引用格式：袁婧, 贾鹿, 许国剑, 等. 基于集成算法的玛湖凹陷油气储层价值预测模型[J]. 油气藏评价与开发, 2025, 15(5): 796-806.
YUAN Jing, JIA Lu, XU Guojian, et al. Ensemble learning-based prediction model for oil and gas reservoir value in Mahu Sag[J]. Petroleum Reservoir Evaluation and Development, 2025, 15(5): 796-806.
DOI: 10.13809/j.cnki.cn32-1825/te.2025.05.009

基于集成算法的玛湖凹陷油气储层价值预测模型

袁 婧[1], 贾 鹿[1], 许国剑[2], 艾 民[1], 李嗣旭[1]
（1.中国石油新疆油田公司数智技术公司，新疆维吾尔自治区 克拉玛依 834000；2.中国石油新疆油田公司玛湖勘探开发项目部，新疆维吾尔自治区 克拉玛依 834000）

摘要：位于新疆准噶尔盆地西北部的玛湖油田，是全球最大的砾岩油田之一，其储量规模已达10亿吨级。然而，油田储层物性较差、非均质性强，给油气资源的高效开发带来了巨大挑战。高效开发油气资源的关键，在于精确识别有工业生产价值的储层，即那些油气产能较高且开发成本较低的区域。针对准噶尔盆地玛湖凹陷油气储层评价的复杂性，研究提出了一种基于集成算法的油气储层价值预测模型——OGRV（Oil and Gas Reservoir Value）。研究首先深入分析了玛湖凹陷的地质特征与油气勘探现状，随后构建了一个融合随机森林（RF）、长短期记忆网络（LSTM）和卷积神经网络（CNN）的集成算法，以此提升储层评价的准确性与泛化能力。在具体实施阶段，通过系统开展预处理与特征工程，提取了关键特征参数，并结合领域专家知识，构建了增维特征，例如烃湿度比、烃平衡比和烃特征比。此外，引入滑动窗口技术追踪特征随深度的变化趋势，利用相似井的类别信息作为先验知识来增强模型的预测能力。最终，通过集成不同模型的优势，构建了一个精确且鲁棒的储层评价算法，该算法能有效识别玛湖凹陷区域中具有工业生产价值的储层，在测试集上的F1分数（F1 Score）、准确率（Accuracy）和曲线下面积（AUC）值分别达到0.847 0、0.772 5和0.781 0。研究还深入探讨了模型的可解释性，旨在为地质学家阐明模型的决策机制，助力其在油气勘探和开发领域中做出更明智的决策。

关键词：储层预测；玛湖凹陷；滑动窗口；集成模型；可解释性
中图分类号：TE122 **文献标识码：**A

Ensemble learning-based prediction model for oil and gas reservoir value in Mahu Sag

YUAN Jing[1], JIA Lu[1], XU Guojian[2], AI Min[1], LI Sixu[1]

(1. Digital Technology Company, PetroChina Xinjiang Oilfield, Karamay, Xinjiang 834000, China; 2. Mahu Exploration and Development Project Department, PetroChina Xinjiang Oilfield Company, Karamay, Xinjiang 834000, China)

Abstract: The Mahu oilfield, located in the northwestern part of the Junggar Basin in Xinjiang, is one of the largest conglomerate oilfields in the world, with reserves exceeding 1 billion tons. However, poor reservoir properties and strong heterogeneity present significant challenges to the efficient development of oil and gas resources. The key to efficient oil and gas development lies in accurately identifying reservoirs with industrial production value, those with higher productivity and relatively lower development costs. To address the complexity of oil and gas reservoir evaluation in the Mahu Sag of the Junggar Basin, this study proposed an oil and gas reservoir value (OGRV) prediction model based on ensemble learning. The study began with an in-depth analysis of the geological characteristics and exploration status of the Mahu Sag. Then, an ensemble model integrating random forest (RF), long short-term memory (LSTM), and convolutional neural network (CNN) was constructed to improve the accuracy and generalization ability of reservoir evaluation. During implementation, key feature parameters were extracted through systematic preprocessing and feature engineering. With expert knowledge, additional augmented features such as hydrocarbon humidity ratio, hydrocarbon balance ratio, and hydrocarbon characteristic ratio were incorporated. In addition, the sliding window technique was introduced to track the trend of features with depth variations, and the category information of similar wells was used as prior knowledge to enhance the model's prediction performance. By leveraging the strengths of different models, a precise and robust reservoir evaluation algorithm was developed. It effectively identified reservoirs with industrial value in the Mahu Sag. The model yielded an F1-score of 0.847 0, accuracy of 0.772 5, and area under the receiver operating characteristic (ROC) curve (AUC) of 0.781 0. The study also investigated model interpretability in depth to help geoscientists better understand the model's decision-making mechanisms and support

收稿日期：2024-07-08。
第一作者简介：袁婧（1998—），女，硕士，助理工程师，现从事油田大数据分析与人工智能研究工作。地址：新疆维吾尔自治区克拉玛依市世纪大道7号，邮政编码：834000。E-mail：sjyuanjing@petrochina.com.cn
通信作者简介：贾鹿（1979—），男，博士，高级工程师，现从事油田信息规划及大数据分析研究工作。地址：新疆维吾尔自治区克拉玛依市世纪大道7号，邮政编码：834000。E-mail：jialu666@petrochina.com.cn
基金项目：新疆维吾尔自治区科学技术厅“天山英才”培养计划项目“油气生产现场智能应用”（2022TSYCJC0032）。

more informed decision-making in oil and gas exploration and development.

Keywords: reservoir prediction; Mahu Sag; sliding window; ensemble model; interpretability

玛湖凹陷位于准噶尔盆地的西北部,是中央坳陷带的次级构造单元,其构造格局主要受到海西、印支、燕山和喜马拉雅四期构造运动的影响[1]。自2012年起在三叠系百口泉组砂砾岩中取得显著的油气勘探突破,尤其是玛湖东、西斜坡地区,已发现多个亿吨级油气田,成为重要的油气勘探区域[2]。玛湖凹陷以扇三角洲沉积为主,岩性较粗,储层物性较差,孔隙度和渗透率相对较低,分别小于10%和1×10^{-3} μm^2,属于致密砾岩储层,且具有极强的非均质性[3]。

高效开发油气资源的关键,在于精确识别有工业生产价值的储层,即那些油气产能较高且开发成本较低的区域。使用中国石油企业标准《油(气)层工业油气流标准及试油结论规定》(Q/SY TZ 0026-2000)作为评判准则,将那些日产油量满足最低工业油流标准、具备适宜的生产气油比、原油密度超过0.8 g/cm^{3}且含水率低于2%的油气层定义为具有工业生产价值的储层。传统的储层评价方法通常基于其物理特性,包括孔隙度、渗透率以及流动性能指标,同时考虑储层的厚度来表征其规模。为了界定不同类型储层的参数,采用了包括油气测试、含油状况分析和基于经验的统计技术在内的多种技术手段[4-10]。在玛湖凹陷的油气勘探与开发实践中,传统方法在面对低孔隙度、低渗透性以及高非均质性等复杂储层特征时,其预测精度往往难以满足高效开发的需求。这些方法主要依赖于经验驱动的知识积累,预测模型的构建和参数优化过程耗时长且缺乏灵活性,难以迅速适应地质条件的多变性。尤其是在玛湖凹陷这类对快速增储上产需求迫切的区域,传统方法的局限性尤为突出,亟须引入更为先进的技术手段,以提高勘探效率和开发成功率。

