Journal of Oil and Gas Technology 石油天然气学报, 2026, 48(1), 1-10
Published Online March 2026 in Hans. https://www.hanspub.org/journal/jogt
https://doi.org/10.12677/jogt.2026.481001

Hans汉斯

基于随机森林的火山岩岩性识别方法

熊　平[1*]，王　磊[2,3]，胡　松[1,2]，刘继龙[1,2]

[1]中国石化石油勘探开发研究院，北京
[2]中国石化测录井重点实验室，北京
[3]中石化经纬有限公司华北测控公司，河南 郑州

收稿日期：2025年12月10日；录用日期：2026年1月22日；发布日期：2026年2月5日

摘　要

随机森林模型是在决策树模型基础上发展而来的一种模式识别技术。与其他方法相比，该模型不仅能解决决策树模型分类过程中出现的过拟合现象，提升模型的泛化能力，还具备更高的分类精度。岩性识别本质上属于分类问题，本文利用随机森林泛化能力强以及分类准确性高的优势进行火山岩岩性识别，首先通过薄片分析、井壁取心、钻井取心等资料确定了X断陷火石岭二段133个岩性样本点以及岩性识别标准，以GR、CNL、DEN、AC、RLLD、RLLS测井曲线为样本属性，然后通过随机森林模型中的最大特征数、最大深度、决策树分类器的个数变化对岩性识别精度影响的数值模拟，建立了基于最大特征数、最大深度、决策树分类器个数的最优随机森林岩性识别模型，最后对15口重点井岩性进行测井识别，可识别安山岩、安山质(沉)火山角砾岩、安山质(沉)凝灰岩、流纹岩、流纹质沉火山角砾岩、流纹质(沉)凝灰岩6种岩性。将识别结果与FMI以及取心岩性进行对比验证，平均符合率为80.2%，结果表明：利用该方法识别岩性比录井确定的岩性更准确，满足X断陷岩性识别的要求。

关键词

火山岩，岩性识别，数值模拟，随机森林模型，泛化能力，FMI

Lithology Identification Method of Volcanic Rocks Based on Random Forest

Ping Xiong[1*], Lei Wang[2,3], Song Hu[1,2], Jilong Liu[1,2]

[1]Petroleum Exploration and Production Research Institute, SINOPEC, Beijing
[2]Well Logging Key Laboratory, SINOPEC, Beijing
[3]Huabei Geosteering & Logging Company Sinopec Matrix Corporation, Zhengzhou Henan

Received: December 10, 2025; accepted: January 22, 2026; published: February 5, 2026

[*]通讯作者。

文章引用: 熊平, 王磊, 胡松, 刘继龙. 基于随机森林的火山岩岩性识别方法[J]. 石油天然气学报, 2026, 48(1): 1-10.
DOI: 10.12677/jogt.2026.481001

Abstract

Random forests model is a pattern recognition technology developed on the basis of decision tree model. Compared with other methods, it can not only solve the over-fitting phenomenon of decision tree model classification, improving the generalization of the model, but also have higher classification accuracy. Lithology identification is essentially a classification problem. Therefore, the advantages of strong generalization of random forest and high classification accuracy are used in this paper. Firstly, 133 lithology samples and lithology identification standard of the second member of Huoshiling formation in X fault depression are determined by thin section analysis, borehole coring and drilling coring data. GR, CNL, DEN, AC, RLLD and RLLS are used to identify lithology as the sample attribute. Then, by simulating the influence of the maximum characteristic number, the maximum depth and the number of decision tree classifiers on lithology recognition accuracy in random forest model, a lithology recognition model based on random forest is established. Finally, the lithology of 15 key wells is identified. Six kinds of lithology including andesite, andesitic volcanic breccia, andesitic tuff, rhyolite, rhyolitic volcanic breccia and rhyolitic tuff are recognised. Compared with FMI and core lithology, the average coincidence rate is 80.2%. The results show that the lithology identification method is more accurate than that determined by mud logging and meets the requirements of lithology identification in X fault depression.

