Artificial Intelligence in Geosciences 2 (2021) 76–81

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences

Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm

Linqi Zhu [a,b], Xueqing Zhou [a,b,*], Chaomo Zhang [c]

[a] Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences, Sanya, 57200, China
[b] Key Lab of Marine Georesources and Prospecting, Hainan Province, Sanya, 572000, China
[c] Key Laboratory of Exploration Technologies for Oil and Gas Resources, Ministry of Education, Yangtze University, Wuhan, 430100, China

ARTICLE INFO

Keywords:
Marine shale gas
Oversampling
Random forest
High-quality reservoir identification

ABSTRACT

The identification of high-quality marine shale gas reservoirs has always been a key task in the exploration and development stage. However, due to the serious nonlinear relationship between the logging curve response and high-quality reservoirs, the rapid identification of high-quality reservoirs has always been a problem of low accuracy. This study proposes a combination of the oversampling method and random forest algorithm to improve the identification accuracy of high-quality reservoirs based on logging data. The oversampling method is used to balance the number of samples of different types and the random forest algorithm is used to establish a high-precision and high-quality reservoir identification model. From the perspective of the prediction effect, the reservoir identification method that combines the oversampling method and the random forest algorithm has increased the accuracy of reservoir identification from the 44% seen in other machine learning algorithms to 78%, and the effect is significant. This research can improve the identifiability of high-quality marine shale gas reservoirs, guide the drilling of horizontal wells, and provide tangible help for the precise formulation of marine shale gas development plans.

1. Introduction

Energy is an eternal necessity, and fossil energy accounts for a huge proportion of total energy usage (Huang et al., 2018; Pang et al., 2021; Aldhuhoori et al., 2021). Conventional oil and gas reserves have become increasingly lower with continuous exploration and development. Unconventional reservoirs represented by shale gas have received growing attention. Countries led by the United States and China have begun to extract natural gas from shale (Zhu et al., 2021a–c; Pang et al., 2021; Aldhuhoori et al., 2021). Research on marine shale gas reservoirs has begun in earnest and the corresponding research understanding has developed rapidly.

Shale gas reservoirs are too tight, requiring natural gas extraction based on horizontal well technology (Sheng et al., 2020; Hou et al., 2021; Wilson et al., 2016). Correspondingly, the efficient and accurate identification of high-quality reservoirs has become a critical step. It is difficult to accurately evaluate the vertical changes of the entire well using only core data, so it is necessary to use logging data for reservoir identification.

At present, the most common high-quality reservoir identification methods are still biased towards qualitative analysis. For example, Ma et al. (2020) proposed the use of the T_2 geometric mean value to evaluate Longmaxi shale gas reservoirs. Wang et al. (2016) analysed and clarified special logging responses such as geochemical logging and nuclear magnetic resonance logging for high-quality shale gas reservoirs in the Niutitang Formation. Yan et al. (2014) discussed the possibility of qualitatively identifying high-quality reservoirs using log response. He et al. (2016) analysed the logging response characteristics of various shale facies and distinguished favourable facies. Although there are many qualitative studies, qualitative identification methods obviously cannot meet the needs of the shale development stage. Quantitative and rapid reservoir identification is more important. Some scholars have proposed a quantitative identification method for linear reservoirs based on maps (Liu and Sun, 2015; Li et al., 2020; Zhu et al., 2021a); however, the relationship between the logging response of most shale reservoirs and the reservoir type has nonlinear characteristics, and the chart method obviously does not meet this demand. Tahmasebi et al. (2017) used data mining and machine learning methods to identify shale dessert layers and performed purely data-driven shale dessert layer recognition. Zhang et al. (2021) used the IPSO-SVM algorithm combined with element logging

* Corresponding author. Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences, Sanya, 57200, China.
E-mail address: Zhouxq@idsse.ac.cn (X. Zhou).

https://doi.org/10.1016/j.aiig.2021.12.001
Received 22 July 2021; Received in revised form 23 November 2021; Accepted 1 December 2021
Available online 6 December 2021