随着数据分析、人工智能等技术在油气智能勘探和开发领域的兴起,油气工业领域迎来了全新的技术手段[11-14]。这些技术能够高效地处理和分析多维地质数据,通过机器学习算法的自适应优化能力,快速识别油气富集区域,并适应地质条件的变化。与传统方法相比,机器学习方法能够从更广泛的特征维度进行综合判断,包括但不限于地质、钻井、录井、测井和生产数据等,从而揭示数据中隐藏的复杂关系和模式。这种多维度的数据分析能力,为提高油气勘探的精度和效率提供了强有力的技术支持。

马海龙等[15]利用测井评价法和模糊聚类法,综合分析测试、测井、生产和地质等多维数据,来刻画储层的展布特征,从而建立了不同等级储层的划分方案。ZHANG等[16]利用地震数据,通过弹性阻抗反演方法直接估算储层流体的性质和岩石的脆性,预测出非常规储层中的“甜点”区域(即高生产潜力区域)和适合进行水力压裂的裂缝区域。聂云丽等[17]为了解决储层分类中存在的多指标考量、主观经验依赖以及耗时费力的问题,提出了一种基于随机森林算法的自动分类方法。李克文等[18]提出了一种结合机器学习算法的新技术,通过自动化特征分析和模式识别来提高油气勘探中有利区域评价的准确性和效率。

除此之外,随着深度学习的不断发展,基于神经网络的储层预测模型也不断被提出,邓少贵等[19]为了提高薄互层的识别精度,研究采用了粒子群优化算法(PSO)对极限学习机(ELM)模型进行优化,同时选取了8个传统测井参数和3个高分辨率测井曲线来表征储层的“三品质”特征,建立了融合PSO和ELM的薄互层识别混合模型。BANSAL等[20]借助人工神经网络,构建了一张揭示地震数据、测井资料、完井参数以及生产特性间复杂相互作用的图谱,旨在精确定位储层中的“有价值”区域。王迪等[21]提出了一种结合先验信息约束的深度学习方法,通过构建一个全连接网络结构和引入地层格架、地震相等先验约束信息,解决了致密砂岩储层参数与地震数据不直接相关的预测问题,从而实现了对致密砂岩“甜点”区域的定量预测和高精度储层参数预测。

以上方法虽然在一定程度上提升了油气储层评价的准确性,但仍存在局限性。比如,单一模型可能难以充分捕捉储层的复杂性与非均质性,且对不同地质条件的适应性欠佳。此外,现有方法在处理大规模多维数据时,可能面临计算效率低下和模型过拟合的问题。为克服这些挑战,研究提出了一种基于集成算法的油气储层价值预测模型(OGRV),该模型综合运用随机森林(RF)、长短期记忆网络(LSTM)和卷积神经网络(CNN)等多种机器学习算法,以提高储层评价的准确性和泛化能力。

1 研究区域概况

玛湖凹陷位于准噶尔盆地西北部,是一个重要的生烃凹陷(图1),其构造位置表现为:西北邻乌夏断裂带和克百断裂带,东南与夏盐凸起和达巴松凸起相邻,东部与三个泉凸起及英西凹陷相邻,该凹陷内发育了多个扇三角洲沉积体系,以三叠系百口泉组、二叠系上乌尔禾组和下乌尔禾组为主要含油层系,这些层系在油气勘探中具有重要的工业价值,区域的地质特征受到多期构造运动的影响,包括海西运动和印支期的构造活动,这些运动不仅塑造了凹陷的构造格局,而且影响了油气的运聚和储集条件[1,22-25]。

玛湖凹陷的地质特征和油气潜力受到了多种地质因素的综合影响。区域的构造活动和沉积环境的变迁为油

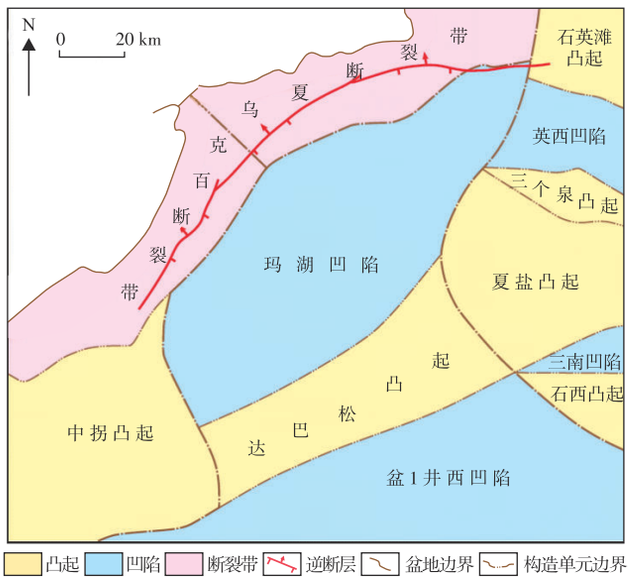

图1　准噶尔盆地玛湖凹陷构造位置（据文献[25]修改）

Fig. 1　Structural location of Mahu Sag in Junggar Basin (modified from Refs.[25])

气资源的分布和勘探提供了重要的宏观背景。湖盆的扩张和收缩，沉积物的搬运和堆积过程，以及随后的成岩作用，共同作用于油气储层的形成和质量的演变[26-27]。这些地质因素的综合作用，使得玛湖凹陷成为油气勘探的重要区域，至今已发现的三级石油地质储量超过1×10^9 t。

2　整体技术路线

在油气勘探和开发过程中，储层预测至关重要，油气储层价值预测问题旨在确定油气藏是否具有工业开采价值。预测模型通常需要处理大量的异构数据，包括钻井数据、录井数据、测井数据、生产数据和其他地质信息，这些数据可能具有不同的尺度、分辨率和不确定性。研究建立的储层预测模型（图2），整体可以分为以下5个阶段：

1）数据准备。数据的精确性和完备性对于评估算法效能起着至关重要的作用。研究收集并整合了包含地质分层、气体测量、9条常规测井曲线、岩性特征、试油数据以及钻井液参数在内的多维数据资料，以便机器学习模型能够从多维度数据中选择与储层分类相关性较高的特征。

2）数据处理。为了提升数据处理的精确度和机器学习算法的兼容性，研究提出了一套分阶段的数据处理框架。该框架主要包括5个步骤：类别特征转换、缺失值处理、特征增维、特征趋势分析和相似井提取。这些步骤共同作用，以提高数据处理的质量和模型的预测性能。

