中国石油大学（北京）

China University of Petroleum Beijing

2022-2023 第一学期

《油气人工智能基础及应用》大作业

姓名：付楷涵

学号：2022211307

班级：研 22-1 班

2022 年 11 月 30 日

基于深度森林方法进行测井岩性识别

姓名：付楷涵[1]

(1. 中国石油大学（北京）理学院，中国 北京，102299)

摘要：我国油气资源对外依存度非常高，致密油气是替代常规油气资源、引导我国油气革命的重要力量。而致密油气储层非常重要的渗流通道之一就是裂缝，单井裂缝的测井解释对分析裂缝期次、厘清裂缝纵横向发育规律、提高三维裂缝网络建模精度等均具有重要意义。如何利用现有的人工智能技术进行裂缝的识别与研究是一个非常重要的课题。本文的主要工作包括以下三个方面：

一、介绍深度森林算法以及 KNN 算法的背景以及算法细节，官方文档中没有多粒度扫描模块的源代码，因此**自实现**了完整的深度森林算法；

二、通过学习深度森林算法中的级联思想以及老师的启发下，将深度森林算法中级联森林结构中的随机森林模块**替换为 KNN 算法**，尝试将基本分类模型进行级联来构造深度学习模型的新方式；

三、利用自实现的完整深度森林算法以及替换为 KNN 模块的算法进行测井岩性识别分类实验，对实验结果进行分析，并比较不同参数对算法效果的影响。进而为利用深度学习处理测井岩性分类问题提供一些思路。

关键词：深度森林；测井岩性识别；KNN；分类模型

1 绪论

1.1 问题研究背景及意义

致密油气是指储集在覆压基质渗透率小于或等于$0.1 \times 10^{-3}\mu m^2$的致密砂岩、致密碳酸盐岩等储集层中的石油气。其是现如今石油工业主要进行开采研究的一个新兴的领域，同样也是全世界非常重要的一种油气资源，我国油气资源对外依存度非常高，而致密油气正是替代常规油气资源、引导我国油气革命的重要力量[1]。致密油气储层非常重要的渗流通道之一就是裂缝，其也是致密储层的有效储集空间。如何识别与研究裂缝影响着致密油气的富集、及时设计与调整裂缝的开发方案以及单井产能与开发效果，是决定致密油气藏是否具有经济开采价值的关键因素[2]。而单井裂缝的测井解释对分析裂缝期次、厘清裂缝纵横向发育规律、提高三维裂缝网络建模精度等均具有重要意义[3,4]。随着人工智能技术的发展，如何将现有的人工智能技术应用到裂缝识别与研究领域是一个非常重要的课题。本次课程论文的主要研究内容在于如何利用人工智能方法建立有效、高精度的测井岩性识别模型。

利用已有的裂缝解释资料可以获得带有裂缝属性的标签信息，其比较常见的标签有：有裂缝、无裂缝、高角度裂缝、低角度裂缝、构造裂缝、成岩裂缝、裂缝密度等等[2]。目前研究测井裂缝识别的主要人工智能方法有有监督学习、无监督学习和半监督学习。而其中有监督学习（Supervised learning）的主要目标是对有标签的数据进行训练，进而对未知的数据进行预测，经典的分类和回归问题都是有监督的学习方法。本文的主要研究问题是当裂缝标签是上述的离散变量时，利用分类方法对测井岩性识别模型进行分类。

而深度森林算法（Deep Forest）是周志华等人[5]提出的非神经网络的深度模型，他们提出传统深度学习得以成功得益于三个重要特性：逐层处理、内置特征转换和充分的模型复杂度，其利用不可微的决策树森林为基本的模块，通过级联的方式构建深度学习模型[6]。其对于大样本数据进行分类任务时具有非常成功的效果。

在了解了深度森林算法的算法流程后，只有了解了其深层次的相连关系后才能对算法的效果进行分析与评估。但在官方的 Github 上只有其中级联森林部分的源代码，对于多粒度扫描部分并没有进行公开。因此，本研究期望在学习深度森林算法后自己实现其源代码并将其应用到测井岩性识别模型的分类问题中，并且在老师的启发下尝试利用 KNN 算法替代级联森林的随机森林模块，得到改良的深度森林算法并将其应用到测井岩性识别模型的分类问题。通过最后的效果对比，为将基本的分类模型进行级联构造深度学习模型处理测井岩性识别模型提供一种新的思路。

