusive facies, effusion

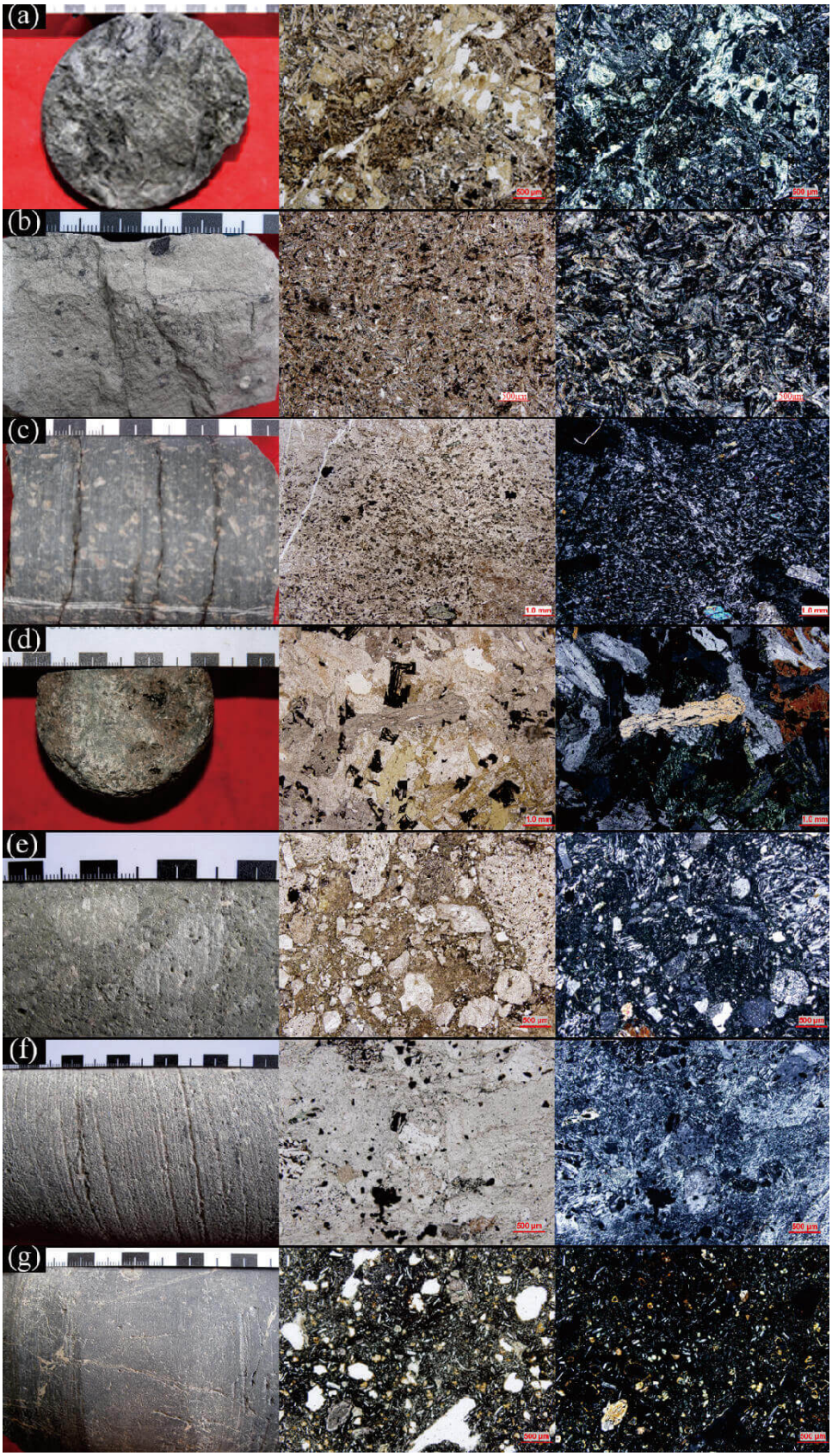

Fig. 2. Photographs of cores and core thin sections of common igneous rocks in the Eastern Depression (orthogonal polarization and single polarization). (a) Andesite, (b) trachyandesite, (c) trachyte, (d) diabase, (e) volcanic breccia, (f) lava and (g) basalt.

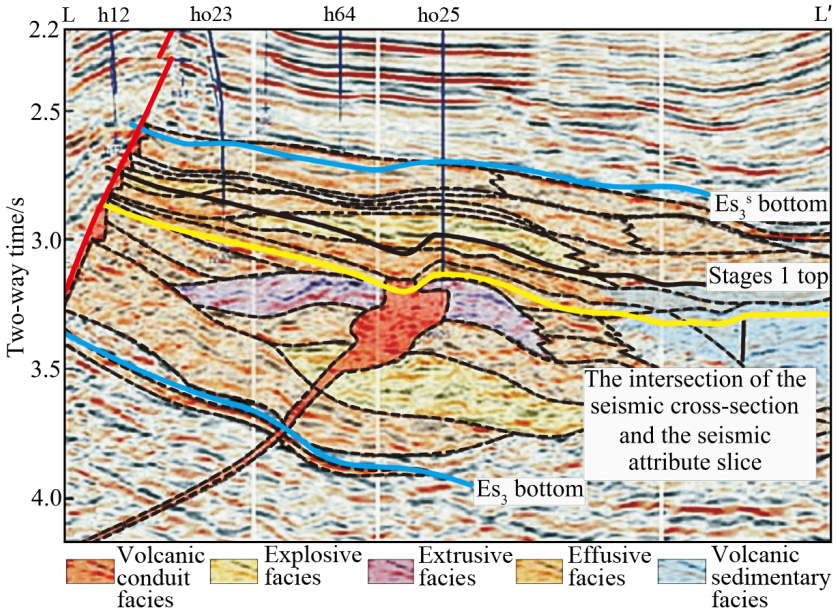

Fig. 3. Seismic interpretation profile of volcanic lithofacies of the Eastern depression (Feng et al., 2016).

Table 1. Classification features of dataset labels.