3）特征工程。在搭建油气储层预测模型时，特征选择的准确性对模型的性能具有决定性的影响。特征选择的核心目标在于识别与储层预测高度相关的特征，并确保所选特征具有较好的区分度，同时避免包含与预测目标关联性较弱的特征。这一过程通过排除无关特征来降低模型的计算复杂度，同时确保关键特征的保留，以防模型性能的损失。通过结合数理统计和领域专家知识，研究选择了皮尔逊相关系数和随机森林算法为特征选择提供参考。

4）算法优选。算法模型的选择是一个关键步骤，需要根据问题的性质和数据的特点来挑选最合适的模型。研究将储层价值预测视为二分类问题，即某段储层有无工业开采价值。鉴于样本的复杂性，研究采用了随机森林、LightGBM、LSTM以及CNN4种不同的机器学习模型对储层进行了分类预测。每个模型都预设了初始参数，并通过训练集进行训练，为了防止过拟合，神经网络模型采用了dropout等技术。通过在验证集上的性能评估，对学习率和正则化强度等超参数进行了调整，以提升模型的泛化性能。之后比较不同算法模型在验证集上的性能，选择性能最佳的算法模型，保存其参数和架构以供后续使用。

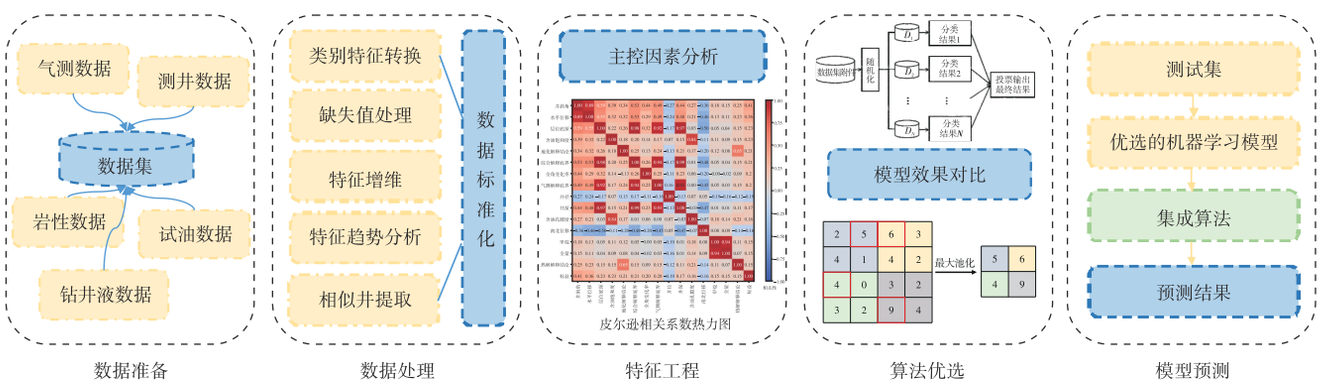

图2　油气储层价值预测建模流程示意图

Fig. 2　Schematic diagram of oil and gas reservoir value prediction modeling workflow

5）模型预测。使用优选的算法模型对测试集进行预测，为了全面评估模型的预测性能，采用了包括准确率（Accuracy）、F1分数（F1 Score）和AUC（Area Under the ROC Curve，即ROC曲线下的面积）在内的多种评价指标。随后，对模型的预测结果进行分析，以深入了解其性能表现及其在实际应用中的潜在价值，最后构建包含随机森林、LSTM和CNN的集成模型OGRV，并输出集成模型的计算过程。

3　数据预处理

3.1　数据集

研究收集了287口井共计15 060个深度点的试油数据。根据试油测试成果，储层样本被划分为具有工业开采价值的油层、气层以及非价值层。具体而言，试油结果表明为工业油层和工业气层的样本被赋予“有价值层”的标签，而其他样本则被标记为“非价值层”。为了系统性地评估模型的泛化性能，数据集根据井号依照8∶1∶1的比例被划分为训练集、验证集和测试集，以便于在模型开发过程中进行有效的训练、调优和性能验证。数据样本的分布情况如表1所示。

表1　数据统计

Table 1　Data statistics

	井数	正样本数	负样本数
训练集	229	9 361	3 636
验证集	29	543	386
测试集	29	829	305

3.2　数据处理

为提升数据处理的精确度并确保其与机器学习算法的兼容性，研究提出了一套分阶段的数据处理框架。该框架包含5个主要环节，分别如下：

1）类别特征转换：对于包含类别信息的特征，如岩石名称和岩屑颜色，进行特征转换，将其转换为数值形式，以便机器学习模型能够识别和处理。

2）缺失值处理：针对数据集中的缺失值，实施两种补救措施。对于缺失数据较少的特征，采用其所在井的样本均值进行填补；而对于缺失数据较多的特征，则选择将其从数据集中剔除。此外，对数据集进行标准化处理，以确保不同特征间数值的一致性。

3）特征增维：依据领域专家的知识，对原始数据进行扩展，增加新的特征维度，如烃湿度比（W_h）、烃平衡比（B_h）和烃特征比（C_h），其计算方法见公式（1）。这些新特征有助于揭示样本的潜在地质信息，增强模型的预测能力。

$$W_h = \frac{C2 + C3 + C4 + C5}{C1 + C2 + C3 + C4 + C5} \times 100$$

$$B_h = \frac{C1 + C2}{C3 + C4 + C5}$$

$$C_h = \frac{C4 + C5}{C3} \tag{1}$$

式中：$C1$为甲烷；$C2$为乙烷；$C3$为丙烷；$C4$为丁烷；$C5$为戊烷。

除此之外，为了同时考虑特征含油饱和度和孔隙度，将两者的乘积视为一个新的特征，并将其命名为含油孔隙度（HYBHD），旨在通过该特征有效减少模型对含油饱和度高但孔隙度极低或孔隙度大但含油饱和度低样本的过高置信评分。

4）特征趋势分析：运用滑动窗口技术来分析样本特征随深度变化的趋势。具体如图3所示，对于每个样本，当滑动窗口大小设定为3时，基于当前样本的井号（JH）和深度（JS）信息，向上下各扩展一个样本，以获取相邻深度点的数据，从而评估样本特征的变化情况。

5）相似井提取：通过K近邻（KNN）算法对样本进行相似性分析，识别与当前样本在特征空间中相似的其他样本。这些相似样本的标签作为先验知识被整合为模型的输入，以提升模型对地质条件变化的适应性和泛化性能。

井号	井深/m	甲烷/%	...	标签
金龙042	3 314	7.501 4	…	1
金龙042	3 315	9.566 2	...	1
金龙042	3 316	10.433 4	…	1
金龙042	3 317	9.591 9	...	1
金龙042	3 318	4.264 0	…	1
金龙042	3 319	6.834 3	...	1
金龙042	3 320	10.642 1	…	1
金龙042	3 321	37.540 9	…	1
…	…	…	…	…