1.2 国内外研究现状

Shi.等人在 2008 年将 SVM 算法应用于南襄盆地泌阳凹陷安棚油田裂缝识别中，得到了较好的分类结果，并通过对比得到其效果要优于 BP 神经网络[7]。2010 年郑军等人同样利用 SVM 算法，通过多个参数调整建立裂缝识别分类模型，在处理阿曼盆地 Daleel 油田碳酸盐岩储层裂缝识别时在速度上较 BP 神经网络有很大提升[8]。4 年后何胡军等人基于 KNN 算法，将其与测井曲线的斜率相融合，将其应用到普光气田长兴组以及飞仙关组的裂缝识别中，取得了较好的结果[9]。Bhattacharya and Mishra 等人在 2018 年利用贝叶斯网络对地缝预测模型的地质意义进行分析预测，将结果与随机森林算法进行比较具有较好的效果[10]。同时，他们利用随机森林算法，利用有限的常规测井资料，利用贡献度表明 Delta_CALL 最为重要。

2021 年，Asmari 组碳酸盐岩储层裂缝识别中，集成算法中随机森林算法和 SVM 相比于决策树算法和多

层感知机算法表现出更好的裂缝识别效果[11]。董少群等人在 2022 年总结了利用人工智能方法在处理致密储层裂缝识别中的应用[2]，图 1 展示了对于分类问题中机器识别裂缝的原理图。

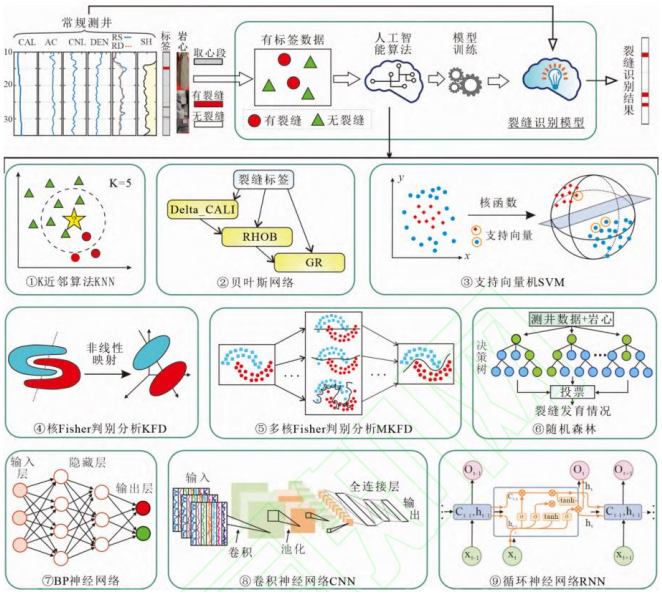

图 1 有监督机器学习机器识别裂缝原理图[2]

Dong 等人在 2022 年应用多核 KFD 算法（MKFD）整合不同特征空间中的裂缝特征信息，更加全面的建立了高鲁棒性的裂缝识别模型，其实验结果相较于单核 KFD 约提升 13%[12]。2015 年，Nouri-Taleghaniden 等人通过遗传算法确定识别结果的权值，利用 RBF 神经网络、MLP 和 LSSVM 对伊朗 Marun 油田的碳酸盐岩储层识别，实验结果通过组合后的识别效果要明显由于单个算法[13]。Tian 等人在 2021 年利用离散小波变换和变点检测法增强声波测井对裂缝的敏感性，通过边界估计将整个测井曲线分割成不重叠的小段，随后建立基于自动编码器和卷积神经网络分类器的深度神经网络模型，将结果与 SVM、随机森林、AdaBoost 算法相比较，其提出的算法与手动裂缝解释结果更加吻合[14]。

总的来看，目前国内外对于测井岩性识别问题中利用深度森林方法的研究较少。如何在学习完深度森林方法的原理后自己实现算法，并将其应用于处理测井岩性识别问题；以及在学习深度森林方法的构造思想后应用其他分类模型进行级联来构造深度模型，是本文的研究重点。

1.3 论文组织结构

本文剩余部分的结构如下。

第 2 章中主要介绍了经典深度森林算法的流程，以及如何利用 KNN 算法的级联构成深度学习模型。

第 3 章利用自实现的深度森林算法以及 KNN 算法级联的深度森林算法进行测井岩性识别的实验，将实验结果与经典的分类算法进行对比，总结算法效果以及对参数进行分析。