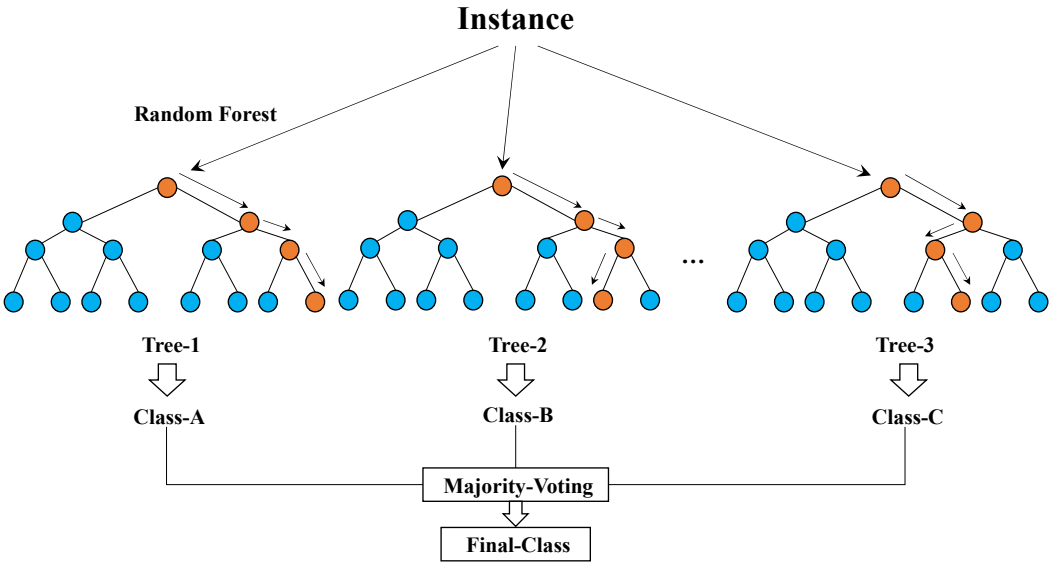

Fig. 1. Random forest prediction process.

curves to identify sweet spots in shale formations. Amosu et al. (2021) used support vector machines to effectively identify high-quality reservoirs in the Eagle Ford Shale. Vikara et al. (2020) used the machine learning-informed ensemble framework to solve the effective classification of Marcellus shale reservoirs. Although a variety of high-quality reservoir identification methods for shale have been proposed, there are still some problems in the identification of sweet beds. First, the biggest problem is that the sample set used for modelling is unbalanced due to the unavoidable biases of core sampling. This will lead to some types of reservoirs that are not easy to predict and is worse when high-quality reservoirs are not easy to identify. Second, the nonlinear approximation capabilities of different algorithms are also different, resulting in differences in prediction effects. This paper proposes a rapid identification framework for high-quality marine shale reservoirs that combines the oversampling method and random forest algorithm to quickly identify high-quality reservoirs in shale gas wells. This method significantly improves the identification effect of high-quality reservoirs and can provide help for the exploration and development of marine shale gas reservoirs.

2. Method

In this study, there are two main techniques applied to the task of reservoir identification: oversampling methods (Tao et al., 2020; Soltanzadeh and Hashemzadeh, 2021) to solve unbalanced data and random forest algorithms (Ao et al., 2020; Feng et al., 2021) to solve classification problems. Combining these can achieve the purpose of quickly identifying high-quality reservoirs and mitigate reservoir identification accuracy problems due to inexperienced formation evaluation personnel and differences in reservoir identification models.

2.1. Random forest

Random forest is an ensemble learning algorithm based on decision trees. It uses bootstrap technology to extract randomized samples from original samples to construct a single decision tree. At each node of the decision tree, a random feature subspace is used to select a sorting point. Finally, these decision trees are combined to obtain the final prediction result through a majority vote. Before we have a deeper understanding of random forests, we need to introduce decision trees first because it is the basic classifier in the random forest algorithm.

Decision trees are a widely used tree-like classifier. It starts from the root node of the tree and continuously splits by selecting the optimal column attributes to construct the tree nodes one by one until the stopping condition for tree building is reached; for example, the data in the leaf nodes are all of the same category. Compared with algorithms such as neural networks, decision trees are very interpretable. Each path from the root node to the leaf node represents a rule, and the decision tree is based on these rules for classification or prediction. When making predictions, the decision tree starts from the root node to determine the only path to the leaf node. The majority category the samples in the leaf node becomes the predicted value of the decision tree. According to the different criteria applied when selecting the attributes for classification, common decision tree generation algorithms are ID3 and C4.5, based on Shannon entropy splitting, and CART, based on Gini coefficient classification (Tewari and Dwivedi, 2019). When used for classification tasks, the C4.5 algorithm is the most commonly used algorithm. Decision tree splitting uses the information gain ratio as the splitting criterion, calculated as:

$$g_r(X,Y)=\frac{g(X,\ Y)}{H(Y)} \tag{1}$$

$$g(X,Y)=H(Y)-H(Y|X) \tag{2}$$

where g(X, Y) refers to the information gain and the degree of uncertainty of the event, which is calculated by subtracting conditional entropy from entropy. The information gain ratio adds a penalty term H(Y|X) on the basis of the information gain. The calculation formula for the penalty is:

$$H\left(Y|X)=\sum_{i=1}^{n}p_iH(Y|X=x_i\right) \tag{3}$$

Correspondingly, the calculation formula of entropy is:

$$H(p)=-\sum_{i=1}^{n}p_ilog_2p_i \tag{4}$$

The greater the entropy, the greater the amount of information contained, and the greater the uncertainty. The establishment of a decision tree essentially differentiates the data layer by layer by formulating rules so that the uncertainty of the data after classification is gradually reduced.

A decision tree is a single classifier. Although it is very interpretable, due to the limitations of its own model, there is a bottleneck in its

performance improvement, which leads to unsatisfactory training and prediction accuracy. To solve this problem, an ensemble learning algorithm is used to integrate multiple decision trees. Bagging and boosting are two categories of algorithms in ensemble learning. Bagging algorithms uniformly randomly sample from the original data so that the training of each basic classifier in the ensemble is independent of the others. The process of the random forest algorithm can be divided into three parts:

① Bootstrap sampling of samples, that is, at the beginning of constructing the tree (training the model), random sampling with replacement is performed from the training dataset D_n, sampling the same number of n data points.
② Carry out the selection of random feature subspace. Each new subdataset formed by sampling is used to construct the corresponding random forest decision tree. When splitting tree nodes, first randomly sample m features (m<D) from the most primitive D features, then select split features and split points from the feature subspace formed by these m features. The selection criterion is a decrease in the information gain ratio. The above process is repeated to construct the tree nodes one by one until the stopping conditions are met. These stopping conditions can be that all the data in the subdatasets are distinguished or the number of samples in the leaf nodes is less than a preset threshold.
③ Voting predictions. The decision tree of each basic classifier is used to make independent predictions first to obtain the discriminative category of the new data, then the results from each tree are used to output the final separate results by means of majority voting. Fig. 1 shows the prediction process of random forest.

2.2. The oversampling method for unbalanced datasets

An unbalanced dataset is a very difficult problem in actual machine learning prediction, especially when the total number of samples is insufficient or the cost of obtaining samples is too high. In the classification task, if the number of samples in different categories in the processed dataset is quite unequal, the direct use of machine learning algorithms for training will lead to poor prediction results. Such a training dataset is called unbalanced. In the unbalanced dataset, the category with a large sample size is called the majority category, and the category with a small sample size is called the minority category. In many practical scenarios, the number of minority types is small and difficult to identify, but these samples may contain critical information. If there are a large number of discriminative errors in the classification of these samples, it may cause a very serious impact, especially when the categories of these samples are the categories we care about most. Traditional machine learning algorithms will assume that the sample sizes of different categories are similar. What these algorithms care about is how to ensure the classification accuracy of the entire dataset, which will bias the model towards categories with the most samples during the classification process. Based on these factors, when dealing with the problem of imbalance, the data need to be specially processed to improve the balance of the dataset.

Oversampling alleviates the imbalance between the categories by expanding the size of the samples of the minority categories while maintaining the original scale for the samples of the majority categories. Oversampling methods include nonheuristic and heuristic algorithms. The typical nonheuristic method is random oversampling. In this way, the size of the minority types of samples is increased by simply copying some of the minority types of samples. The random oversampling method is relatively simple and convenient to apply and can help the prediction result. In this paper, the oversampling method is selected to rebalance datasets in the reservoir identification problem based on logging as much as possible.

Table 1
Relationship between reservoir types and corresponding reservoir parameters.

Reservoir type	TOC content	Porosity	Clay content
Type I	≥4	≥6	≤35
Type II	3–4	5–6	35–45
Type III	2–3	3–5	45–55
Type IV	<2	<3	≥55

Table 2
Core classification results of different reservoir types in different wells.