井号	井深/m	甲烷[井深-1]/%	…	甲烷/%	甲烷[井深+1]/%	…	标签
金龙042	3 315	7.501 4	…	9.566 2	10.433 4	…	1

⋮

井号	井深/m	甲烷[井深-1]/%	…	甲烷/%	甲烷[井深+1]/%	…	标签
金龙042	3 320	6.834 3	…	10.642 1	37.540 9	…	1

图3　滑动窗口

Fig. 3　Sliding window

3.3 主控因素分析

在构建油气储层预测模型时，主控因素的选择对于模型的表现至关重要。此步骤的目标是挑选出与预测目标紧密相关的特征，确保这些特征具有高度的辨识度，并剔除与预测目标关联性不强的因素。这样不仅减少了模型的计算复杂度，而且确保了关键特征的获取，进而提升了模型的预测精度和运算效率。初始构建的数据集包含了154个特征维度，其中混杂了大量的干扰信息，这不仅增加了计算量，也可能会对模型的性能产生负面影响。因此，采用相关性分析对特征参数进行优化筛选，在降低计算复杂度的同时，提升模型的预测性能。

在进行主控因素分析时，研究不仅关注了储层的地质特性，也考虑了施工参数的影响，以确保模型能够全面捕捉影响油气储层价值的因素。尽管井斜角、水平位移、层位的顶底深度和井径等特征初看似乎更偏向于工程领域，但它们实际上与储层的地质质量和开发效率紧密相关。这些特征直接影响钻探过程中对油气层的接触和穿透效果，进而决定了油气的开采效率和产量。

3.3.1 皮尔逊相关系数

皮尔逊相关系数（Pearson correlation coefficient），记作r，是评估2个连续变量间线性相依性的一种统计指标。该系数通过比较两组数据的协方差与它们各自标准差的乘积，来量化这2个变量在统计上的协同变化程度。r的值域介于−1与1之间，分别代表着变量间可能的最强正相关（r=1）、最强负相关（r=−1）以及不存在线性相关（r=0）。对于2个变量X和Y，皮尔逊相关系数r的计算公式如式（2）所示：

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{2}$$

其中：X_i和Y_i代表数据集中的观测值，而$\bar{X}$和$\bar{Y}$则是这2个变量的样本均值。i为循环变量，指代当前数据点的序号；n为样本总量，决定了求和的范围。

如图4所示，通过皮尔逊相关系数分析，并将与储层预测最相关的前15个特征可视化可以发现，与储层价值相关性较高的特征主要为井基本信息数据、测井数据、深度数据和解释结论相关的数据。然而，皮尔逊相关系数只衡量线性关系，对于非线性关系可能无法有效反映2个变量之间的实际关联程度。如果数据集中的变量关系是非线性的，皮尔逊相关系数可能会给出误导性的结果。

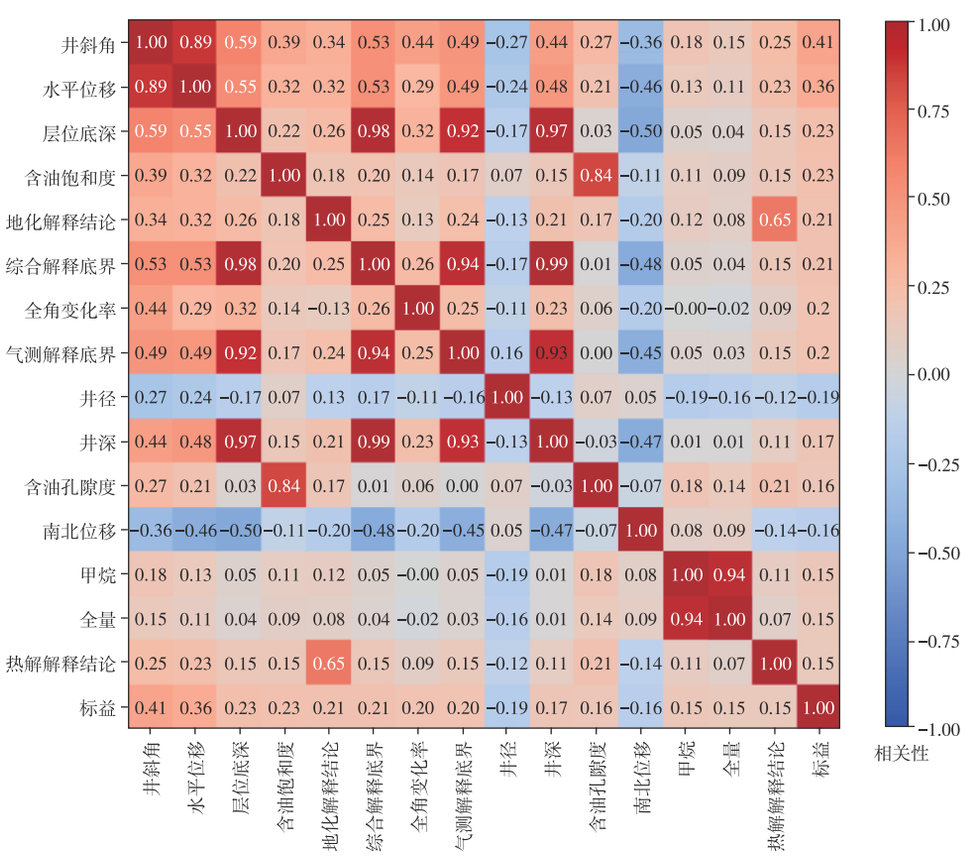

图4　皮尔逊相关系数热力图

Fig. 4　Pearson correlation coefficient heat map

3.3.2　随机森林

随机森林算法是一种集成学习方法，它通过构建并结合多个决策树来进行模式识别和数据分类。该算法通过在训练过程中引入随机性，不仅增强了模型的泛化能力，而且有效降低了过拟合的风险。在构建每棵决策树时，随机森林采用自助采样（bootstrap sampling）获取数据子集，并在分裂过程中随机选择特征子集，从而提高了模型的多样性和鲁棒性。此外，随机森林能够评估各个特征对于预测结果的贡献度，为特征选择和模型解释提供了重要信息。采用随机森林算法对输入特征的重要性进行分析，并对在训练中贡献前15的特征进行排序，结果如表2所示。结果表明：井基本信息数据、测井数据和深度数据对预测结果的贡献较高。通过随机森林算法可以获取特征与目标变量之间的非线性关系，防止剔除非线性相关但对预测结果贡献较大的特征。

表2　随机森林特征贡献度

Table 2　Feature contribution rates in random forest

特征名称	贡献度/%	特征名称	贡献度/%
井斜角	6.83	井径	2.93
水平位移	6.64	测井层位	2.79
出口电导率	4.53	自然伽马	2.25
冲洗带电阻率	3.74	综合解释顶界	2.17
层位名称	3.40	补偿中子	2.16
方位角	3.33	烃特征比	2.12
层位底深	3.01	丙烷	2.09
层位顶深	2.93		