第 4 章主要说明了本次结课论文中做了哪些工作，以及对应论文题目中的相应要求。

第 5 章给出了一些结论。

2 深度森林方法及改进

本章主要介绍了经典深度森林算法（Deep Forest）算法的实现过程，以及如何对于经典的深度森林算法进行改良——将级联森林模块中的随机森林模型替换为 KNN 算法，为利用基本分类模型的级联构成深度学习提供新的思路。

2.1 深度森林算法

深度森林算法（Deep Forest）是由 Zhou 等人于 2017 年提出的[5]，Zhou 等人揭示出深度学习得以成功的三个重要特性：逐层处理、内置特征转换和充分的模型复杂度。基于这三个特性，Zhou 等人提出了第一个非神经网络的深度模型——gcforst 模型。该模型基于不可微的决策树森林为基本模块，通过级联的方式构建深度模型[6]。深度森林模型的提出为深度学习打开了一扇新的大门，展示了不依赖于梯度以及误差反向传播而构造深度学习的可行性。

为了实现上述深度学习所具备的三个特性，深度森林算法可以分为两个部分——多粒度扫描（Multi-grained scanning）和级联森林（Cascade forest structure）。先对多粒度扫描模块进行介绍：

1、多粒度扫描

深度神经网络在处理特征方面具有强大的能力，gcForest 受到卷积神经网络在处理图像数据这类原始像素存在位置关系的任务上的处理方式的启发，增加了多粒度扫描模块来处理特征间的相互关系。图 2 展示了多粒度扫描模块的总体进程，其共分为三个环节。

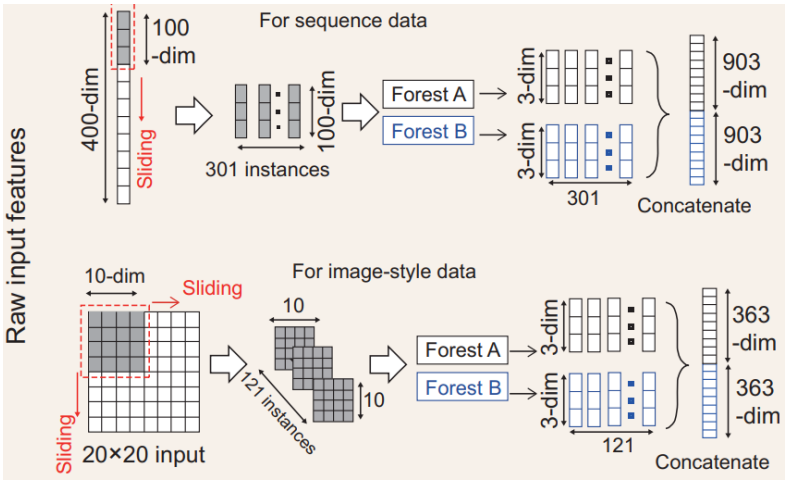

图 2 gcForest 的多粒度扫描示意图[5]

在第一个环节中，滑动窗口被用于对数据原始的属性特征进行扫描。以图 2 中的上方的数据处理方式为例，原始的数据特征为 400 维，若使用 100 维的滑动窗口依次对数据进行扫描，最后可以得到 301 个维度为 100 的新的特征向量。对于下方的例子中，原始数据的维度为 20×20，若采用 10×10 的滑动窗口进行扫描则可以生成 121 个新的特征向量。

在生成新的特征向量后，第二个环节中将新的特征向量视作新的训练样本，利用这些新的训练样本对两种不同的类型的森林进行训练，而每个森林生成每个新样本所对应的类别分布样本，即记载着每个样本所属各个类别的概率的列表。在最后一个环节中，将所有新样本的类别分布向量列拼接起来，作为该原始样本的

重新表示。对于图 2 中的例子，在上方图例中最终会将 400 维的原始数据重新扫描转化为 1806 维的特征向量，而对于下方图例最终转化为新的 726 维的特征向量。

2、级联森林

为了实现深度学习模型中的逐层处理的特性，gcForest 提出了一个级联结构。每层接受上一层的输出信息作为输入信息。级联森林结构的每层是决策树森林的集成，即集成之后的集成，如图 3 所示。

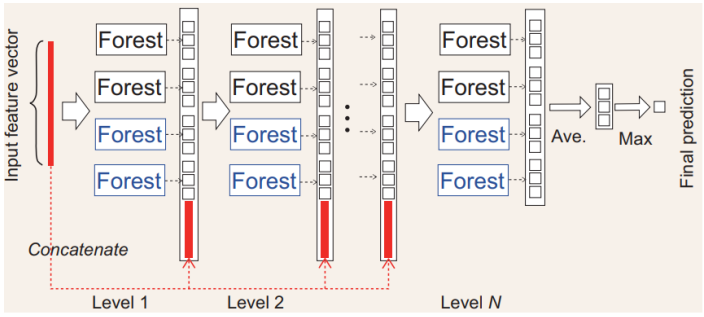

图 3 gcForest 的级联森林结构示意图[5]