Reservoir type	Total	Well A	Well B	Well C
Type I	8	1	4	3
Type II	13	1	5	7
Type III	25	5	8	12
Type IV	30	2	2	26
Type I proportion	10.5%	11.10%	21.05%	6.25%

3. Data

3.1. Block

Block X in the study area, located in the southern part of the Sichuan Basin, is a marine overmature shale gas reservoir. High-quality reservoirs are located in the Lower Ordovician Longmaxi Formation-Upper Silurian Wufeng Formation, with a burial depth of approximately 3500 m or more, and belong to deep shale gas reservoirs (Zhu et al., 2019). The shale in the study area belongs to the main shale gas development horizon in China and is dominated by black shale with relatively developed nanoscale organic pores. The total organic carbon content in the study area is 0–8%, the porosity ranges from 0 to 9.44%, and the clay content ranges from 10 to 59%. The total gas content of high-quality reservoirs can be higher than 6 m^3/t. Since most of the shale gas reservoirs in the Sichuan Basin will be constructed in horizontal wells during the development stage, this makes the rapid determination of high-quality reservoirs of great significance, so this paper attempts to quickly identify high-quality reservoirs using our algorithms.

3.2. Data characteristics

We selected the logging data and core data of 3 key wells for rapid identification of high-quality reservoirs. A total of 76 cores in these 3 wells were subjected to helium porosity experiments, TOC content experiments, and X-ray diffraction experiments. The results of these experiments are usually used to divide high-quality reservoirs. We combine the high-quality reservoir classification standards in the study area given in the literature and core data to give the high-quality reservoir types corresponding to all cores, as shown in Table 1. Among them, the selected standards are TOC content, porosity, and clay mineral content.

Combined with the standards in this table, the reservoir types corresponding to the 76 cores can be determined. After this classification, the number of cores corresponding to different reservoir types is counted. The statistical results are shown in Table 2.

Two key pieces of information can be obtained from the classification results of reservoir categories in different wells. The first point is that there are certain differences in the distribution of reservoir types between different wells. For Well B, type I reservoirs account for the largest proportion of all reservoirs. This is consistent with the current understanding of the block. Well B has a higher single-well production and is also the most studied well. The difference in the proportion of Type I reservoirs between different wells is large, which also emphasizes the task of reservoir identification, especially for high-quality reservoirs that have a small number of cores and are relatively short for development plans. Recognition is a very meaningful work. The second point is that there are differences in the number of different reservoir types in any

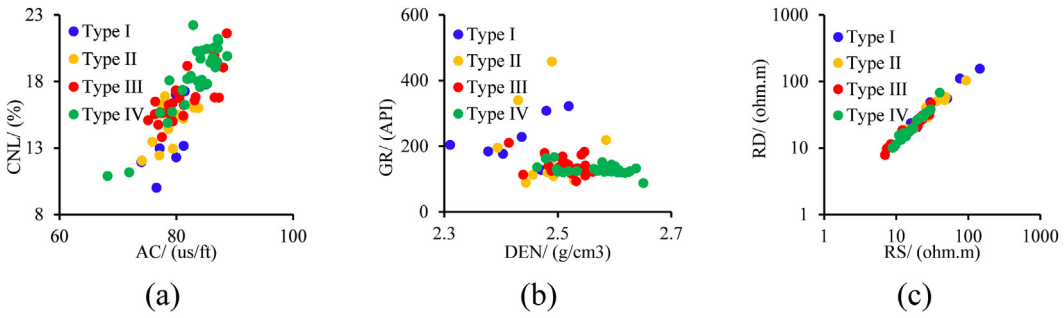

Fig. 2. The difference between different types of data after data expansion.

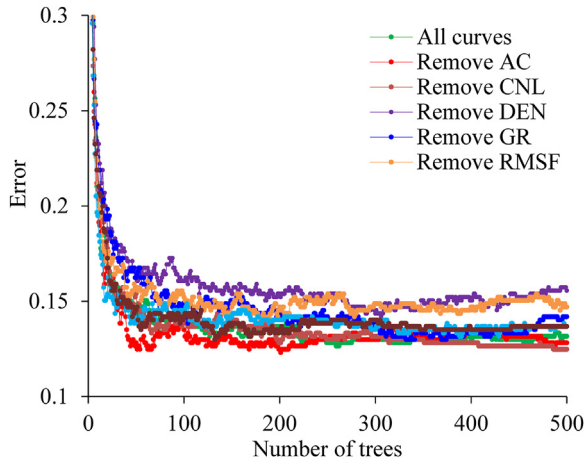

Fig. 3. Display of the results of out-of-bag errors under different modelling conditions.

well, and the number of core samples corresponding to Type I reservoirs is relatively small. This is because the thickness of the highest quality reservoirs in the study area is relatively lower than that of other types of reservoirs. They are thin, so it is very important to find Type I reservoirs and drill horizontal wells. The small number of core samples corresponding to Type I reservoirs leads to the fact that when modelling the reservoir identification process, the model tends to classify Type I reservoirs as other types of reservoirs, resulting in the inability to accurately find Type I reservoirs. The exploration and development of shale gas is very challenging.