Keywords

Volcanic Rock, Lithology Recognition, Numerical Simulation, Random Forests Model, Generalization Ability, FMI

1. 引言

随着石油勘探日益发展，火山岩油气藏逐渐成为勘探开发的主要目标。X 地区火山岩储层油气资源丰富，是油气增储上产的重要阵地，然而该区火山岩岩性复杂，各岩性的测井响应特征交叠，给火山岩岩性识别带来巨大挑战。准确且有效识别火山岩岩性对于提高该区产能具有重大意义。

岩性识别方法主要有图版法、识别方程判别法以及基于机器学习的模式识别方法。其中，图版法具有操作简单，易于实现的优点，主要利用测井解释人员的主观思想人为地确定岩性分类界限进行岩性判断，其识别的准确与否受人为因素影响[1]-[4]；识别方程判别法则是通过测井参数建立岩性识别方程来对岩性进行自动识别，计算的结果较图版法更精确，但是可调节的参数过多，从而造成实际的应用效果较差[5] [6]。模式识别方法很多，其中基于样本中心距离概念的聚类分析方法虽可以进行岩性识别，但不同样本数量的聚类中心对岩性识别精度影响较大[7] [8]；神经网络模型内部属于“黑箱”，可视化程度弱，网络的结构难以确定[9] [10]；SVM 模型能够通过内积运算将低维空间线性不可分的样本转换为高维空间可分，但其训练速度慢，核函数的选择缺乏依据，这在一定程度上制约着 SVM 模型的应用[11] [12]；决策树模型虽然能够提高模型的可视化程度，但其对数据分布特征的表征不够准确，同时存在严重的过拟合现象，往往对特征较多的样本预测效果不够好[7] [13]。随机森林模型是在决策树模型的基础上发展起来的，它能够解决决策树模型的过拟合现象，而且该模型具有一定的泛化能力，因此它得到了学术界的广泛认可，经过长时间的发展，随机森林模型已经被用作岩性识别，但是很少有学者对随机森林模型的

关键参数进行调节，因此，生成的随机森林模型往往不是稳定的、最优的分类模型[14] [15]。

本文通过调节随机森林中节点分裂准则，决策树最大深度、最大特征数和决策树分类器的个数，建立了适用于 X 断陷复杂火山岩岩性识别的稳定的、最优的随机森林模型，最终通过投票的方式确定测试样本的分类结果。有效地对火山岩岩性进行了自动识别。

2. 随机森林基本原理

随机森林模型分类准确率高、模型泛化能力强[15]。随机森林模型首先通过有放回抽样的方法对训练集进行抽样，假设待抽取的训练集个数为 D，待生成的决策树分类模型个数为 K，则生成 K 个相互独立决策树模型的随机向量，即用 D 个样本去训练得到相互独立的 K 棵决策树分类模型。在决策树的生成过程中，基于训练集抽取的样本子集中的一组随机特征空间，利用信息增益率作为节点分类准则来选取最优属性，生成 K 棵相互独立的决策树分类模型；在决策树的分类过程中，将测试集输入到随机森林模型对 K 棵树的分类结果进行投票，则确定最终的分类结果[16]。随机森林生成过程如图 1 所示，随机森林模型的分类过程如图 2 所示。

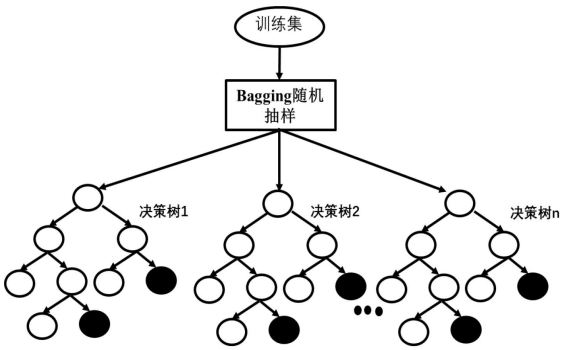

Figure 1. Schematic diagram of random forest training process
图 1. 随机森林训练过程示意图

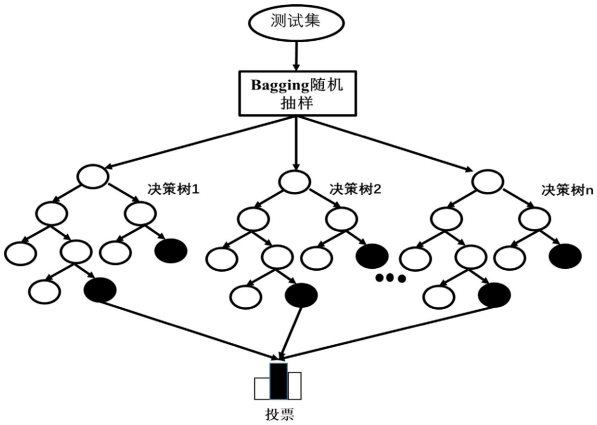

Figure 2. Schematic diagram of random forest classification process
图 2. 随机森林分类过程示意图