如图所示，在级联森林结构中使用多种不同的森林结构来增强模型的复杂度——两个随机森林和两个完全随机森林。其中随机森林中每颗决策树在结点划分时随机选择$\sqrt{d}$维的属性作为候选属性，其中d是总的特征属性个数；完全随机森林的每颗决策树在结点划分时随机选择一个属性特征，即随机选择划分点进行划分。

对于给定的样本，每个森林会生成该样本分布的一个估计。如图 3 所示，样本最终会落到每棵树的一个叶子节点，将该叶节点中所包含的训练样本的各个类别的比重作为这个树所属类别的估计，最终将所有树的类别分布估计的平均值作为该森林生成的类别分布向量。在这个级联结构中，每一层级所包含的森林生成的类别分布向量会和原始的特征向量拼接在一起，作为下一层的输入。为了减少过拟合（over-fitting）的风险，每一个森林均采用 k-折交叉验证的方式生成类别分布向量，直到验证集上显示模型性能不再有显著的提升时，结束整个训练过程。图 4 展示了 gcForest 模型的整体结构。

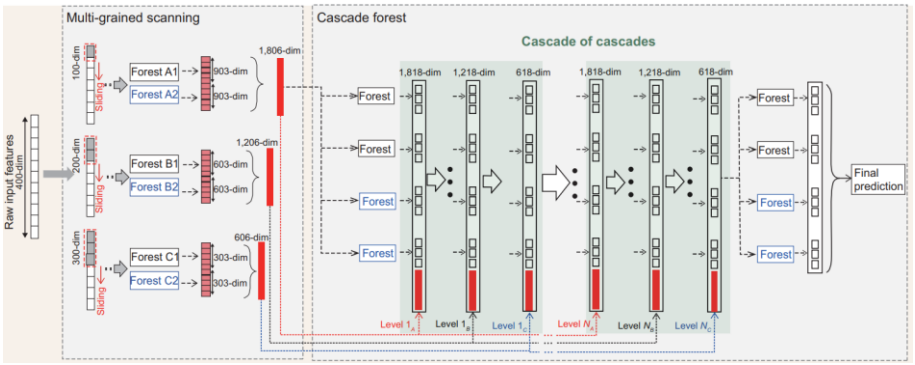

图 4 gcForest 的整体过程示意图[5]

我们假定原始数据的特征属性维度为 400 维，通过多粒度扫描模块，每一个样本会被首先扫描生成一个新的样本集合，将新样本所得类别分布向量拼接。利用三种不同大小的滑动模块，最后分别生成 1806 维、

1206 维和 606 维的新的特征转换。最终的模型可以看作是一个级联的级联，将这三组新的特征转换结合上一层的输出特征分别作为第t层级联中的t_A、t_B和t_C进行输入。如果当验证集上显示性能不再有明显的提升后，层数不再增长，此时终止训练过程。

在训练结束后，对于给定的测试样本，多粒度扫描生成相应的特征转换表示，作为后续层级级联的输入。最终类别分布向量为级联最终层的输出的平均值，而将类别分布向量中样本所属各类别概率最大的类别做为最终的类别预测。

2.2 基于 KNN 改进的深度森林方法

在实现了深度森林算法后，在老师启发下，尝试利用 KNN 算法替换级联森林模块中的随机森林以及完全随机森林算法，下面对 KNN 算法进行介绍以及给出改进的深度森林算法的流程图。

K 近邻算法（K-nearest neight, KNN）是一种基于实例的学习算法，其特点是不依赖于参数的学习算法。基于实例学习的模型以记忆训练集为特征，是一种懒惰的学习模型，其算法模型的特点是在于它不是从训练数据中学习判别函数，而是靠记忆训练过的数据集来完成分类任务。

KNN 算法的的工作机制非常简单：对于给定的测试样本，基于某种距离度量找出训练集中与其最靠近的 k 个训练样本，然后基于这 k 个“邻居”的信息来进行预测。通常，在分类任务中可以使用“投票法”，即选择这 k 个样本中出现最多的类别标记作为预测结果[15]。通过“投票法”我们可以意识到在深度森林中的随机森林模块也是通过决策树输出的分类结果后利用“投票法”进行推选出最后的结果，KNN 算法的原理如图 5 所示。