Lithofacies	Lithology	Fluid
Explosive	Trachyte, diabase, volcanic breccia	Dry layer, oil layer, low-yield oil layer
Volcanic conduit	Trachyandesite, basalt, lava	Dry layer, oil layer, low-yield oil layer
Extrusive	Trachyte	Dry layer, low-yield oil layer
Effusion	Andesite, trachyandesite, volcanic breccia, lava, basalt, trachyte	Dry layer, oil layer, low-yield oil layer

facies, and volcanic sedimentary facies, while the other set includes volcanic conduit facies, effusion facies and volcanic sedimentary facies. Additional results revealed that the near-caldera assemblage of volcanic conduit facies, extrusive facies and explosive facies have excellent hydrocarbon signatures. Therefore, based on the actual conditions of the reservoir, a total of four lithologies were classified in this study: Effusion facies, explosive facies, extrusive facies, and volcanic conduit facies (Yue et al., 2021). Table 1 shows the classification characteristics of the dataset labels.

In order to construct the dataset, this study used the cores and core thin section identification results from consecutive cores as labels for the lithology and the lithofacies. The formation test results were taken as labels for fluid classes. A total of 9 features were included in the data, namely, acoustic, borehole diameter, compensated neutron-porosity logging, density, gamma ray log, deep lateral resistivity (RLLD), shallow lateral resistivity (RLLS), micro lateral resistivity (RMLL), and spontaneous potential (SP).

For the comparison of model generalization performance for all possible outcomes, this study constructed simulated well datasets with all possible outcomes. After selection, two multi-label datasets were assembled from a total of 2,557 instances in 27 different depth intervals from 13 wells in the Eastern Depression. Of these, 2,337 and 220 instances were used to form the training set and simulated well dataset, respectively. The training set and simulated wells are not mutually exclusive in terms of well views, as they have data from different depth intervals of the same well. It is important to note that the spatial linkage of the dataset sources may interfere with the prediction results to some extent and cannot adequately evaluate the generalization ability of the model. Therefore, to validate the model application effect, this study also forms actual well datasets that are mutually exclusive with the training set and simulated wells. However, it is critical to note that the actual wells will only show some of the results and not all of the labels will be of interest.

The multi-label dataset consists of three sets of labels: Lithology, lithofacies and fluids. Fig. 4 depicts the Circos plot generated from the training set data. Circos diagrams are circular charts that precisely show the correlation between variables. They are widely used to visualize genomic data (Krzywinski et al., 2009). In this study, this plot type was used to demonstrate the complexity and imbalance of the multi-label dataset. The Circos diagram uses different colors and widths to represent the different features, and the presence or absence of the different features are denoted by lines. Each feature corresponds to a sector of a circle, and the size of the sector will be scaled according to the number of rock types, with larger numbers of features taking up a larger

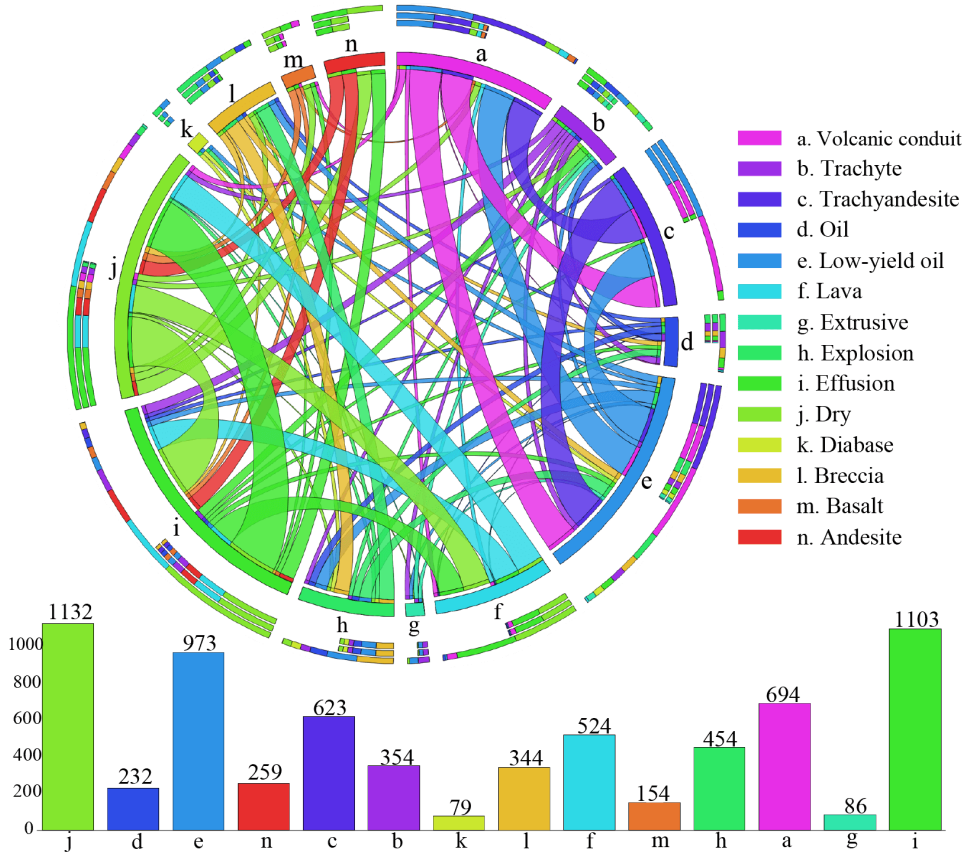

Fig. 4. Circos plot drawn for the training set to demonstrate the interactions of different series of labels. The histogram reflects the instance size counts.

sector area. Each line starting from the center represents a specific feature. These feature lines are attached to the circle corresponding to the rock type, and the thickness and color of the line can indicate the extent or number of the feature present in the different features. The size of the arc for each label reflects the frequency of that label. The dry layer and effusion facies labels occupy a significant portion of the arcs, indicating that these labels, as the majority class of the dataset, have a more dramatic effect on the other minority classes of labels. In the histogram, considering the label series, the minority class labels are obviously the oil layer, diabase, basalt, and extrusive facies. In fact, basalt is a common igneous lithology in the study area, and the distribution of dataset labels does not correspond to the actual situation in the area, which may lead to training effects deviating from the actual expectation. Therefore, the balance of the dataset should not be neglected.