In this study, Well A was used as the predicted well, and Wells B and C were used as for model training. Because the amount of core data for different reservoir types is unequal, the sample needs to be expanded. According to the oversampling method, the number of cores in Type I, Type II, and Type III reservoirs was expanded by randomly selected methods in the datasets from Wells B and C, and the total number of cores in the four types of reservoirs was increased to 30. The corresponding core intersection diagram for the expanded dataset is shown in Fig. 2. The different types of reservoirs clearly cannot be directly divided by linear models and the traditional method based on intersection graphs cannot obtain reliable results.

4. Result

The random forest algorithm is used to try to build a model based on the core data of Wells B and C. Well A is used as a blind well that does not participate in the modelling process to test the algorithm. This is also because the data volume of Well A is smaller than those of Wells B and C and is relatively unsuitable for modelling. The sample numbers of Wells A, B, and C are 9, 19, and 48, respectively, so the number of samples involved in modelling is 67 and in testing is 9. This is also shown in Table 2. Since the split criteria of the basic classifier-decision tree algorithm are relatively fixed, in a strict sense, the parameters that need to be determined most should be the number of random forest trees and the selected logging curve characteristics. We use the out-of-bag error as a test of the modelling effect to analyse the selection of hyperparameters (Otchere D. A., 2021). The calculation method for out-of-bag error is to use the established decision tree to predict the samples that were not drawn when each decision tree was established and count the prediction results of these samples. We used the data of Wells B and C to model and explore the best combination of input logging curves and the best number of decision trees. We used all logging curves to model, chose a log curve to remove, then and remodelled to compare the results. The corresponding out-of-bag error results are shown in Fig. 3. Out-of-bag error is actually a method of model validity testing because out-of-bag data do not participate in the establishment of each decision tree in the random forest.

It is clear from Fig. 3 that for the random forest algorithm, the AC logging curve has a negative effect on the accuracy of the model. When the AC curve is removed, the prediction accuracy of the model improves. Current research shows that sonic logging in the study area is not closely related to the TOC content of the reservoir, the porosity of the reservoir, or the clay content of the reservoir (Zhu et al., 2021a). It is suspected to be caused by factors such as pore pressure and fractures. The information provided by the CNL logging curve is not helpful for improving the accuracy of the model. Other curves have played a positive role in the prediction results, especially DEN and RMSF logging. Therefore, we choose the 5 curves of density log (DEN), gamma ray log (GR), microsphere focused resistivity logging (RMSF), RD, and RS to establish the final model. Combining the results of out-of-bag errors, we can determine that as the number of decision trees in the random forest increases, the out-of-bag errors basically decrease or remain stable. Results also show that random forest is very suitable for the task of high-quality reservoir identification and there is no obvious overfitting phenomenon. Based on

Table 3
The prediction results of different methods in Well A.

Accuracy	RF+oversampling	RF	Bayes+oversampling	Bayes	KNN+oversampling	KNN
Type 1	100%	0%	0%	0%	0%	0%
Type 2	100%	100%	0%	0%	100%	100%
Type 3	60%	60%	60%	40%	20%	20%
Type 4	100%	100%	100%	100%	100%	100%
Total	78%	66%	56%	44%	44%	44%

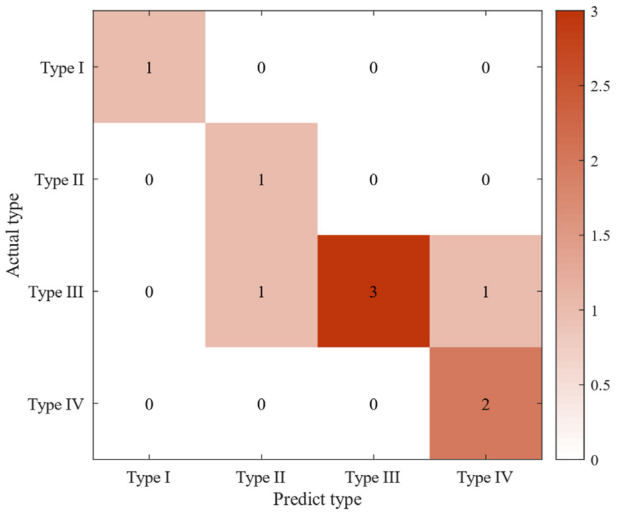

Fig. 4. Confusion matrix of all data prediction results of Well A.

these results, we chose 300 decision trees for modelling. After modelling, the Well A data were predicted (Table 3). The corresponding RF+oversampling prediction results are shown in Fig. 4.