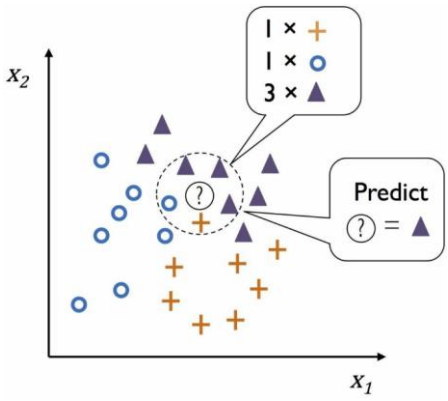

图 5 KNN 算法示意图[15]

以图 4 为例，当我们设置 k=5 时，会选取预测样本周围距离最近的 5 个样本，选取出现类别最多的样本类别作为最后的预测结果，如图所示，三角类别的样本共有 3 个，因此最后预测样本的类别同为三角。我们的想法是对于深度森林的多粒度扫描模块我们不做更改，仍然保持利用随机森林和完全随机森林对输入的样本进行特征转换。对于级联森林模块，我们利用 KNN 算法来替换级联森林中的随机森林和完全随机森林。对于多粒度扫描生成相应的特征转换表示，作为后续层级级联的输入。最终类别分布向量为级联最终层的输出的平均值，而将类别分布向量中样本所属各类别概率最大的类别做为最终的类别预测。值得注意的是，为了增加模型的复杂度，我们将设置跨度较大的 k 值来保持结果的多样性，适当的增加层数进而保证深度学习

模型具有足够的复杂度，下面给出算法步骤：

步骤 1 设置模型的参数，包括多粒度扫描模块的两种随机森林的子树的数量 n_estimators，节点划分最小不纯度 min_samples_split， KNN 分类器的邻居数量 n_neighbors，度量距离的方式 metric 等；

步骤 2 设置不同的滑动窗口的大小d_1，d_2和d_3，通过步骤 1 中设置的随机森林的参数来扫描样本，生成相应的特征转换；

步骤 3 利用步骤 1 设置的 KNN 算法的参数，将新生成的样本的特征转换做为新的样本分布，通过多层的 KNN 的级联做相应的特征处理，最后输出带有样本预测概率的列表，将概率最大的类别作为最后的预测结果。

2.3 本章小结

在第二章中主要对深度森林算法的细节以及构成算法的两个模块——多粒度扫描模块以及级联森林模块进行了介绍。第二节中对 KNN 算法进行了描述，以及企望将经典深度森林算法中级联森林结构中的随机森林算法替换为 KNN 算法，并给出了新提出算法的流程图，以便为后续进行测井岩性识别分类问题的实验提供理论基础。

3 测井岩性识别实验

本章将深度森林算法在 python3.8.0 上运行，并利用 sk-learn 库将测试实例的 80%划分为训练样本，剩余 20%划分为测试样本，利用提出的算法进行分类效果测试；通过调整参数测试算法的不同效果，并与经典的分类算法进行比较。

3.1 算法结果对比

首先对数据集进行介绍，实验所应用的数据集是老师提供的测井岩性识别数据。数据共分为七个维度，分别为：GR、AC、DEN、CNL、LLD、LLS 和 CAL。一共有 357 个样本可供模型训练，模型分为三类，分别为：C、M 和 PS。表 1 展示了其部分数据。

表 1 测井岩性识别部分数据

GR	AC	DEN	CNL	LLD	LLS	CAL	Type
154.069	260.635	2.579	23.774	61.011	49.593	24.356	C
157.943	276.435	2.514	25.108	62.09	49.673	24.389	C
155.394	293.842	2.397	27.541	71.385	55.793	24.404	C
145.404	304.742	2.241	32.11	95.975	74.62	24.491	C
133.034	304.978	2.092	38.263	136.621	111.535	24.524	C
124.25	294.219	2.017	39.696	177.732	159.079	24.544	C

在进行实验时，将测试集进行拆分——80%作为训练集，其余 20%作为测试集，并设置随机数种子以保证每次划分的样本集合不同。本论文中对分类算法的效果采用的评价指标有分类准确率和 F1 score，其中对于准确率有：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

其中，TP 代表样本被判定为正样本，实际为正样本；TN 代表样本被判定为负样本，实际为负样本；FP 代表样本被判定为正样本，实际为负样本；FN 代表样本被判定为负样本，实际为正样本。