2.1. 随机森林随机化过程

随机森林模型是由多个相互独立的决策树集成得到的，模型通过样本以及特征属性的随机抽样得到，从整体上提高了单棵决策树模型的准确率和泛化能力。

(1) 样本随机化

样本随机化是指每棵决策树模型输入的样本都是采用有放回的抽样的方法从原始样本空间抽取样本，这个随机过程使得每个决策树模型都具有唯一的一个分类结果，最终各个决策树模型通过投票的方式确定最终的分类结果。

(2) 特征随机化

特征随机化是指利用样本属性阈值来完成模型的训练时，创造一个随机属性集合，在此集合范围内利用节点分裂准则从样本属性集合中有放回地随机抽取最优属性阈值，现假设属性集合有 M 个子特征空间，随机森林选取的最优子属性阈值为 F 个。F 的大小直接决定了随机森林模型各棵决策树分类的准确性以及各决策树模型之间的相关程度[15]。F 值越大，随机森林的分类准确度越大，各棵决策树模型的相关性越强，模型的泛化能力越弱，因此合理选择 F 值意义重大。

2.2. 节点分裂准则

在随机森林模型中，每棵决策树模型都通过不断调节样本属性阈值来完成模型的训练过程，属性以及阈值的选取就叫做节点分裂准则。随机森林模型通过对随机抽取的 F 个最优属性进行选择后，对样本相似性进行度量，选择度量结果最优的属性及阈值作为分裂节点。经典的决策树模型包括 ID3 算法、C4.5 算法以及 CART 算法[17]。

(1) ID3 算法

ID3 算法以最大信息增益为节点分裂准则，最大信息增益考虑到节点分裂前后信息熵值的差异。下面详述一下最大信息增益的计算过程。在一个带有标签的样本数为 D 的决策树训练集中，包含 m 种类别，其类别是第 i 类的概率为 P_i，$i=1$，2，3…m，训练集中存在 n 种属性，对于任意一种属性，存在 k 种不同的备选值，训练样本 D 的信息增熵为：

$$Info(D) = -\sum_{i=1}^{m} P_i \log(P_i) \tag{1}$$

而按照节点分裂准则将训练集进行划分，得到第 j 种属性作为划分依据，得到划分后样本子集的信息熵公式为：

$$Info_j(D) = \sum_{i=1}^{k} \frac{D_{j,i}}{D} Info(D_{j,i}) \tag{2}$$

因此，属性 j 的信息增益为：

$$Gain(A) = Info(D) - Info_j(D) \tag{3}$$

(2) C4.5 算法

由于 ID3 算法受属性特征以及属性备选值个数的影响，该算法通常会选择属性备选阈值多的特征，而且 ID3 算法在处理连续属性数据时会出现明显的不足，因此 C4.5 算法得以应用，C4.5 算法以信息增益率最大为节点分裂准则，消除了备选属性阈值数量对信息不确定度的影响，对于第 j 种属性 A 的信息熵表达式为：

$$SplitInfo(A) = SplitInfo(D_j) = -\sum_{i=1}^{k} \frac{|D_{j,i}|}{|D|} \times \log_2\left(\frac{|D_{j,i}|}{|D|}\right) \tag{4}$$