4　储层价值预测模型

在油气勘探领域，油气储层预测模型的建立对于评估油气资源的潜在价值极为关键。准确的预测模型能够从复杂的地质数据中识别出具有开采价值的油气藏，从而优化勘探策略，降低开发成本，并提高资源利用效率。因此，开发一个能够准确处理和分析测井、钻井等多维数据的预测模型，对于提高资源开发的精确性和效率，具有显著的实践意义。

4.1　评价指标

如图5所示，二分类的混淆矩阵是一个二维表格，展示了实际类别与模型预测类别之间的关系，包括真正例（TP）、假正例（FP）、假反例（FN）和真反例（TN）。

4.1.1　准确率

准确率（Accuracy）是模型正确分类的样本数量与总样本数量之比。这一指标在类别分布均匀的数据集中是一种简单直观的性能度量，但在类别不平衡的情况下，高准确率可能掩盖了对少数类别的低效识别。其计算过程如式（3）所示。

真实结果	预测结果	
	有价值层（Positive）	无价值层（Negative）
有价值层（Positive）	真正例（P_T）	假反例（N_F）
无价值层（Negative）	假正例（P_F）	真反例（N_T）

图5　混淆矩阵

Fig. 5　Confusion matrix

$$A = \frac{P_T + N_T}{P_T + P_F + N_F + N_T} \tag{3}$$

式中：A为准确率。

4.1.2　精确率

精确率（Precision）是指模型预测为正类的样本中实际为正类的比例。它衡量的是模型预测正类的能力，即尽量减少将负类预测为正类的错误，因此精确率又被称为查准率。其计算过程如式（4）所示。

$$P = \frac{P_T}{P_T + P_F} \tag{4}$$

式中：P为精确率。

4.1.3　召回率

召回率（Recall）是指实际为正类的样本中被模型正确预测为正类的比例。它衡量的是模型能否将所有的正类都正确预测出来，因此召回率又被称作查全率。其计算过程如式（5）所示。

$$R = \frac{P_T}{P_T + N_F} \tag{5}$$

式中：R为召回率。

4.1.4　F1分数

F1分数（F1 Score）是对精确率和召回率的综合评估，是精确率和召回率的调和平均数，已知两个数的调和平均数会更靠近较小的那一个数，因此，越高的F1分数，能保证分类模型同时具有较高的精确率和召回率，F1分数的取值范围为[0，1]，越靠近1表明分类效果越好。其计算过程如式（6）所示。

$$F = \frac{2 \times P \times R}{P + R} \tag{6}$$

式中：F为F1分数。

4.1.5 AUC

AUC(Area Under the ROC Curve，即ROC曲线下的面积)是一种用于评估分类模型性能的指标。ROC曲线是一种以真正例率(True Positive Rate，又称召回率)为纵轴，假正例率(False Positive Rate)为横轴的图形，用于表示在不同分类阈值下模型的性能。AUC值是ROC曲线下方的面积，其取值范围在0到1之间。AUC值越接近1，说明模型性能越好；而AUC值越接近0.5，则说明模型性能越差，类似于随机猜测。其示意图如图6所示。

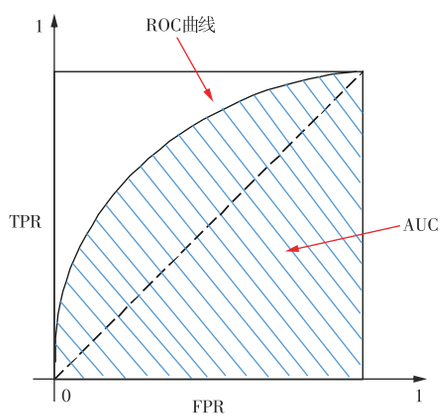

图6 ROC曲线示例图

Fig. 6 ROC curve diagram

AUC对类别不平衡相对不敏感。在正负样本比例失衡的情况下，AUC依然能够提供一个公平的性能评估，而其他指标如准确率可能会受到样本不平衡的影响。

4.2 模型搭建

4.2.1 随机森林

随机森林是一种集成学习算法，通过构建多个决策树并对它们的预测结果进行集成来提高整体模型的准确性和鲁棒性。在随机森林中，每棵树都是独立训练的，并在训练过程中引入随机性，例如通过随机选择特征子集来构建每棵树，这有助于减少过拟合并增强模型的泛化能力。在分类任务中，随机森林通过投票机制确定最终类别，而在回归任务中则通常取平均预测值。随机森林的优势在于其易于实现、对超参数不太敏感，并且能够处理高维数据和缺失值，同时提供了特征重要性的评估，这使得它在各种数据挖掘和机器学习任务中广受欢迎。

4.2.2 LightGBM

LightGBM是一种高效的机器学习算法，它采用基于梯度提升的方法来进行模型训练和预测。与随机森林等集成学习方法不同，LightGBM不是通过投票机制来集成多个决策树，而是通过顺序地构建一系列决策树，每个树都试图减少前一个模型的预测误差。算法会逐步添加新的决策树，每一轮都会考虑到之前树的预测结果，并通过优化损失函数来提升整体模型的性能。这种逐步提升的方式使得LightGBM能够聚焦于模型的弱点，并逐步改进。

此外，LightGBM还引入了基于直方图的梯度提升技术，这种技术可以在不牺牲模型性能的前提下，显著提高训练速度和减少内存使用。这使得LightGBM特别适合处理大规模数据集，同时保持了模型的高效性和准确性。

4.2.3 LSTM

LSTM[28]是一种针对序列数据建模的深度学习架构，它通过精巧的门控机制优化了传统循环神经网络(RNN)在处理长期依赖问题上的性能局限。LSTM单元内部的输入、遗忘和输出门控制着信息的流动，实现了对关键信息的持久化存储与适时遗忘，从而在序列学习任务中表现出卓越的时间动态特征捕捉能力。

4.2.4 CNN

CNN是一种深度学习架构，专注于从数据中自动学习空间特征。通过使用一系列的卷积层，CNN能够捕捉输入数据中的局部模式，并逐步构建出更高层次的特征表示，由于其强大的特征提取能力，CNN被广泛应用于多种领域，包括语音识别、自然语言处理、时间序列分析以及任何需要识别空间或时间上局部模式的数据处理任务。CNN的设计使其在保持模型参数数量的同时，能够处理大规模和高维度的数据，从而在机器学习社区中得到了广泛的认可和应用。

4.3 实验结果

基于3.1节按照8∶1∶1切分好的训练集分别使用以上4种机器学习算法进行训练，并使用验证集进行超参数的调整，最后使用测试集对模型的分类性能进行评估，结果如表3和表4所示。

表3展示了仅使用数据集本身含有的特征进行训练的结果。根据主控因素分析，研究从154维特征中优选了45个特征作为模型最终的输入，主要包含井基本信息数据、气测数据、深度数据、解释结果数据和9条测井曲线数据。表4则加入了增维数据，包括烃湿度比(W_h)、烃平衡比(B_h)和烃特征比(C_h)，含油饱和度(HYBHD)，相似井信息以及滑动窗口数据。