对于 F1 score：

$$F_1 = 2\frac{precision \times recall}{precision+recall} \tag{2}$$

其中，precision 为查准率，可以用$precision = \frac{TP}{TP+FP}$来定义，而 recall 为召回率，用$recall = \frac{TP}{TP+FN}$定义。

现利用自实现的深度森林算法、官方的深度森林包以及基于 KNN 算法的深度森林算法对测井岩性识别数据进行独立的 20 次，计算其模型分类结果的准确率以及 F1 score，结果如图 6 所示。

通过图片我们可以看出，对于三种算法的分类准确率和 F1 score 的变化趋势是相同的。对于基于深度森林算法原理而完整自实现的两个算法在整体性能上是劣于官方给出的深度森林算法的，利用官方给出的深度森林算法分类准确率可以达到 93%，而自实现的两种算法基本分类准确率可以达到 85%左右。

另外，通过实验结果可以发现，当级联森林模块中的子模块是随机森林时，算法拥有较高的稳定性，在 20 次实验过程中，两个深度森林算法进行分类时只有 2 次不同的预测结果，即算法对样本具有较高的鲁棒

性。在级联模块替换为 KNN 算法后，在实验过程中分类结果具有较大的波动性。因此，通过分析我们可以知道在级联多个基本分类模型来构建深度学习模型时，利用随机森林和完全随机森林比利用多个不同参数的 KNN 算法来增加模型的复杂度更稳定，即随机森林算法相较于 KNN 算法在多层级联后具有较高的稳定性。

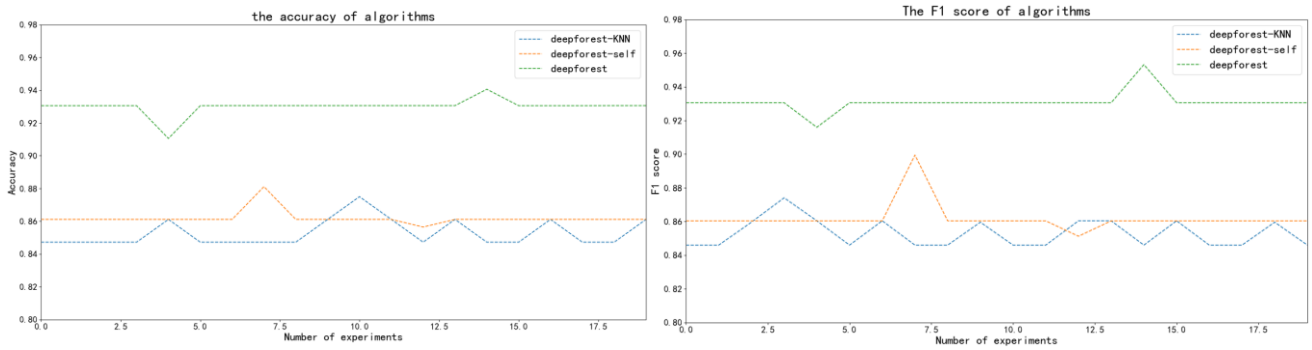

图 6 三种算法的分类准确率（左）和 F1 score（右）

下面我们通过绘制算法分类结果的决策边界图来观察三种算法的分类效果，由于测井岩性识别数据是 7 维数据，不易用图片的方式直观表示，因此在此选取其 2 个维度的数据进行实验，所选取的维度分别是 DEN 和 CNL。图 7，图 8 和图 9 分别展示了官方提供的深度森林方法、自实现的深度森林算法和利用 KNN 算法替换之后的深度森林算法的分类结果决策边界图。