3. Method

3.1 Multi-label data augmentation cascade forest

In multi-label classification, each instance outputs not only a single feature but also a set of feature vectors consisting of labels (Zhang and Zhou, 2013). In classification scenarios where the number of samples in some classes differs significantly from that in other classes, this problem is called unbalanced learning. Typically, the design of the learner to pursue global accuracy will sacrifice the less representative minority class dataset in the classification. At the same time, noise labeling complicates the effect on the classification results in unbalanced learning. The classifier is less effective when using unbalanced datasets, while label imbalance is common in multi-label data. The common approaches to solve multi-label classification at this stage include three ideas: Resampling, algorithmic adaptation and cost-sensitive classification in the data preprocessing step. As a branch of resampling techniques, original sample generation has proven to be superior to alternative methods (Lopez et al., 2013).

In single-label imbalance learning, the synthetic minority oversampling technique is the most popular oversampling method, and its core idea is to interpolate between the nearest neighbors of minority class samples to generate a certain number of original samples, that is, the data to reach the class equilibrium state. In multi-label classification, the presence of multiple labels leads to an increase in the number of minority class groups. At the same time, the need to generate label vectors rather than individual labels in multi-label classification poses a challenge to the rebalancing process. The MLSMOTE (Charte et al., 2015) has been shown to have significant advantages in dealing with extremely unbalanced classifications (Zhang et al., 2018). It uses the imbalance ratio (I) and the mean imbalance (M) ratio to identify minority class groups, defined as:

Algorithm 1: Multi-label synthesis of minority oversampling methods.

Input: D
k

begin
  $Y \leftarrow$ labels in D
  $M \leftarrow$ calculate $M(D,Y)$
  foreach y ***in*** Y **do**
    $I(y) \leftarrow$ calculate $I(D,y)$
    if $I(y) > M$ **then**
      $D_{\min} \leftarrow$ get all instances of y
      $D_i \leftarrow$ sample i in D
      for i ***in*** $D_{\min}$ **do**
        $d \leftarrow$ calculate distance $(D_i, D_{\min})$
        sort smaller to largest (d)
        $n \leftarrow$ get head items (d,k)
        $r_n \leftarrow$ get rand neighbor (n)
        $D_n \leftarrow$ new sample (D_i, r_n, n)
        $D = D + D_n$
      end
    end
  end
end

$$I(y) = \frac{\underset{y=Y_1}{\overset{Y_{|Y|}}{\arg\max}} \left(\sum_{i=1}^{|D|} X(y, Y_i) \right)}{\sum_{i=1}^{|D|} X(y, Y_i)}, X(y, Y_i) = \begin{cases} 1 & y \in Y_i \\ 0 & y \notin Y \end{cases} \quad (1)$$

where D represents the multi-label data, X denotes an indicator function, Y denotes the feature vector, Y_1 represents the first label in the Y vector, and i is the ordinal number of the sample. The parameter $I(y)$ is the ratio of majority class label samples to y label sample counts in the dataset, which is used to reflect the imbalance of the dataset:

$$M = \frac{1}{|Y|} \sum_{y=Y_1}^{Y_{|Y|}} (I(y)) \quad (2)$$

In MLSMOTE, M is used as the threshold value to judge the minority class samples. When the I value of a label is greater than M, this means that the count of that label is lower than the average count of the labels within the multi-label data, and this label is considered a minority class label.

Each label partitions the multi-label data into two subsets via M. A small number of samples are set in which each sample determines the set of several nearest neighbors based on the Euclidean distance, where k is used to denote the number of nearest neighbors. Furthermore, a new sample is generated by interpolating the concatenation of the samples with their nearest neighbors. The new samples are compared with the label counts in the reference sample and the neighboring labeled samples, and the new sample set is considered feasible when the labels in the new sample set are present in most of the comparison samples. The pseudo-code for the data enhancement process in this study is detailed in Algorithm 1.

After obtaining the balanced dataset, the data also need to be pre-processed. The model effect is inversely proportional to the difference in data distribution across the labels (Galar et al., 2011a), so the RLLD, RLLS, and RMLL data are logarithmically transformed. The SP data are also normalized to a single well. To significantly reduce the effect of differences in data distribution on the model, the entire dataset is normalized, so that it follows a standard normal distribution:

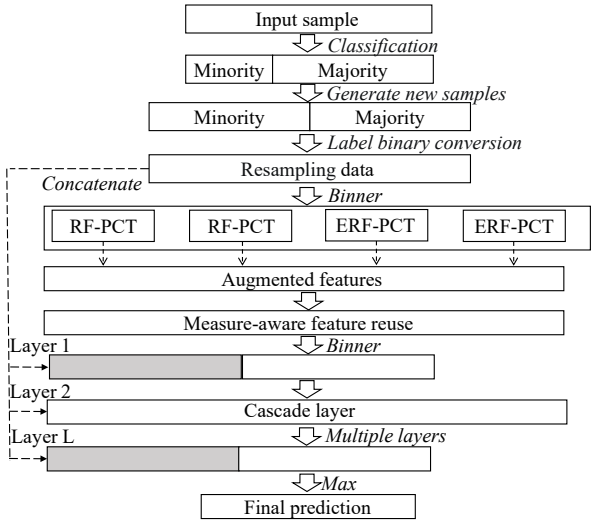

Fig. 5. MLDACF schematic diagram.

$$Z = \frac{D - \bar{D}}{S} \quad (3)$$

where $\bar{D}$ represents the mean of the data, Z represents the standard value, and S is the standard deviation of the dataset.

In addition to preprocessing the dataset, the label set must also be converted from multivariate labels to binary labels to accommodate multi-label learners and evaluation metrics (Galar et al., 2011b). This means that each label corresponds to a property. When the result is 1, the instance is considered to have this property, while when the result is 0, the instance is considered not to have this property in the current model. When the result is 0 for all of the instances in a set of labels, the instance is considered to be others.