由实验结果可以看出，在测试集上，随机森林模型无论是在包含还是不包含增维特征的情况下，都表现出较

表3　分类性能-不含增维特征

Table 3　Classification performance-without augmented features

算法	样本集	F1分数	精确率	AUC值
随机森林	训练集	0.889 1	0.827 1	0.926 1
	验证集	0.792 0	0.710 4	0.808 1
	测试集	0.868 4	0.785 7	0.750 1
LightGBM	训练集	0.954 6	0.932 1	0.989 5
	验证集	0.815 0	0.750 3	0.861 4
	测试集	0.853 1	0.761 9	0.745 5
LSTM	训练集	0.897 7	0.848 9	0.914 8
	验证集	0.750 0	0.668 5	0.784 9
	测试集	0.825 2	0.733 7	0.697 3
CNN	训练集	0.892 1	0.841 6	0.916 0
	验证集	0.779 8	0.704 0	0.803 1
	测试集	0.783 7	0.691 4	0.686 5

表4　分类性能-含增维特征

Table 4　Classification performance-with augmented features

算法	样本集	F1分数	精确率	AUC值
随机森林	训练集	0.989 3	0.984 5	0.999 4
	验证集	0.795 0	0.715 8	0.828 9
	测试集	0.861 8	0.784 8	0.773 7
LightGBM	训练集	0.982 4	0.974 4	0.997 6
	验证集	0.781 1	0.706 1	0.823 0
	测试集	0.824 1	0.727 5	0.727 6
LSTM	训练集	0.953 7	0.932 8	0.978 0
	验证集	0.804 0	0.747 0	0.803 4
	测试集	0.825 4	0.738 1	0.721 7
CNN	训练集	0.919 2	0.882 4	0.943 8
	验证集	0.793 4	0.732 0	0.813 2
	测试集	0.767 2	0.677 2	0.700 5

好的分类效果。这是因为随机森林模型对多特征数据的处理能力较强，能够捕捉到特征之间的复杂关系。除LightGBM外，在新增特征后，随机森林、LSTM和CNN在指标AUC上均有一定的提升，在测试集上分别增长了2.36、2.44和1.40个百分点。根据数据分布可知，研究中正负样本的分布并不均衡，不能单纯依赖准确率(Accuracy)指标来准确评估模型的有效性。值得指出的是，AUC指标在面对类别不平衡的情况下，受到的影响相对较小，在测试集上指标AUC的提升进一步证实了新增特征在提升模型性能方面的有效性。

为了进一步的提升模型的预测效果，研究采用了集成算法中的堆叠法(图7)，通过结合多个基础模型的预测结果，训练出1个元模型，从而整合各个基础模型的优势，实现更为精确的预测。具体来说，将随机森林、LSTM和CNN模型的输出构建为下一层模型的输入，原因在于逻辑回归能够提供明确的输出公式，通过这种方法，模型能够在综合不同模型的预测结果的同时，保持对预测过程的可解释性。由于LightGBM在新增特征后AUC值不增反降，因此将其移出集成算法OGRV的构建过程中。实验结果如表5所示，可以看出相较表4中效果最好的模型，集成算法OGRV在指标AUC上在测试集上提高了0.73个百分点。

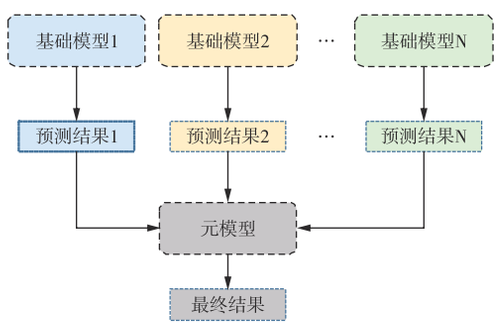

图7　模型堆叠

Fig. 7　Model stacking

表5　分类性能

Table 5　Classification performance

算法	样本集	F1分数	精确率	AUC值
集成算法OGRV	训练集	0.993 7	0.990 9	0.999 6
	验证集	0.781 1	0.725 5	0.812 5
	测试集	0.847 0	0.772 5	0.781 0

输出集成算法的计算过程如下：

$$y=\frac{1}{1+\exp-[(-9.97+17.97x_1+3.17x_2-3.54x_3)]} \tag{7}$$

式中：x_1、x_2、x_3分别表示模型随机森林、LSTM和CNN的预测结果，y为集成模型OGRV最终的预测结果，由公式可以看出随机森林对于预测结果的影响程度最大，LSTM和CNN的贡献基本一致。

4.4　应用示例

传统方法通常需要地质专家通过多次现场勘查和分析，依据孔隙度、渗透率、含油饱和度等参数设定阈值区间，以确定储层的“甜点”等级。这种方法不仅耗时，而且结果的准确性高度依赖于专家的主观经验。特别是当勘探区域的地质条件发生变化时，需要重新划分阈值，这导致工作量巨大且难以保证结果的一致性。

相较而言，提出的基于集成算法的储层价值预测模型OGRV则展现出显著的优势。该模型能够自动处理和分析包括测井、录井、钻井在内的多源地质数据。通过机

器学习算法的应用，模型能够快速识别油气富集区域，并适应地质条件的变化。与传统方法相比，OGRV模型在数据处理的广度和深度上都有显著提升，能够揭示数据中隐藏的复杂关系和模式，从而显著提高勘探的精度和效率。此外，在勘探区域发生变化时，OGRV将先使用全量数据预训练模型，在新区块预测时选用代表性井做模型微调，便可迅速适应新的地质条件，较人工而言，其“经验”累积与更新速度更快。

表6以A井为例，展示了OGRV的预测结果，输入相应的特征参数，模型将会给出预测的结果。可以看出OGRV模型对于储层内部大多数深度点的预测结果都是准确的，显示出模型具有良好的预测能力。然而，在储层的交界处，模型的预测出现了一定的偏差。这种偏差可能源于多个因素，包括数据特性的变化、模型泛化能力的局限、特征工程的不足以及地质条件的复杂性。因此，未来的研究将集中在这些方面以进一步提升OGRV在油气储层价值预测中的准确性和可靠性。

表6 OGRV预测示例

Table 6 OGRV prediction examples

井号	井深/m	试油结果	试油结论	预测结果	是否预测准确
A井	3 622	干层	干层	非价值层	是
	3 623	干层	干层	非价值层	是
	3 624	干层	干层	非价值层	是
	3 625	干层	干层	非价值层	是
	3 626	干层	干层	非价值层	是
	3 627	干层	干层	非价值层	是
	3 628	干层	干层	非价值层	是
	3 629	干层	干层	非价值层	是
	3630	干层	干层	非价值层	是
	3 631～3 662				
	3 663	油层	工业油层	非价值层	否
	3 664	油层	工业油层	非价值层	否
	3 665	油层	工业油层	有价值层	是
	3 666	油层	工业油层	有价值层	是
	3 667	油层	工业油层	有价值层	是
	3 668	油层	工业油层	有价值层	是
	3 669	油层	工业油层	有价值层	是
	3 670	油层	工业油层	有价值层	是
	3 671	油层	工业油层	有价值层	是
	3 672	油层	工业油层	有价值层	是
	3 673	油层	工业油层	有价值层	是
	3 674	油层	工业油层	非价值层	否