It can be seen from Table 3 and Fig. 4 that the model obtained by using the modelling method proposed in this paper has a prediction accuracy of 78% for the core sample of Well A and an error rate of 22%. Moreover, the type of misjudgement mainly appears between Type II and Type III, and the accuracy of identifying Type I reservoirs is 100%. To facilitate the comparison, we use the original sample set and the oversampled set as the training data for the KNN algorithm, the support vector machine, and the random forest algorithm, then predict the core of Well A with each model. The results of the reservoir types predicted based on the logging data of the entire interval of Well A are shown in Fig. 5.

It can be seen from Fig. 5 that the prediction results of the whole well section of different methods are still very different. Compared with other algorithms, KNN may be less suitable for the identification of high-quality marine shale gas reservoirs. The prediction results given will jump throughout the depth section, especially for Type III and Type IV reservoirs. The problem with Bayes' prediction results is that its evaluation effect is not good for the best quality reservoir section. From the core reservoir classification results of Well A, it can be seen that high-quality reservoirs are mainly concentrated at the bottom of the entire reservoir, and this depth section is also the place where the reservoir type changes most frequently. When drilling a horizontal well, the best drilling depth is likely to come from this depth section. Using the prediction results of the Bayes algorithm to determine the drilling position of a horizontal well is less likely to be accurately determined. The random forest algorithm combined with the oversampling method and the random forest algorithm without the oversampling method have more reliable prediction

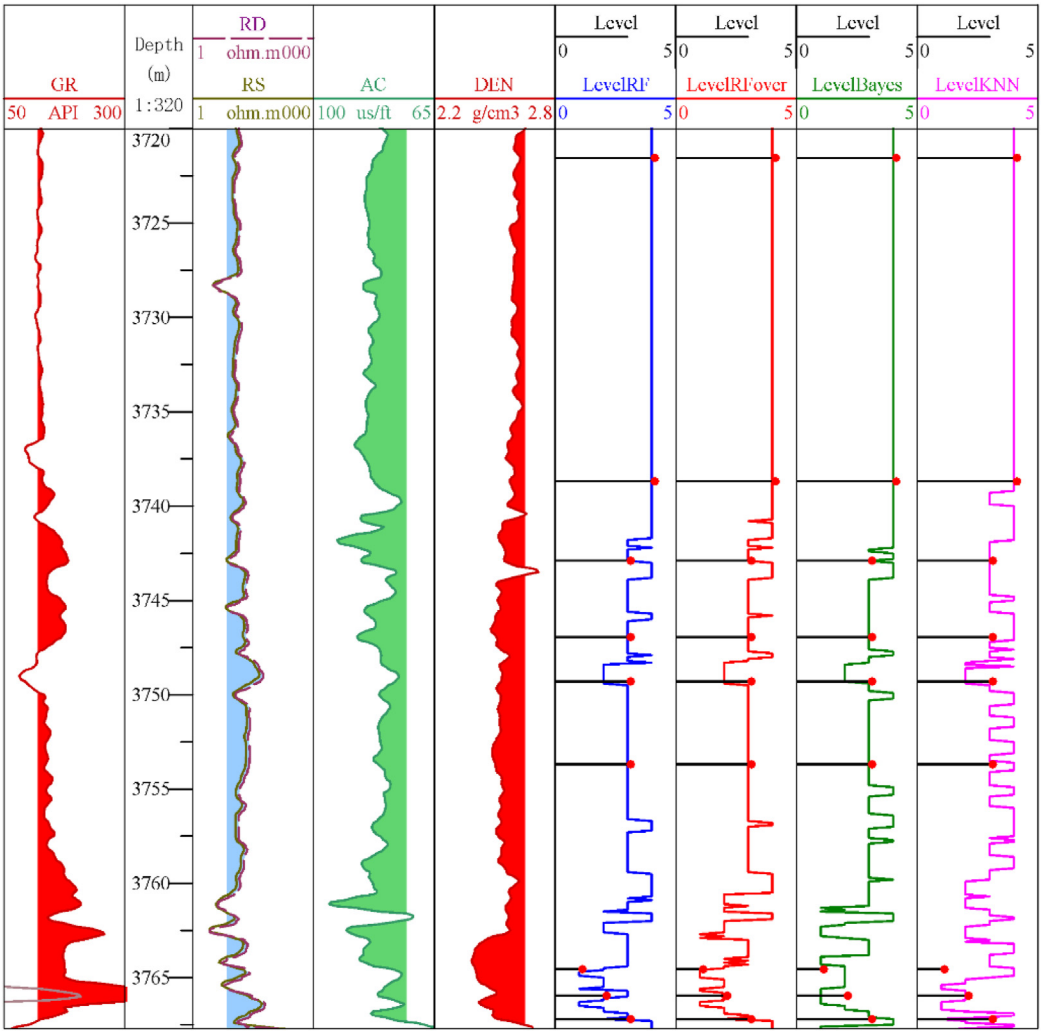

Fig. 5. Reservoir type prediction results of Well A based on logging.