因此，训练集 D 被划分到第 j 种属性的 k 个阈值后的信息增益率为：

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{5}$$

2.3. 基于随机森林的岩性二分类问题的实现

常见的分类问题一般都是二分类问题，关于二分类问题的随机森林构建过程，根据前人的描述，确定为以下步骤[17]：

(1) 设定决策树模型的数量 T，节点分裂属性数目 s 小于 S (其中 S 为样本属性总数目)；

(2) 基于节点分裂准则确定第 t 棵决策树模型的根节点，将样本输入到模型中；

(3) 如果样本中只含一类，那么返回标签等于相应种类的单叶节点决策树，如果节点样本空间为空，则标签等于种类数量最多的标签；否则继续执行步骤(4)；

(4) 随机选择 S 个特征作为子特征空间以该特征空间结合抽样子集形成子样本空间，根据节点分裂准则选取最优的属性阈值作为分裂节点；

(5) 根据分裂属性，在内部节点中建立左子树和右子树；

(6) 在左子树和右子树中，执行步骤(3)的判断作为叶节点的生成条件，若条件不成立，则执行步骤(5)；

(7) 当 t < T 时，执行步骤(2)，否则结束。

2.4. 基于随机森林的岩性多分类问题的实现

常见的分类问题都是在二分类问题的基础上进行的，利用随机森林模型同样能够实现样本的二分类，但事实上 X 断陷需要识别的火山岩类型多样，需要对输入数据进行分步处理才能实现火山岩岩性的划分，因此，识别的步骤主要分为三步，第一，识别火山岩成分特征；第二，识别火山熔岩类型；第三，识别火山角砾和凝灰结构特征；具体流程图如图 3 所示。

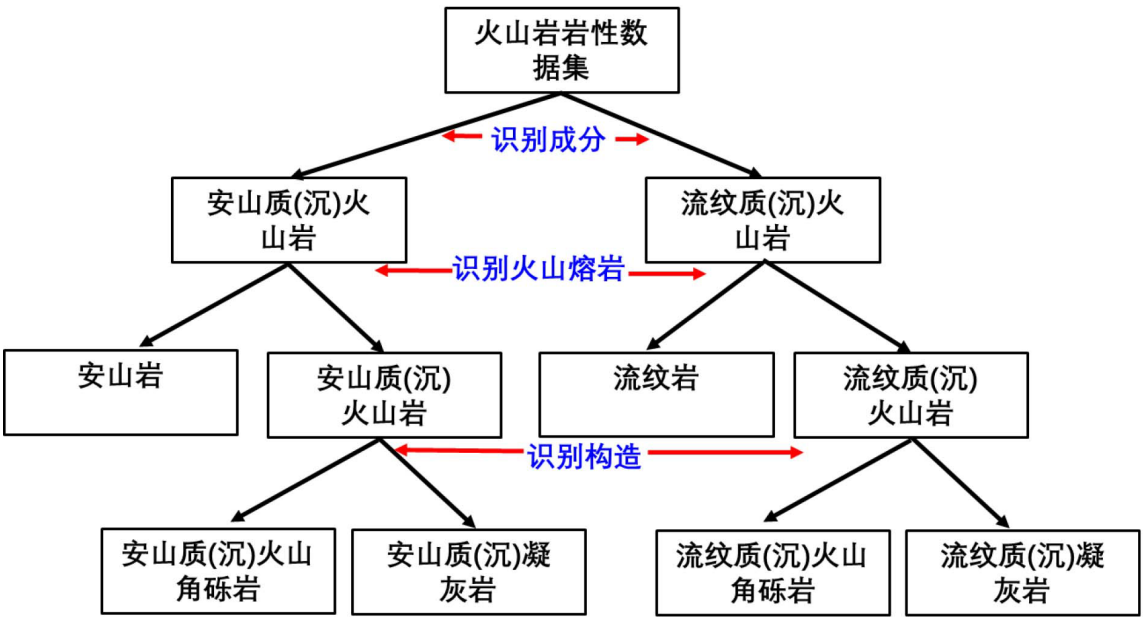

Figure 3. Schematic diagram of volcanic rock lithology identification based on the random forest algorithm
图 3. 基于随机森林算法的火山岩岩性识别流程图

3. 实际应用

3.1. 火山岩测井响应特征

本文研究区位于 X 地区，基于 13 口井 133 层的火山岩岩性数据，其中安山质(沉)火山岩 57 层，流

纹质(沉)火山岩 76 层；根据井壁取心，薄片鉴定，确定了火山岩岩性按成分主要分为安山质(沉)火山岩以及流纹质(沉)火山岩两大类，按照结构，进一步划分为安山岩，安山质(沉)火山角砾岩，安山质(沉)凝灰岩，流纹岩，流纹质沉火山角砾岩，流纹质(沉)凝灰岩六小类，各岩性的测井响应特征范围详见表 1。