After preprocessing, the dataset and labels are fed into the cascade forest. The multi-label cascade forest is a tree-integrated multi-label deep learning method (Yang et al., 2020). It has a multi-layer cascade structure, with each layer consisting of multiple complete random forests and regular random forests. The basic unit of random forest is the predictive clustering tree. The complete random forest uses random features for branching, and the regular random forest selects the features with the highest Gini coefficients. The feature vector output from several random forests in each layer is used as the input for the next layer, and after several layers, the mean of the output results is used as the final prediction result. Fig. 5 shows a schematic diagram of MLDACF.

The multi-label cascade forest uses the metric-aware feature reuse method in each layer of the cascade structure to determine the availability of the output feature vector. The metric-aware feature reuse effectively improves the model effectiveness in terms of evaluation metrics. The stopping rule

Table 2. Performance of back-judgment models.

Algorithm	Hamming loss	One-error	Coverage	Ranking loss	Average precision	F1	Macro-AUC
MLDACF	0.0079	**0.003**	0.1477	**0.0033**	0.9929	0.9794	**0.9998**
MLExtraTrees	0.0024	0.0036	0.1504	0.0111	0.9927	0.994	0.9941
MLRidgeCV	0.0788	0.0716	0.4387	0.3138	0.7771	0.5977	0.7685
MLRF	0.1174	0.2017	0.5712	0.4966	0.6602	0.3571	0.6592
MLMLP	0.1415	0.1874	0.6088	0.5943	0.5952	0.2678	0.6229
MLKNN	**0.0012**	0.0036	**0.1456**	0.0044	**0.9976**	**0.9976**	0.998
MLDT	0.189	0.0179	0.1897	0.0498	0.9579	0.9489	0.9679

of the cascade forest is metric-aware layer growth, which sets an initial model evaluation metric P. In this study, Hamming loss is used as a metric to evaluate the stopping mechanism. This metric measures the number of incorrectly predicted labels as a percentage of the number of labels in all samples. For each additional layer of the model, the current model evaluation metric P_c is used to compare with P. If P_c is better than P, the model is updated; if P_c is better than P, P_c is updated to P; if P_c is not updated, the model moves to the next layer. The stopping mechanism is triggered when the P_c is not updated for three consecutive cascade levels. Metric-aware layer growth can prevent the occurrence of overfitting. In this study, the number of nearest neighbors is set to 5, the number of forests is set to 4, and the number of PCT in each forest is set to 15.

3.2 Comparison methods and evaluation indexes

Six typical multi-label machine learning methods are used for comparison, of which multi-label Decision Trees (MLDT), multi-label Extra Trees (MLExtraTrees), and multi-label Random Forest (MLRF) are represented as tree-based learners; Multi-label K-Nearest Neighbor (MLKNN) is a typical multi-label learning method (Zhang and Zhou, 2007); Multi-label Ridge with cross-validation (MLRidgeCV) is represented as a support vector machine learner; and multi-label Multi-layer Perceptron (MLMLP) is a neural network multi-label method (Chu et al., 2023). The parameter settings of the different methods are listed below. For the tree-based learner, the random states are all set to 0 to make the stochastic process consistent. For MLExtraTrees, the number of trees is set to 1,000. For MLRF, the maximum depth is set to 2. For MLKNN, the number of nearest neighbors is set to 3. For MLRidgeCV, alpha is set to 1e-3, 1e-2, 1e-1, and 1. For MLMLP, alpha is set to 1e-5, 5 neurons are set for the first hidden layer, and 2 neurons are set for the second hidden layer.

Performance evaluation in multi-label learning is a more complex process than in traditional single-label learning. The traditional, commonly used metrics such as accuracy and F-measure are not suitable for direct use. To evaluate the performance, this study uses seven widely used multi-label evaluation metrics (Wu and Zhou, 2017), namely, Hamming loss, one-error, coverage, ranking loss, average precision, F-measure (F1), and the macro area under curve (macro-AUC). These evaluation metrics are considered from different perspectives of modeling effectiveness. F1 and macro-AUC are labeled perspective evaluation metrics. Among the other instance-based evaluation metrics, Hamming Loss is concerned with the correctness of each element and is a classification-based evaluation metric, while the rest are all ranking-based metrics. For Hamming Loss, ranking loss, one-error, and coverage, the smaller the value, the better the model performance. For the other metrics, the larger the value, the better the model performance.

4. Model validation

4.1 Model performance

During training, 30% of the instances in the training set were randomly selected as the test set to evaluate the model, and the model was subjected to a random sampling process 10 times to take the mean of the evaluation metrics. The final performance of the models is shown in Table 2. The optimal evaluation indicators are bolded. First, MLDACF and MLExtraTrees perform well in most metrics, such as Hamming Loss, one-error and average precision, which means that they are able to predict and rank labels accurately. In addition, these two models also perform well in F1 and macro-AUC metrics, indicating that they have advantages in balancing precision and recall. Second, MLRidgeCV and MLRF perform slightly worse. Although they perform moderately well on some metrics, their higher one-error, coverage and F1 values and their lower average precision and macro-AUC compared to the first two models imply that they have some difficulties in predicting the correct labels. Finally, MLMLP, MLKNN and MLDT are three models with poor performance in most metrics, with MLMLP in particular performing the worst in terms of F1 and macro-AUC, showing a clear challenge in balancing accuracy and recall. In summary, based on these performance metrics, MLDACF and MLExtraTrees are the best choices, which perform well in several metrics and demonstrate higher accuracy and recall in the label prediction task. Considering the possibility of overfitting in the back-judgment model, the generalization ability of the model needs to be evaluated.

In order to further evaluate the model effectiveness, all labels were considered, and to evaluate the generalization ability

Table 3. Performance of different simulated wells.