results. Of the two prediction results, the random forest algorithm combined with the oversampling method is more reliable than the core result. Moreover, it also predicted the existence of Type I reservoirs at 3762.49–3763.02 m, showing higher accuracy and having higher guiding significance for the drilling of horizontal wells. Using the random forest algorithm with the oversampling method, the potential optimal horizontal well drilling horizons can be quickly determined as 3762.49–3763.02 m, 3764.5–3765.03 m, 3765.3–3765.83 m, and 3766.09–3766.62 m.

5. Conclusion

An increasing number of blocks of shale gas reservoirs have passed the exploration stage and have begun to enter the development stage. This puts forward high requirements for the rapid identification of high-quality marine shale gas reservoirs based on logging. This article conducts research on this task and proposes a rapid identification study of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. Through a series of studies in this article, the following conclusions can be drawn:

① Since the thickness of high-quality reservoirs is usually smaller than that of general reservoirs, training sample sets underrepresent these critical samples and model predictions tend to be more accurate for general reservoirs. In this paper, the oversampling method is used to make the number of training samples of each type of reservoir equal and 67 core samples were expanded to 112 samples. The increase in the number of samples reduces the bias of model building.

② Using the random forest algorithm, nearest neighbour algorithm, and support vector machine algorithm to predict high-quality reservoirs, we showed that the random forest algorithm achieved the best prediction results. In addition to the prediction accuracy of the other algorithms being lower than that of the random forest algorithm, they also have relatively poorer prediction effects for the highest quality Type I reservoirs. The random forest algorithm can achieve the best prediction results, increasing the prediction accuracy from 44% to 78%.

Finally, this research showed that the combination of the oversampling method and random forest algorithm can quickly identify high-quality marine shale gas reservoirs and thereby provide tangible help for the development of marine shale gas reservoirs.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, “Rapid identification of high-quality marine shale gas reservoirs based on oversampling method and random forest algorithm”.

Acknowledgements

This project was funded by the Laboratory for Marine Geology, Qingdao National Laboratory for Marine Science and Technology, (MGQNLM-KF202004); China Postdoctoral Science Foundation (2021M690161, 2021T140691); Postdoctoral Funded Project in Hainan Province (General Program); Chinese Academy of Sciences- Special Research Assistant Project; the Open Fund of Key Laboratory of Exploration Technologies for Oil and Gas Resources (Yangtze University), Ministry of Education (No.K2021–03, K2021-08). The authors would like to express their sincere thanks to the editor Prof. Hua Wang for his enthusiasm, patience, and tireless efforts. The authors are very grateful to the anonymous reviewers for their constructive advice on how to improve the paper.

References

Aldhuhoori, M.A., Belhaj, H., Alkuwaiti, H.K., Ghosh, B., Fernandes, R., Qaddoura, R., 2021. Role of viscous, diffusion and inertial mechanisms in modeling fluid flow through unconventional reservoir. J. Petrol. Sci. Eng. 205, 108772.

Amosu, A., Imsalem, M., Sun, Y., 2021. Effective machine learning identification of TOC-rich zones in the Eagle Ford Shale. J. Appl. Geophys. 188, 104311.

Ao, Y., Zhu, L., Guo, S., Yang, Z., 2020. Probabilistic logging lithology characterization with random forest probability estimation. Comput. Geosci. 144, 104556.

Feng, R., Grana, D., Balling, N., 2021. Imputation of missing well log data by random forest and its uncertainty analysis. Comput. Geosci. 152, 104763.

He, J., Ding, W., Jiang, Z., Li, A., Wang, R., Sun, Y., 2016. Logging identification and characteristic analysis of the lacustrine organic-rich shale lithofacies: a case study from the Es_3^L shale in the Jiyang Depression, Bohai Bay Basin, Eastern China. J. Petrol. Sci. Eng. 145, 238–255.