Table 1. Logging response characteristics of volcanic rocks
表 1. 火山岩测井响应特征数据

岩性	GR (API)	RLLD (Ω·m)	CNL (%)	DEN (g/cm^3)	AC (μs/m)
安山岩	42~82	13~227	10~29	2.43~2.62	203~247
安山质火山角砾岩	51~92	8~142	3~32	2.36~2.65	197~265
安山质凝灰岩	45~102	13~133	5~25	2.37~2.7	203~251
流纹岩	109~199	11~305	1~31	2.34~2.64	195~273
流纹质火山角砾岩	96~160	10~299	1~22	2.39~2.62	207~271
流纹质凝灰岩	7~173	19~100	5~24	2.32~2.59	211~268

3.2. 基于随机森林的火山岩岩性识别

基于研究区 13 口井 133 层的岩性样本，选取 GR、AC、CNL、DEN、RLLD、RLLS 作为样本属性，以 bagging 随机抽样形式，用样本数量的 80%作为训练集，以样本数量的 20%作为测试集，对样本数据进行随机化处理，采用分级划分岩性的思想，建立了逐级划分岩性的火山岩岩性识别随机森林模型，需要调节的参数为：节点分裂准则，决策树的最大深度，最大特征数以及决策树分类器的个数。

(1) 节点分裂准则

由于岩性识别受到流体性质等多种因素的影响，选用单一属性信息会对分类结果带来偏差，考虑到信息增益率会选择预测样本中的更多属性，因此，选择信息增益率作为节点分裂准则。

(2) 决策树最大深度

随机森模型中单棵决策树的最大深度表示决策树分类器节点深度的最大值，分别模拟了不同决策树最大深度的 10 次测试的精度，具体结果见表 2。从表中可以看到：最大深度为 8 时，随机森林模型识别火山岩成分特征的准确率为 90.4%，识别效果较好。以此类推，分别确定了识别安山岩、安山质(沉)火山角砾岩与安山质(沉)凝灰岩的最大深度分别为 9、6，识别流纹岩、流纹质沉火山角砾岩、流纹质(沉)凝灰岩的决策树最大深度为 7、5。具体参数详见表 3。

Table 2. Maximum depth characteristics analysis for volcanic rock lithology identification
表 2. 火山岩岩性识别最大深度特征分析

决策树最大深度	测试集准确率	训练集准确率
1	90.7%	90.6%
2	87.8%	90.6%
3	93.3%	94.2%
4	88.5%	99.5%

续表

5	88.1%	100%
6	90%	100%
7	87.8%	100%
8	90.4%	100%
9	89.6%	100%
100	88.5%	100%

Table 3. Parameters of random forest model for volcanic rock lithology identification
表 3. 火山岩岩性识别随机森林模型参数

火山岩分类识别		属性			识别率	
		最大特征数	最大深度	决策树分类器个数	训练集平均识别率	测试集平均识别率
安山质火山岩	流纹质火山岩	6	8	50	100%	92.3%
安山岩	安山质火山碎屑岩	3	9	20	100%	87.5%
流纹岩	流纹质火山碎屑岩	3	7	20	100%	70.1%
安山质凝灰岩	安山质火山角砾岩	3	6	30	100%	75.0%
流纹质凝灰岩	流纹质火山角砾岩	6	5	10	89.2%	63.8%

(3) 最大特征数

随机森林模型最大特征数的选取一般有 3 种方式：原始最大特征数，原始最大特征数开根号，以 2 为底原始最大特征数的对数，三种计算结果分别为 6、2、3，图 4 分别模拟了最大特征数是 2、3、6 情况下的岩性识别的精度，从图中可以看到：火山岩成分识别的最大特征数为 6 时，火山岩精度最高，并且火山岩识别的稳定性最好，因此确定了火山岩岩性成分识别的最大特征数为 6，以此类推，识别安山岩、安山质(沉)火山角砾岩与安山质(沉)凝灰岩的最大特征数分别为 3、3；识别流纹岩、流纹质(沉)凝灰岩与流纹质沉火山角砾岩的最大特征数为 3、6，具体信息详见表 3。