Algorithm	Hamming loss	One-error	Coverage	Ranking loss	Average precision	F1	Macro-AUC
MLDACF	0.0565	**0.05**	**0.1906**	**0.0336**	**0.9308**	0.853	**0.9816**
MLExtraTrees	**0.0316**	0.0868	0.2495	0.1312	0.9155	**0.931**2	0.9416
MLRidgeCV	0.1438	0.2466	0.5698	0.5064	0.5064	0.4901	0.7092
MLRF	0.1722	0.3973	0.6976	0.6924	0.5197	0.2406	0.5894
MLMLP	0.1882	0.4064	0.7061	0.7418	0.4883	0.2334	0.5963
MLKNN	0.0385	0.0776	0.2166	0.0973	0.9252	0.9147	0.951
MLDT	0.0691	0.0959	0.2945	0.1771	0.8515	0.7852	0.8781

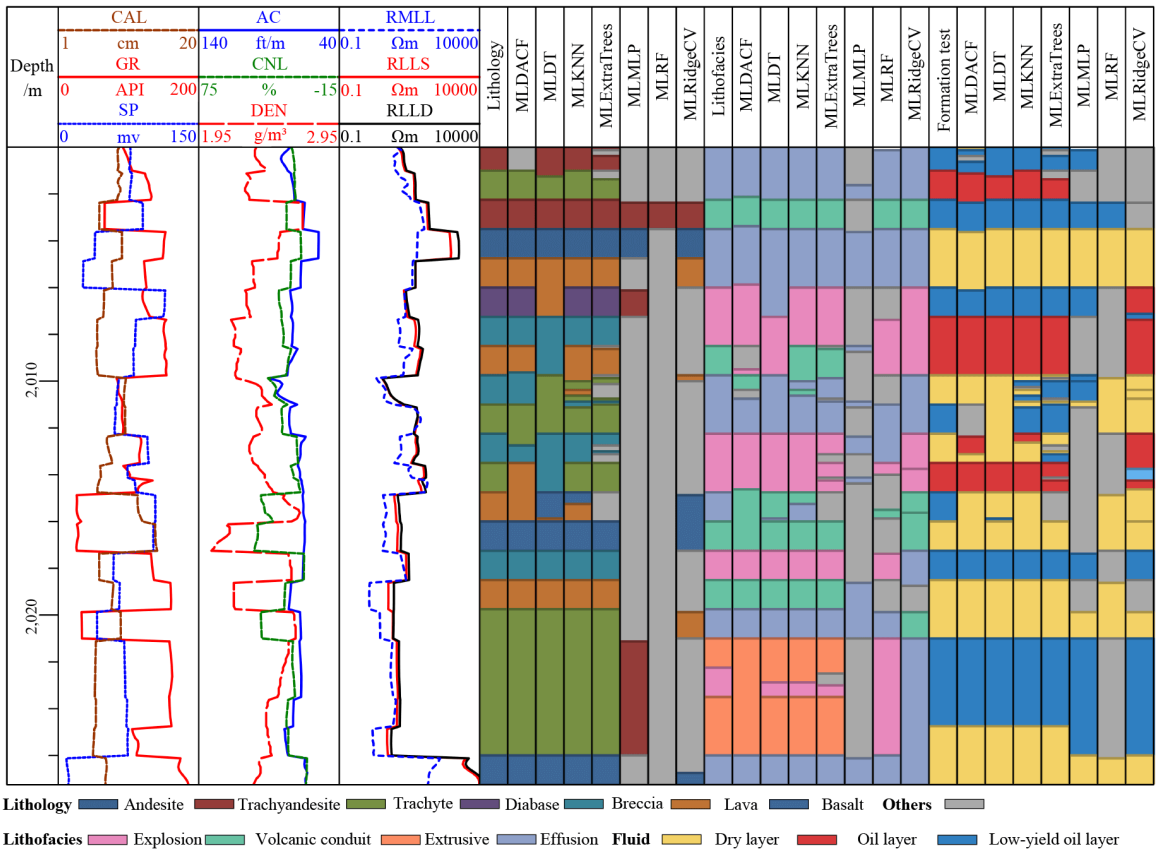

Fig. 6. Graph of predicted results of simulated well.

of the model, multi-label predictions were performed using the simulated well dataset. Table 3 includes the comparison of the prediction performance of simulated wells for each model. It can be found that five of the seven metrics are superior to MLKNN. MLExtraTrees is slightly inferior for Hamming loss and F1 metrics. MLDACF performs the best in all four evaluations metrics based on ranking. In addition, MLDACF performs the best on the label-based macro-AUC metrics.

Fig. 6 shows a graph of the prediction results of the simulated wells, providing a visual comparison of the actual prediction results of each method under each label series. For lithology identification, MLRidgeCV, MLRF and MLMLP identify the vast majority of instances as others. MLExtraTrees has multiple layer sections and yields identification results as others. MLKNN has a large number of identification errors in the 2,007.5-2,016 m interval. The identification errors of MLDT are mainly concentrated in the 2,006-2,016 m interval. MLDACF identifies some of the trachyandesite as other lithologies and some of the trachyte as lava. In lithofacies identification, MLRF, MLMLP and MLRidgeCV produce a large number of identification errors. MLExtraTrees identifies instances of six layers as others. MLKNN identifies some of the effusion instances as volcanic conduit, and also it identifies the explosion instances as extrusive. MLDT yields the same error. MLDACF partially identifies volcanic conduit facies as other lithofacies and partially identifies effusion facies as volcanic conduit facies. In terms of fluid identification, MLRF, MLMLP, MLExtraTrees, and MLRidgeCV produce

a large number of identification errors. In the layer segment below 2,018 m, the identifications of MLDACF, MLDT and MLDACF are highly accurate, but all three methods feature dry layer identification errors. MLDACF partially identifies low-yield oil layers as other fluids and dry layers, and partially identifies dry layers as oil layers. Overall, the simulated well identification by MLDACF is in line with the expectations.

4.2 Discussion of model effects

Combining the model back-judgment performance and simulated well performance, the evaluation metrics of each method become less effective in the simulated well data. As shown in Fig. 7, the mean absolute deviation is the average of the absolute values of changes in the evaluation metrics of each model. The smaller the mean absolute deviation in this study, the weaker the model degradation effect after using the simulated well dataset, and the stronger generalization performance. The mean absolute deviation of MLDACF is 0.0536, which is the lowest among the seven methods. In contrast, MLKNN and MLExtraTrees outperform MLDACF in some of the model metrics, but MLDACF dominates in more metrics and has better generalization performance.