数据采集按照一米一段的精细度进行，旨在精准捕捉玛湖凹陷区域储层特征的微妙变化，尤其是在该区域非均质性显著的情况下。这种连续的数据采集方式能够细致地映射储层的微观波动，从而为预测模型提供详尽的信息，增强预测的准确性。在预测结果的呈现阶段，将具有连续相同预测结果的层段进行聚合，仅展示这些层段的顶底深度，以此简化数据输出。该策略不仅提升了数据的可读性，还确保了地质解释的精确度，从而为油气储层的评价工作提供了精确且可靠的预测结果。

5 结论

针对玛湖凹陷油气储层价值预测问题，研究构建了集成算法OGRV模型。该模型融合随机森林、LSTM和CNN 3种算法的优势，综合了井基本信息数据、气测数据、深度数据、解释结果数据和9条测井曲线数据等多维度特征，有效提高了储层评价的准确性和泛化能力。通过预处理和特征工程，提取了关键特征参数，并结合领域专家知识构建了增维特征，同时引入滑动窗口技术和相似井信息，进一步增强了模型的预测能力。模型评估结果显示，OGRV模型在测试集上取得了AUC为0.781 0的优异性能，能够有效识别具有工业生产价值的储层区域。此外，研究还探讨了模型的可解释性，旨在为地质学家阐明模型的决策机制，助力其在油气勘探和开发领域做出更明智的决策。

参 考 文 献

[1] 钟厚财, 朱俊梅, 林煜, 等. 玛湖凹陷三叠系百口泉组地层再认识与勘探潜力分析 [J]. 特种油气藏, 2024, 31 (2): 28-36.
Zhong Houcai, Zhu Junmei, Lin Yu, et al. Re-understanding and exploration potential analysis of the Triassic Baikouquan Formation in Mahu Depression [J]. Special Oil and Gas Reservoirs, 2024, 31 (2): 28-36.

[2] 操应长, 燕苗苗, 葸克来, 等. 玛湖凹陷夏子街地区三叠系百口泉组砂砾岩储层特征及控制因素[J]. 沉积学报, 2019, 37(5): 945-956.
CAO Yingchang, YAN Miaomiao, XI Kelai, et al. The characteristics and controlling factors of glutenite reservoir in the Triassic Baikouquan formation, Xiazijie area, Mahu depression[J]. Acta Sedimentologica Sinica, 2019, 37(5): 945-956.

[3] 唐勇, 宋永, 郭旭光, 等. 准噶尔盆地玛湖凹陷源上致密砾岩油富集的主控因素[J]. 石油学报, 2022, 43(2): 192-206.
TANG Yong, SONG Yong, GUO Xuguang, et al. Main controlling factors of tight conglomerate oil enrichment above source kitchen in Mahu sag, Junggar Basin[J]. Acta Petrolei Sinica, 2022, 43(2): 192-206.

[4] 邵长新, 王艳忠, 操应长. 确定有效储层物性下限的两种新方法及应用: 以东营凹陷古近系深部碎屑岩储层为例[J]. 石油天然气

学报, 2008, 30(2): 414-416.
SHAO Changxin, WANG Yanzhong, CAO Yingchang. Two new methods used to determine the low limits of effective reservoir physical properties and their applications: A case study on deep clastic reservoir of Palaeogene in Dongying depression[J]. Journal of Oil and Gas Technology, 2008, 30(2): 414-416.

[5] 刘之的, 石玉江, 周金昱, 等. 有效储层物性下限确定方法综述及适用性分析[J]. 地球物理学进展, 2018, 33(3): 1102-1109.
LIU Zhidi, SHI Yujiang, ZHOU Jinyu, et al. Review and applicability analysis of determining methods for the lower limit of physical properties of effective reservoirs[J]. Progress in Geophysics, 2018, 33(3): 1102-1109.

[6] 赵政璋, 杜金虎, 邹才能, 等. 大油气区地质勘探理论及意义[J]. 石油勘探与开发, 2011, 38(5): 513-522.
ZHAO Zhengzhang, DU Jinhu, ZOU Caineng, et al. Geological exploration theory for large oil and gas provinces and its significance [J]. Petroleum Exploration and Development, 2011, 38(5): 513-522.

[7] 李新宁, 马强, 梁辉, 等. 三塘湖盆地二叠系芦草沟组二段混积岩致密油地质特征及勘探潜力[J]. 石油勘探与开发, 2015, 42(6): 763-771.
LI Xinning, MA Qiang, LIANG Hui, et al. Geological characteristics and exploration potential of diamictite tight oil in the second Member of the Permian Lucaogou Formation, Santanghu Basin, NW China[J]. Petroleum Exploration and Development, 2015, 42(6): 763-771.

[8] 邹才能, 张国生, 杨智, 等. 非常规油气概念、特征、潜力及技术: 兼论非常规油气地质学[J]. 石油勘探与开发, 2013, 40(4): 385-399.
ZOU Caineng, ZHANG Guosheng, YANG Zhi, et al. Geological concepts, characteristics, resource potential and key techniques of unconventional hydrocarbon: On unconventional petroleum geology [J]. Petroleum Exploration and Development, 2013, 40(4): 385-399.

[9] 刘叶轩, 刘向君, 丁乙, 等. 考虑隔层影响的页岩油储层可压性评价方法[J]. 油气藏评价与开发, 2023, 13(1): 74-82.
LIU Yexuan, LIU Xiangjun, DING Yi, et al. Evaluation method of fracability of shale oil reservoir considering influence of interlayer[J]. Petroleum Reservoir Evaluation and Development, 2023, 13(1): 74-82.

[10] 张进, 田洪波, 胡宇新. 深层油藏大斜度井深抽技术对策[J]. 油气藏评价与开发, 2023, 13(2): 247-253.
ZHANG Jin, TIAN Hongbo, HU Yuxin. Technical countermeasures for deep pumping of highly deviated wells in deep reservoir[J]. Petroleum Reservoir Evaluation and Development, 2023, 13(2): 247-253.

[11] 林伯韬, 郭建成. 人工智能在石油工业中的应用现状探讨[J]. 石油科学通报, 2019, 4(4): 403-413.
LIN Botao, GUO Jiancheng. Discussion on current application of artificial intelligence in petroleum industry[J]. Petroleum Science Bulletin, 2019, 4(4): 403-413.