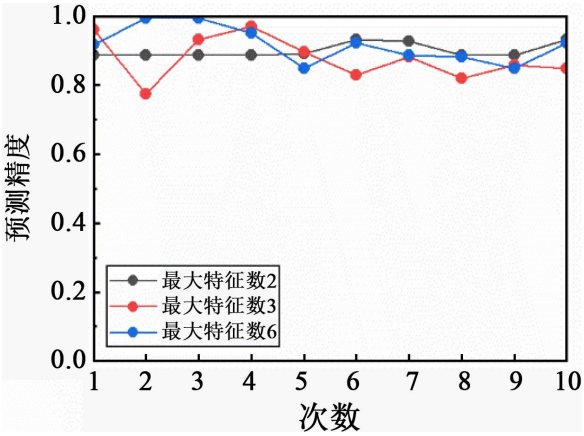

Figure 4. Schematic diagram of the variation in simulation prediction accuracy of the maximum characteristic number of volcanic rocks
图 4. 火山岩最大特征数模拟预测精度变化示意图

(4) 决策树分类器的个数

决策树分类器的个数是随机森林模型的又一参数，但它对于随机森林的表现往往是微弱的提升，而且，分别模拟火山岩随机森林模型中决策树分类器的个数影响预测精度的变化(图 5)，从图中可以看出：当火山岩随机森林模型的决策树分类器个数为 50 时，模型预测的精度最高，且预测结果较为平稳。依次类推，确定了安山岩、安山质(沉)火山角砾岩与安山质(沉)凝灰岩、流纹岩、流纹质沉火山角砾岩与流纹质(沉)凝灰岩随机森林模型的决策树分类器个数分别为 20、30、20、10。

针对火山岩岩性复杂的特点，调节了火山岩随机森林模型的节点分裂准则，最大特征数，决策树的最大深度以及决策树分类器的个数等参数，确定了适用于该地区的火山岩识别岩性的最优的随机森林模型参数，并给出了随机森林模型的岩性识别准确率(表 3)，从表 3 中可以看出：对于安山质(沉)火山岩，随机森林模型的识别精度略高，测试集识别率为 87.5%、75.0%，而流纹质火山岩识别精度略低，测试集识别率平均为 70.1%、63.8%。

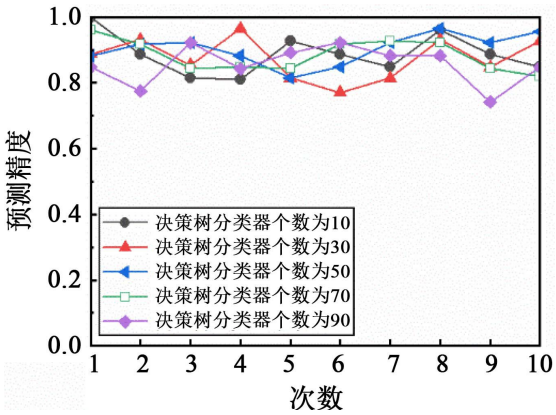

Figure 5. Schematic diagram of the variation in simulation prediction accuracy of the number of decision tree classifiers for volcanic rocks
图 5. 火山岩决策树分类器个数模拟预测精度变化示意图

3.3. 随机森林模型岩性识别准确性分析

火山岩随机森林模型的调节参数主要为节点分裂准则，最大特征数，决策树最大深度以及决策树分类器的个数。

由于岩性识别受流体性质等多种因素的影响，选用单一的属性信息会对分类结果造成偏差，考虑到信息增益率会选择预测样本中的更多属性，因此，选择信息增益率作为节点分裂准则。

最大特征数对岩性识别影响较大，最大特征数太小，随机抽取的特征空间就会存在一定的相关性，进而使得随机森林模型中的每棵决策树存在一定的相关性，降低了模型的泛化能力，最大特征数太大，往往形成的单棵决策树结构复杂，模型的泛化能力较弱。

决策树分类器的个数对火山岩岩性分类是个弱提升，因此，通过枚举法来确定决策树分类器的个数得到结果更可靠。

4. 效果分析

将进行了岩性深度归位的钻井取心、井壁取心岩性信息结合岩石薄片鉴定结果，确定火山岩岩性的名称，保证了岩性定名的可靠性，选取的岩性数据都来自井眼稳定的层段，避免了扩径的影响，建立了岩性与测井曲线间可靠的映射关系。