5. Practical strata applications

In order to verify the applicability of the MLDACF model, it was applied to actual wells. Fig. 8 shows the prediction results for 2,921-3,179 m in well A. The results show that the main lithofacies in the basaltic section is the effusion facies and the reservoir is the dry layer. With increasing depth, the lithofacies of the trachyte section changes to explosive facies, and the reservoir is mainly oil layer accompanied by some dry and low-yielding oil layers. From the predicted results, the appearance of oil can be related to the change in lithofacies and lithology. The results from successive cores show that there is a transition from basalt to trachyte in lithology from top to bottom, and from effusion facies to explosive facies in lithofacies from top to bottom. This part of explosive pyroclastic flow moves along the surface with the hot clastic mixture under the promotion of the subsequent ejecta and its own gravity, forming a high-density gravity flow accumulation. This type of stratum is usually accompanied by the development of intergranular and dissolution pores. The identification results of core at 3,001.1 m indicate that the lithology is trachytic breccia lava interspersed with trachyte, and this lithology transition may be the reason why part of the lithology is predicted to be different. There is a collapse of well diameter at 3,040 m, resulting in a significant change in the density curve, so the lithology is identified as other lithologies. In general, the prediction of lithology and lithofacies in well A is basically consistent with the core data, and the predicted fluid property results are consistent with the formation test results.

Fig. 9 illustrates the predicted results for 2,636-2,684 m in well B, which show that this formation section is trachyte that changes from effusion facies to explosive facies from top to bottom. In the effusion facies section, it mainly consists of dry layer and low-yield oil layer, and when the lithofacies changes

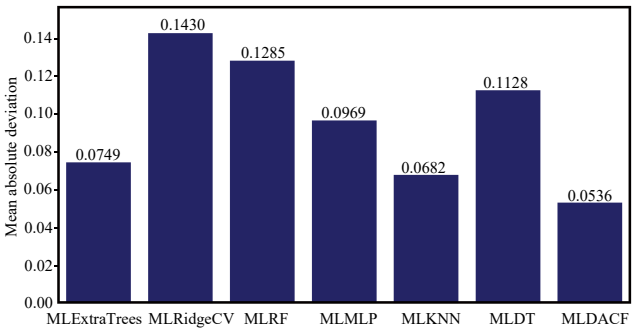

Fig. 7. Mean absolute deviation of the model in the evaluation metrics of the two datasets.

to explosive facies, the reservoir changes to oil layer. From the predicted results, the transition from dry layer to oil layer can be related to the transition of lithofacies. The continuous coring results show that the upper part of the interval is dense trachyte that transitions to trachyte. The lithofacies changes from effusion facies to explosive facies from top to bottom. The upper dense part may be the result of the rapid flow of magma along the surface after the overflow of this section from the channel to form a thick slab lava flow. As the lithofacies changes to explosive facies, the slab lava flow becomes a massive lava flow. At the same time, core identification data from 2,682.44 m show that an alteration has occurred there, which may be the reason for the change in reservoir fluid class. Combined with the curve, the lithology of the upper part of the formation is predicted to be different, probably due to dense formation. In general, the predicted lithology and petrography results of well B are basically consistent with the core data, and the predicted fluid properties are consistent with the formation test results.

6. Conclusions

In volcanic systems, the lithology, lithofacies and fluids of igneous rocks are correlated. Using conventional logging data, this study proposes a multi-label method for lithology, lithofacies and fluid identification. This method employs the MLSMOTE for data augmentation and the multi-label cascade forest for prediction. The MLDACF performs optimally in one-error, ranking loss and macro-AUC in the back-judgement model. In the simulated well, MLDACF performs the best in terms of one-error, coverage, ranking loss, average precision, and macro-AUC. Meanwhile, the lowest mean absolute deviation among all mentioned methods means that MLDACF has the best generalization ability. Overall, MLDACF has the best prediction among the mentioned methods. In actual wells, MLDACF predictions are consistent with the formation tests and core results, and the data show that MLDACF can be used as a reliable method for the multi-label prediction of igneous lithology, lithofacies and reservoir fluids.

The reasons for the “other” formation prediction results were further analyzed using real data, and it was found that more data support is needed to reduce this effect, which will be the focus of our next research.