Hou, L., Yu, Z., Luo, X., Lin, S., Zhao, Z., Yang, Z., Wu, S., Cui, J., Zhang, L., 2021. Key geological factors controlling the estimated ultimate recovery of shale oil and gas: a case study of the Eagle Ford shale, Gulf Coast Basin, USA. Petrol. Explor. Dev. 48 (3), 762–774.

Huang, S., Wu, Y., Meng, X., Liu, L., Ji, W., 2018. Recent advances on microscopic pore characteristics of low permeability sandstone reservoirs. Adv. Geo-Energy Res. 2 (2), 122–134.

Li, J., Wang, M., Lu, S., Chen, G., Tian, W., Jiang, C., Li, Z., 2020. A new method for predicting sweet spots of shale oil using conventional well logs. Mar. Petrol. Geol. 113, 104097.

Liu, Z., Sun, Z., 2015. New brittleness indexes and their application in shale/clay gas reservoir prediction. Petrol. Explor. Dev. 42 (1), 129–137.

Ma, X., Wang, H., Zhou, S., Feng, Z., Liu, H., Guo, W., 2020. Insights into NMR response characteristics of shales and its application in shale gas reservoir evaluation. J. Nat. Gas Sci. Eng. 84, 103674.

Otchere, D.A., Ganat, T.O.A., Gholami, R., Lawal, M., 2021. A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction, 91, p. 103962.

Pang, X., Shao, X., Li, M., Hu, T., Chen, Z., Zhang, K., Jiang, F., Chen, J., Chen, D., Peng, J., Pang, B., Wang, W., 2021. Correlation and difference between conventional and unconventional reservoirs and their unified genetic classification. Gondwana Res. 97, 73–100.

Sheng, G., Su, Y., Zhao, H., Liu, J., 2020. A unified apparent porosity/permeability model of organic porous media: coupling complex pore structure and multi-migration mechanism. Adv. Geo-Energy Res. 4 (2), 115–125.

Soltanzadeh, P., Hashemzadeh, M., 2021. RCSMOTE: range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem. Inf. Sci. 542, 92–111.

Tahmasebi, P., Javadpour, F., Sahimi, M., 2017. Data mining and machine learning for identifying sweet spots in shale reservoirs. Expert Syst. Appl. 88, 435–447.

Tao, X., Li, Q., Guo, W., Ren, C., He, Q., Liu, R., Zou, J., 2020. Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering. Inf. Sci. 519, 43–73.

Tewari, S., Dwivedi, U.D., 2019. Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. Comput. Ind. Eng. 128, 937–947.

Vikara, D., Remson, D., Khanna, V., 2020. Machine learning-informed ensemble framework for evaluating shale gas production potential: case study in the Marcellus Shale. J. Nat. Gas Sci. Eng. 84, 103679.

Wang, R., Ding, W., Wang, X., Cui, Z., Wang, X., Chen, E., Sun, Y., Xiao, Z., 2016. Logging identification for the lower Cambrian Niutitang shale reservoir in the Upper Yangtze region, China: a case study of the Cengong block, Guizhou Province. J. Nat. Gas Geosci. 1 (3), 231–241.

Wilson, M.J., Wilson, L., Shaldybin, M.V., 2016. Clay mineralogy and unconventional hydrocarbon shale reservoirs in the USA. II. Implications of predominantly illitic clays on the physico-chemical properties of shales. Earth Sci. Rev. 158, 1–8.

Yan, W., Wang, J., Liu, S., Wang, K., Zhou, Y., 2014. Logging identification of the Longmaxi mud shale reservoir in the Jiaoshiba area, Sichuan Basin. Nat. Gas. Ind. B 1, 230–236.

Zhu, L., Ma, Y., Cai, J., Zhang, C., Wu, S., Zhou, X., 2021a. Key factors of marine shale conductivity in southern China—Part II: the influence of pore system and the development direction of shale gas saturation models. J. Petrol. Sci. Eng. 109516.

Zhu, L., Ma, Y., Zhang, C., Wu, S., Zhou, X., 2021b. New parameters for charactering the gas-bearing properties of shale gas. J. Petrol. Sci. Eng. 201, 108290.

Zhu, L., Ma, Y., Cai, J., Zhang, C., Wu, S., Zhou, X., 2021c. Key factors of marine shale conductivity in southern China—Part I: the influence factors other than porosity. J. Petrol. Sci. Eng. 205, 108698.

Zhu, L., Zhang, C., Zhang, Z., Zhou, X., Liu, W., 2019. An improved method for evaluating the TOC content of a shale formation using the dual-difference ΔlogR method. Mar. Petrol. Geol. 102, 800–816.