基于随机森林模型，对全区 15 口重点井进行了处理。其中：图 6 给出了 XX 井随机森林模型岩性识别实例，从图中可以看出：火山岩成分为流纹质(沉)火山岩，XX15~XX16.5m 井段，随机森林模型解释岩性为流纹岩，FMI 上显示为流纹岩特征的流动构造，因此将其校正为流纹岩。XX16.5~XX25m 井段，随机森林模型解释岩性为流纹质(沉)凝灰岩，FMI 上显示为凝灰岩特征的块状构造，因此将其校正为流纹质(沉)凝灰岩。

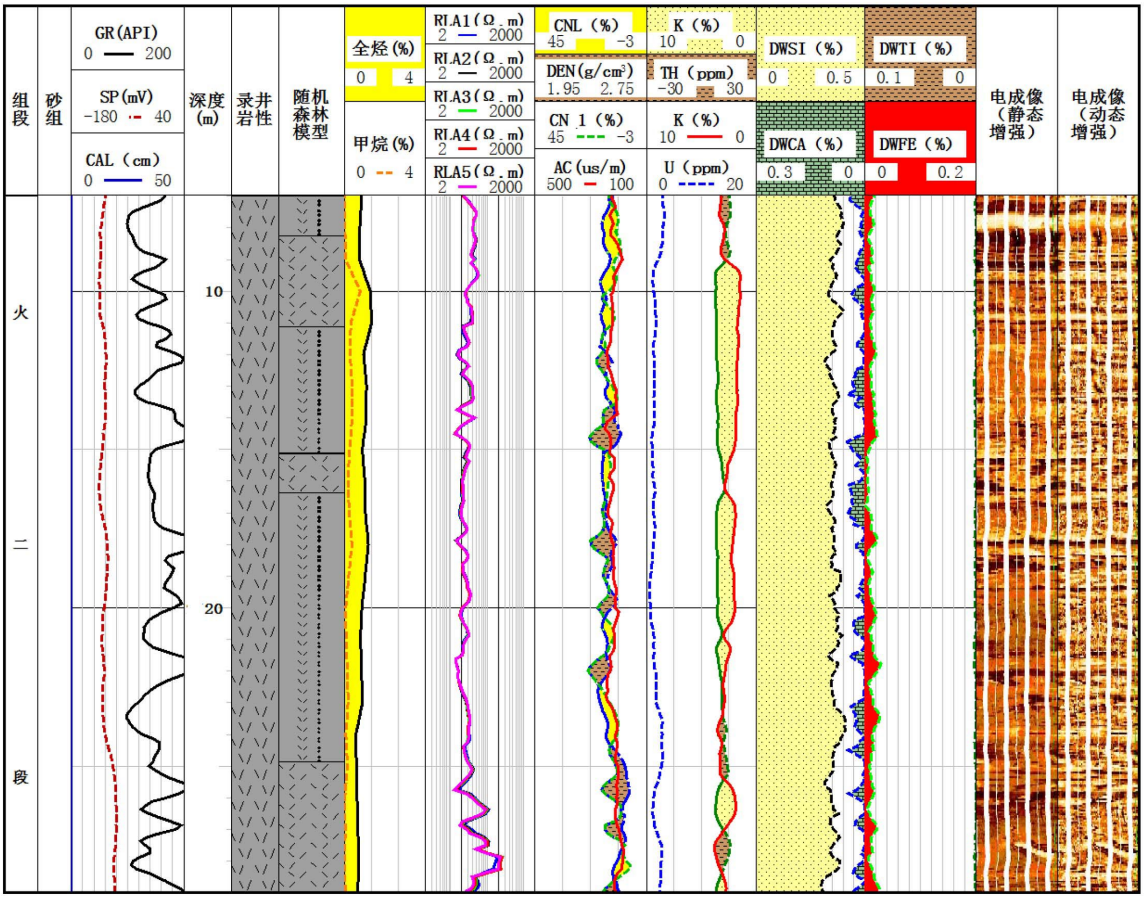

Figure 6. Case analysis of volcanic rock lithology identification
图 6. 火山岩岩性识别实例分析

5. 结论

(1) 通过系统模拟最大深度、最大特征数及决策树数量等参数对随机森林模型识别精度的影响，优化并确定了模型关键参数，结合适当的节点分裂准则，构建了兼具较高准确性与泛化能力的最优随机森林模型。