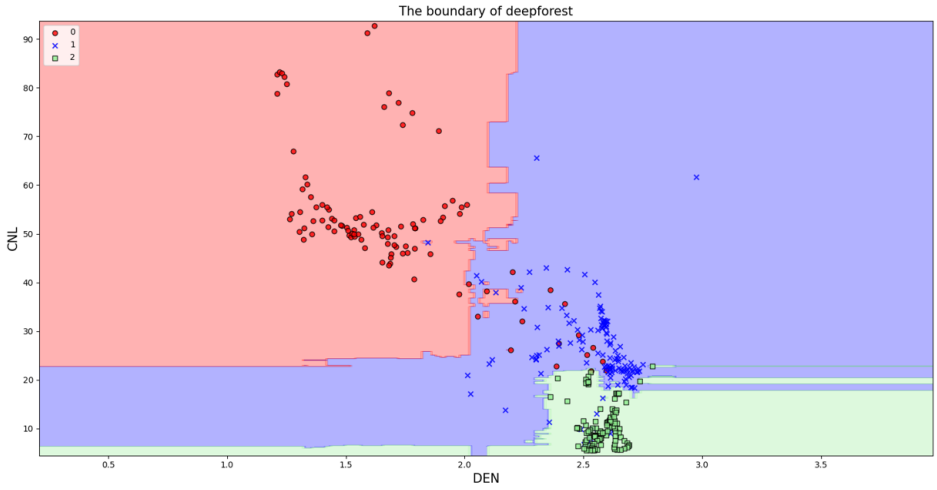

图 7 gcForest 算法分类边界

通过对比图片我们可以直观的看到三种算法的分类结果，通过分类边界我们可以直观的看到分类算法在处理二维数据的效果。我们同样可以验证官方提供的算法库的分类效果是最好的，对于图中的绿色类和红色类的分类边界均较准确。且运算结果也较为稳定，分类边界出现较多的锯齿也是因为在级联结构中的决策树算法是利用二叉树将样本进行分类。而对于自实现的深度森林算法其分类效果相比较而言较差，但对于样本分类而言分类结果基本准确，但对于分类边界而言出现了一些不合理的地方，比如数据分布的右侧在蓝色区域中间又出现了一道红色区域。整体而言自实现的算法的效果是没有官方提供的算法效果好的。而对于利用

KNN 算法构成级联结构的算法在稳定性以及分类准确度方面都是较差的。由于算法在执行过程中是根据不同参数的 KNN 算法来增加模型的复杂度的，因此对于距离较远的样本很有可能因为“投票法”的判定原因而导致最后的误分类。而如何选取较为有效的多个 KNN 模型来提高整体算法的分类效果，将在第二节中进行详细的实验。