[12] 李宁, 徐彬森, 武宏亮, 等. 人工智能在测井地层评价中的应用现状及前景[J]. 石油学报, 2021, 42(4): 508-522.
LI Ning, XU Binsen, WU Hongliang, et al. Application status and prospects of artificial intelligence in well logging and formation evaluation[J]. Acta Petrolei Sinica, 2021, 42(4): 508-522.

[13] 杨午阳, 魏新建, 李海山. 智能物探技术的过去、现在与未来[J]. 岩性油气藏, 2024, 36(2): 170-188.
YANG Wuyang, WEI Xinjian, LI Haishan. The past, present and future of intelligent geophysical technology[J]. Lithologic Reservoirs, 2024, 36(2): 170-188.

[14] 马乔雨, 张欣, 张春雷, 等. 基于一维卷积神经网络的横波速度预测[J]. 岩性油气藏, 2021, 33(4): 111-120.
MA Qiaoyu, ZHANG Xin, ZHANG Chunlei, et al. Shear wave velocity prediction based on one-dimensional convolutional neural network[J]. Lithologic Reservoirs, 2021, 33(4): 111-120.

[15] 马海龙, 王兆生, 王爱霞, 等. 准噶尔盆地玛湖凹陷西北斜坡区三叠系百口泉组二段储层分类与"甜点"预测[J]. 大庆石油地质与开发, 2023, 42(4): 30-36.
MA Hailong, WANG Zhaosheng, WANG Aixia, et al. Reservoir classification and "sweet sports" prediction of the 2nd member of Triassic Baikouquan Formation of northwestern slope area in Mahu Sag of Junggar Basin[J]. Petroleum Geology & Oilfield Development in Daqing, 2023, 42(4): 30-36.

[16] ZHANG S, HUANG H, DONG Y, et al. Direct estimation of the fluid properties and brittleness *via* elastic impedance inversion for predicting sweet spots and the fracturing area in the unconventional reservoir[J]. Journal of Natural Gas Science and Engineering, 2017, 45: 415-427.

[17] 聂云丽, 高国忠. 基于随机森林的页岩气"甜点" 分类方法[J]. 油气藏评价与开发, 2023, 13(3): 358-367.
NIE Yunli, GAO Guozhong. Classification of shale gas "sweet spot" based on Random Forest machine learning[J]. Petroleum Reservoir Evaluation and Development, 2023, 13(3): 358-367.

[18] 李克文, 周广悦, 路慎强, 等. 一种基于机器学习的有利区评价新方法[J]. 特种油气藏, 2019, 26(3): 7-11.
LI Kewen, ZHOU Guangyue, LU Shenqiang, et al. A new method for favorable zone evaluation based on machine learning[J]. Special Oil & Gas Reservoirs, 2019, 26(3): 7-11.

[19] 邓少贵, 张凤姣, 陈前, 等. 基于混合机器学习算法的页岩薄互层识别方法[J]. 石油学报, 2023, 44(7): 1097-1104.
DENG Shaogui, ZHANG Fengjiao, CHEN Qian, et al. Identification of shale thin interbeds based on hybrid machine learning algorithm [J]. Acta Petrolei Sinica, 2023, 44(7): 1097-1104.

[20] BANSAL Y, ERTEKIN T, KARPYN Z, et al. Forecasting well performance in a discontinuous tight oil reservoir using artificial neural networks[C]//Paper SPE-164542-MS presented at the SPE Unconventional Resources Conference, Texas, USA, April, 2013.

[21] 王迪, 张益明, 张繁昌, 等. 利用先验信息约束的深度学习方法定

量预测致密砂岩"甜点"[J]. 石油地球物理勘探, 2023, 58(1): 65-74.

WANG Di, ZHANG Yiming, ZHANG Fanchang, et al. Quantitative prediction of tight sandstone sweet spots based on deep learning method with prior information constraints[J]. Oil Geophysical Prospecting, 2023, 58(1): 65-74.

[22] 庞宏, 尤新才, 胡涛, 等. 准噶尔盆地深部致密油藏形成条件与分布预测: 以玛湖凹陷西斜坡风城组致密油为例[J]. 石油学报, 2015, 36(增刊2): 176-183.

PANG Hong, YOU Xincai, HU Tao, et al. Forming conditions and distribution prediction of deep tight reservoir in Junggar Basin: A case study from tight reservoir of Fengcheng Formation in the west slope of Mahu sag[J]. Acta Petrolei Sinica, 2015, 36(Suppl. 2): 176-183.

[23] 姜福杰, 黄任达, 胡涛, 等. 准噶尔盆地玛湖凹陷风城组页岩油地质特征与分级评价[J]. 石油学报, 2022, 43(7): 899-911.

JIANG Fujie, HUANG Renda, HU Tao, et al. Geological characteristics and classification evaluation of shale oil in Fengcheng Formation in Mahu sag, Junggar Basin[J]. Acta Petrolei Sinica, 2022, 43(7): 899-911.

[24] 陈静, 陈军, 李卉, 等. 准噶尔盆地玛中地区二叠系—三叠系叠合成藏特征及主控因素[J]. 岩性油气藏, 2021, 33(1): 71-80.

CHEN Jing, CHEN Jun, LI Hui, et al. Characteristics and main controlling factors of Permian: Triassic superimposed reservoirs in central Mahu Sag, Junggar basin[J]. Lithologic Reservoirs, 2021, 33(1): 71-80.

[25] 雷海艳, 齐婧, 周妮, 等. 玛湖凹陷玛页1井风城组富硅页岩成因及其油气意义[J]. 新疆石油地质, 2022, 43(6): 724-732.

LEI Haiyan, QI Jing, ZHOU Ni, et al. Genesis and petroleum significance of silica-rich shale in Fengcheng formation of well Maye-1, Mahu Sag[J]. Xinjiang Petroleum Geology, 2022, 43(6): 724-732.

[26] 杜猛, 向勇, 贾宁洪, 等. 玛湖凹陷百口泉组致密砂砾岩储层孔隙结构特征[J]. 岩性油气藏, 2021, 33(5): 120-131.

DU Meng, XIANG Yong, JIA Ninghong, et al. Pore structure characteristics of tight glutenite reservoirs of Baikouquan Formation in Mahu Sag[J]. Lithologic Reservoirs, 2021, 33(5): 120-131.

[27] 马永平, 张献文, 朱卡, 等. 玛湖凹陷二叠系上乌尔禾组扇三角洲沉积特征及控制因素[J]. 岩性油气藏, 2021, 33(1): 57-70.

MA Yongping, ZHANG Xianwen, ZHU Ka, et al. Sedimentary characteristics and controlling factors of fan-delta of the Upper Urho Formation of Permian in Mahu Sag[J]. Lithologic Reservoirs, 2021, 33(1): 57-70.

[28] 任欢, 王旭光. 注意力机制综述 [J]. 计算机应用, 2021, 41 (增刊1): 1-6.

Ren Huan, Wang Xuguang. Review of attention mechanisms [J]. Computer Applications, 2021, 41 (Suppl. 1): 1-6

(编辑 徐佩)