(2) 基于分级识别思想，首先，利用随机森林模型识别火山岩成分特征，进而识别其结构特征，简化了识别模型流程。对研究区 15 口重点井进行岩性识别，可以准确地识别出安山岩、安山质(沉)火山角砾岩、安山质(沉)凝灰岩、流纹岩、流纹质沉火山角砾岩、流纹质(沉)凝灰岩 6 种岩性。识别结果平均准确率为 80.2%。

(3) 本文设计的模型建立在 X 断陷特定地质背景下的数据基础上，若推广至其他地质条件差异显著的地区，可借鉴本文方法重新训练和验证，以避免因地质背景变化而引起的识别性能下降。后续研究可

扩充样本规模、融合多源数据，以进一步提升模型泛化能力与稳健性。

参考文献

[1] 赵建, 高福红. 测井资料交会图法在火山岩岩性识别中的应用[J]. 世界地质, 2003, 22(2): 136-140.

[2] 张大权, 邹妞妞, 姜杨, 等. 火山岩岩性测井识别方法研究——以准噶尔盆地火山岩为例[J]. 岩性油气藏, 2015, 27(1): 108-114.

[3] 张丽华, 张国斌, 齐艳萍, 等. 准噶尔盆地西泉地区石炭系火山岩岩性测井识别[J]. 新疆石油地质, 2017, 38(4): 427-431.

[4] 张永禄, 德勒恰提·加娜塔依, 张明玉, 等. 西泉 C 井区石炭系火山岩储层岩性测井识别[J]. 西部探矿工程, 2020, 32(12): 143-146.

[5] 刘西雷, 王玉环, 王可君, 等. 基于多级 Bayes 判别的砂砾岩体岩性识别方法[J]. 价值工程, 2017, 36(30): 162-164.

[6] 韩玉娇, 袁超, 范宜仁, 等. 基于经验模态分解和能量熵判别的火成岩岩性识别方法——以春风油田石炭系火成岩储层为例[J]. 石油与天然气地质, 2018, 39(4): 759-765.

[7] 邢贝贝, 马世忠. 三塘湖盆地火山岩岩性测井识别方法[J]. 科学技术与工程, 2011, 11(33): 8292-8294.

[8] 宋延杰, 王团, 付健, 等. 雷 64 区块砂砾岩储层岩性识别方法研究[J]. 哈尔滨商业大学学报(自然科学版), 2015, 31(1): 73-79.

[9] 张莹, 潘保芝. 基于主成分分析的 SOM 神经网络在火山岩岩性识别中的应用[J]. 测井技术, 2009, 33(6): 550-554.

[10] 范宜仁, 朱雪娟, 邓少贵, 司兆伟. 南堡 5 号构造火山岩岩性识别技术研究[J]. 地球物理学进展, 2012, 27(4): 1440-1447.

[11] 国景星, 彭雪还, 李飞. 交会图和 BP 神经网络技术在碎屑岩识别中的应用[J]. 甘肃科学学报, 2016, 28(6): 13-17.

[12] 牟丹, 王祝文, 黄玉龙, 许石, 周大鹏. 基于最小二乘支持向量机测井识别火山岩类型: 以辽河盆地中基性火山岩为例[J]. 吉林大学学报(地球科学版), 2015, 45(2): 639-648.

[13] 张昭杰, 方石. 基于遗传算法优化的支持向量机在岩性识别中的应用[J]. 世界地质, 2019, 28(2): 1-6.

[14] 刘继龙, 宋延杰, 孙红, 等. X 断陷火二段火山岩储层岩性识别技术研究[J]. 天然气与石油, 2019, 37(6): 81-86.

[15] 周雪晴, 张占松, 张超模, 等. 基于粗糙集—随机森林算法的复杂岩性识别[J]. 大庆石油地质与开发, 2017, 36(6): 027-133.

[16] 康乾坤, 路来君. 随机森林算法在测井岩性分类中的应用[J]. 世界地质, 2020, 39(2): 398-405.

[17] 苏瑞. 基于随机森林的 F 区块水淹层测井定性识别[D]: [硕士学位论文]. 大庆: 东北石油大学, 2019.