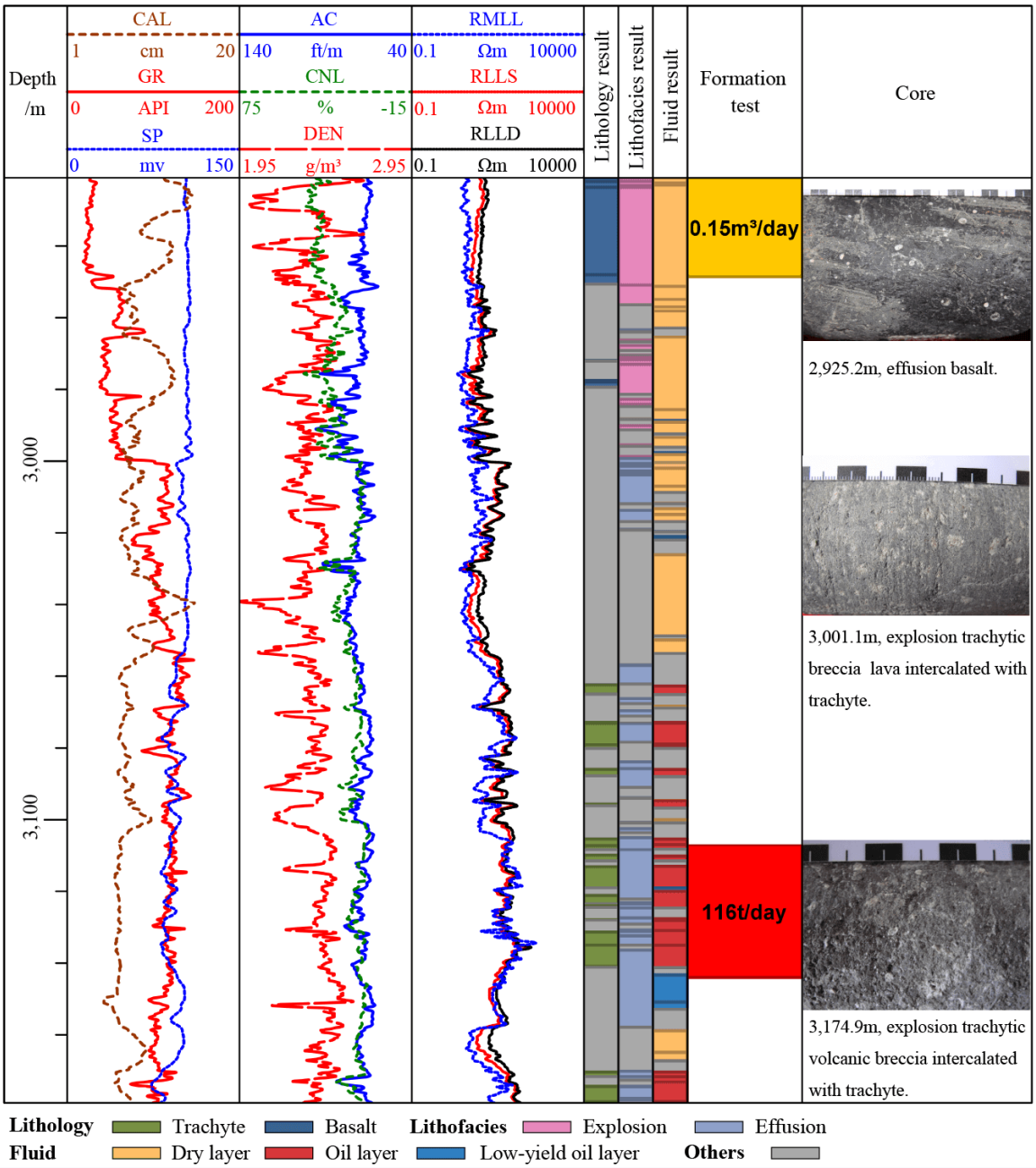

Fig. 8. Predicted results of well A for 2,921-3,179 m.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Nos. 42204122, 41874135, and 41790453) and the "Deep-time Digital Earth" international grand science program.

Conflict of interest

The authors declare no competing interest.

References

Busby, C. J., Bassett, K. N. Volcanic facies architecture of an intra-arc strike-slip basin, Santa Rita Mountains, southern Arizona. Bulletin of Volcanology, 2007, 70(1): 85-103.

Cai, J., Zhao, L., Zhang, F., et al. Advances in multiscale rock physics for unconventional reservoirs. Advances in Geo-energy Research, 2022, 6(4): 271-275.

Chang, J., Li, J., Kang, Y., et al. SegLog: Geophysical logging segmentation network for lithofacies identification. IEEE Transactions on Industrial Informatics, 2021, 18(9): 6089-6099.

Charte, F., Rivera, A. J., del Jesus, M. J., et al. MLSMOTE: Approaching imbalanced multilabel learning through

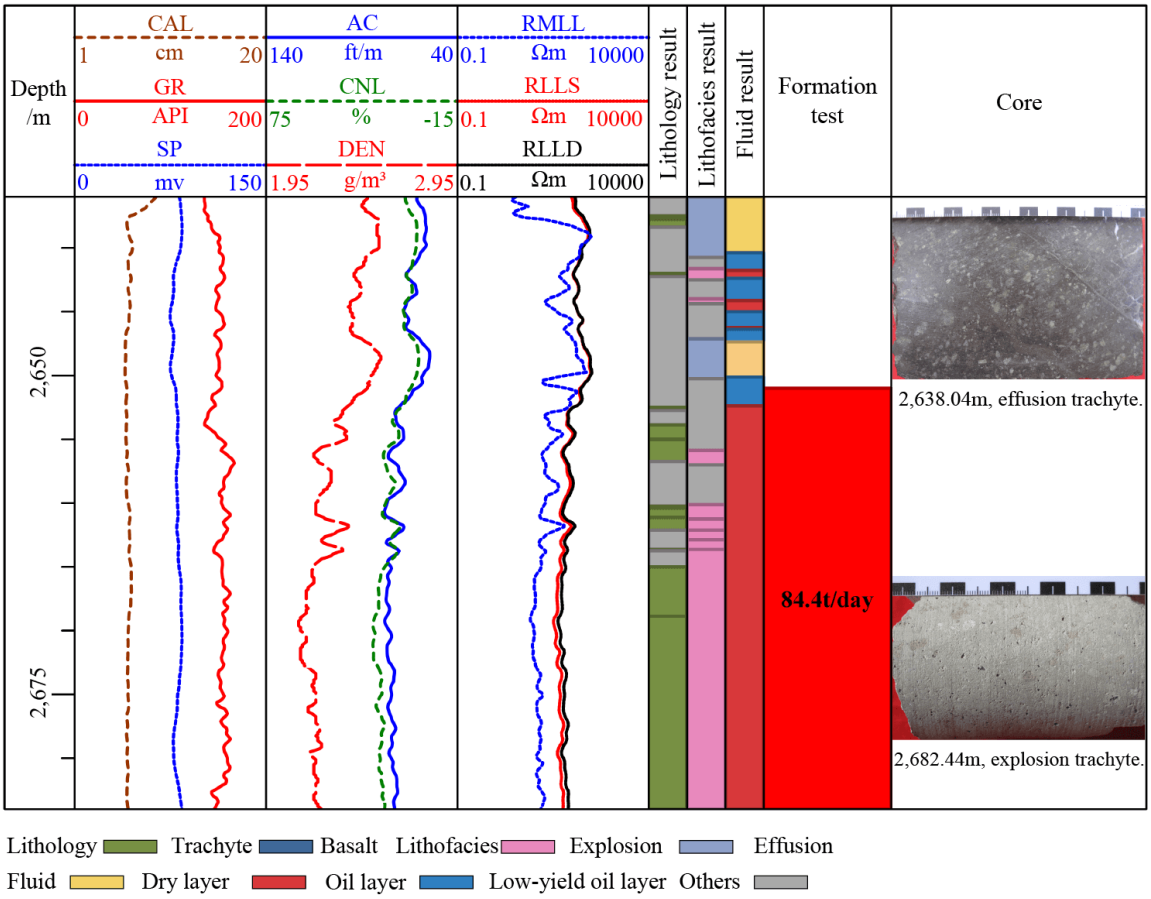

Fig. 9. Predicted results of well B for 2,636-2,684 m.

synthetic instance generation. Knowledge-Based Systems, 2015, 89: 385-397.

Chu, H., Dong, P., Lee, W. J. A deep-learning approach for reservoir evaluation for shale gas wells with complex fracture networks. Advances in Geo-Energy Research, 2023, 7(1): 49-65.

Duan, Y., Xie, J., Su, Y., et al. Application of the decision tree method to lithology identification of volcanic rocks-taking the Mesozoic in the Laizhouwan Sag as an example. Scientific Reports, 2020, 10(1): 19209.

Ehsan, M., Gu, H. An integrated approach for the identification of lithofacies and clay mineralogy through Neuro-Fuzzy, cross plot, and statistical analyses, from well log data. Journal of Earth System Science, 2020, 129(1): 101.

Falivene, O., Auchter, N. C., De Lima, R. P., et al. Lithofacies identification in cores using deep learning segmentation and the role of geoscientists: Turbidite deposits (Gulf of Mexico and North Sea). AAPG Bulletin, 2022, 106(7): 1357-1372.

Feng, Y., Bian, W., Gu, G., et al. A drilling data-constrained seismic mapping method for intermediate-mafic volcanic facies. Petroleum Exploration and Development, 2016, 43(2): 251-260.

Galar, M., Fernandez, A., Barrenechea, E., et al. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition, 2011a, 44(8): 1761-1776.

Galar, M., Fernandez, A., Barrenechea, E., et al. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems Man and Cybernetics, Part C-Applications and Reviews, 2011b, 42(4): 463-484.

Giordano, G., Cas, R. A. F. Classification of ignimbrites and their eruptions. Earth-Science Reviews, 2021, 220: 103697.

He, M., Gu, H., Wan, H. Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir. Journal of Petroleum Science and Engineering, 2020, 194: 107498.

Ho, M., Idgunji, S., Payne, J. L., et al. Hierarchical multi-label taxonomic classification of carbonate skeletal grains with deep learning. Sedimentary Geology, 2023, 443: 106298.

Huang, Y., Shan, J., Bian, W., et al. Facies classification and reservoir significance of the Cenozoic intermediate and mafic igneous rocks in Liaohe Depression, East China. Petroleum Exploration and Development, 2014, 41(6): 734-744.

Krzywinski, M., Schein, J., Birol, I., et al. Circos: An information aesthetic for comparative genomics. Genome Research, 2009, 19(9): 1639-1645.

Kuhn, S., Cracknell, M. J., Reading, A. M., et al. Identification

of intrusive lithologies in volcanic terrains in British Columbia by machine learning using random forests: The value of using a soft classifier. Geophysics, 2020, 85(6): B249-B258.

Li, P., Chen, P., Zhang, D. Cross-modal feature representation learning and label graph mining in a residual multi-attentional CNN-LSTM network for multi-label aerial scene classification. Remote Sensing, 2022, 14(10): 2424.

Liu, Z., Wu, H., Chen, R. Evaluation of volcanic reservoir heterogeneity in eastern sag of Liaohe Basin based on electrical image logs. Journal of Petroleum Science and Engineering, 2022, 211: 110115.

Liu, B., Zhao, X., Fu, X., et al. Petrophysical characteristics and log identification of lacustrine shale lithofacies: A case study of the first member of Qingshankou Formation in the Songliao Basin, northeast China. Interpretation, 2020, 8(3): SL45-SL57.

Lopez, V., Fernandez, A., Garcia, S., et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 2013, 250: 113-141.

Wu, B., Ji, X., He, M., et al. Mineral identification based on multi-label image classification. Minerals, 2022, 12(11): 1338.

Wu, X., Zhou, Z. A unified view of multi-label performance measures. Paper Presented at 34th International Conference on Machine Learning, Sydney, Australia, 6-11 August, 2017.

Xiang, M., Qin, P., Zhang, F. Research and application of logging lithology identification for igneous reservoirs based on deep learning. Journal of Applied Geophysics, 2020, 173: 103929.

Xiao, L. The fusion of data driven machine learning with mechanism models and interpretability issues. Geophysical Prospecting for Petroleum, 2022, 61(2): 205-212. (in Chinese)

Yang, L., Wu, X., Jiang, Y., et al. Multi-label learning with deep forest. Paper FAIA200274 Presented at 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 29 August-8 September, 2020.

Yue, Q., Shan, X., Zhang, X., et al. Quantitative characterization, classification, and influencing factors of the full range of pores in weathering crust volcanic reservoirs: Case study in bohai bay basin, China. Natural Resources Research, 2021, 30(2): 1347-1365.

Yue, W., Tao, G. A new method for reservoir fluid identification. Applied Geophysics, 2006, 3(2): 124-129.

Zhang, L., Chen, L., Hu, R., et al. Subsurface multiphase reactive flow in geologic CO_2 storage: Key impact factors and characterization approaches. Advances in Geo-energy Research, 2022, 6(3): 179-180.

Zhang, M., Li, Y., Liu, X., et al. Binary relevance for multi-label learning: An overview. Frontiers of Computer Science, 2018, 12(2): 191-202.

Zhang, L., Pan, B., Shan, G., et al. Method for identifying fluid property in volcanite reservoir. Oil Geophysical Prospecting, 2008, 43(6): 728-730. (in Chinese)

Zhang, M., Zhou, Z. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40(7): 2038-2048.

Zhang, M., Zhou, Z. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(8): 1819-1837.

Zheng, Z., Zhang, L., Cheng, M., et al. Lithofacies logging identification for strongly heterogeneous deep-buried reservoirs based on improved Bayesian inversion: The Lower Jurassic sandstone, Central Junggar Basin, China. Frontiers in Earth Science, 2023, 11: 1095611.

Zhou, K., Zhang, J., Ren, Y., et al. A gradient boosting decision tree algorithm combining synthetic minority oversampling technique for lithology identification. Geophysics, 2020, 85(4): WA147-WA158.