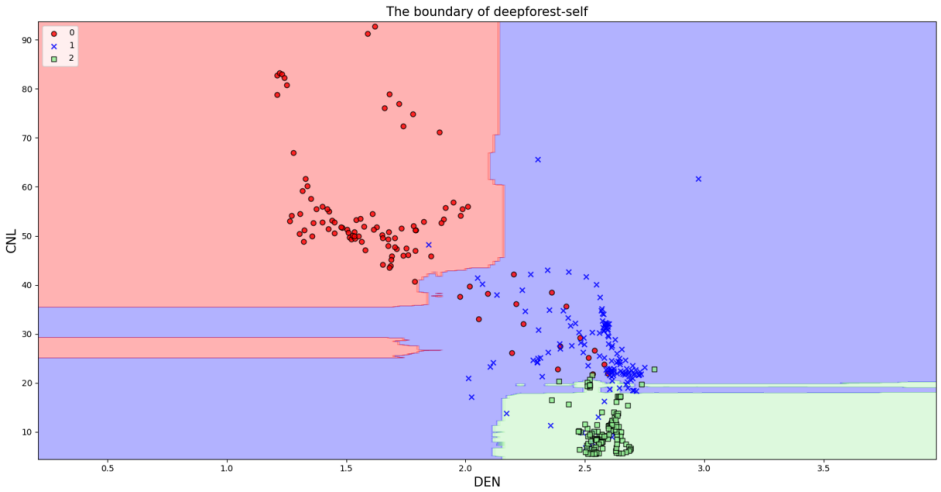

图 8 自实现 gcForest 算法分类边界

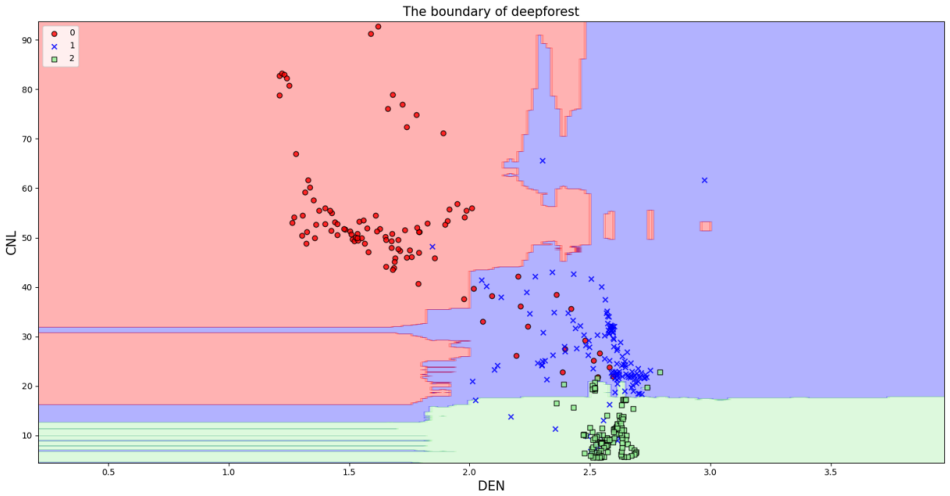

图 9 利用 KNN 算法替换后的 gcForest 算法分类边界

3.2 算法参数分析

在本小节中我们所实验的重点问题是模型参数对于深度学习模型的分类效果的影响，由于官方提供的方法模型中有很多参数都是提前设定好的内嵌在模型里的，因此在本小节中我们仅讨论通过自己实现的深度森林算法以及利用 KNN 算法构成级联结构算法中不同参数对于模型分类结果的影响。表 2 记录了本小节中需要调整的参数。

表 2 模型参数

	多粒度扫描
共同模块	滑动窗口大小

算法	级联森林
自实现深度森林	n_estimators
KNN 算法级联	n_neighbors

通过表 2 我们可以看出本小节针对算法参数分析共涉及 3 个参数，其分别为：多粒度扫描模块的滑动窗口大小、级联森林模块中深度森林算法两种随机森林的子树的数量 n_estimators，以及 KNN 分类器的邻居数量 n_neighbors。其实每个子算法中有很多可以设置的参数，比如随机森林中刻画结点不纯度的方式，KNN 算法中刻画距离的方式，但以上 3 个参数是比较重要的参数，因此在此只针对以上提及的 3 个参数研究其对算法分类效果的影响。

1、滑动窗口大小

由于两种算法对于多粒度扫描模块的部分是没有进行更改的，因此对于这部分的参数调整对两个算法的影响是共同的。在 2017 年 Zhou 团队在论文中提到共利用 3 种滑动窗口，设训练数据是 d 维向量，则利用：$\frac{d}{4},\frac{d}{9}$和$\frac{d}{16}$四种滑动窗口大小。现针对不同的滑动窗口进行实验。

实验过程如下，设置多粒度扫描模块中滑动窗口大小分别为：$d,\frac{d}{4},\frac{d}{9},\frac{d}{16}$以及 3 种滑动窗口共同搭配对两种算法进行独立的 20 次实验，以算法分类结果的准确率作为最后的评判标准。图 10 分别代表自实现深度森林算法以及利用 KNN 算法级联算法不同滑动窗口 20 次实验结果：

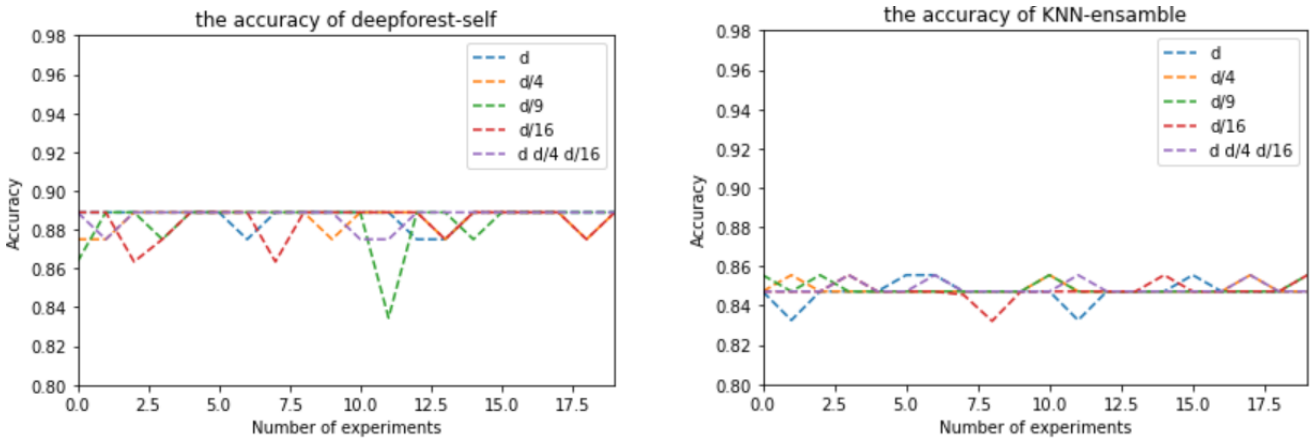

图 10 滑动窗口大小对自实现深度森林算法（左）和 KNN 级联算法（右）的影响

通过图片我们可以看出不同滑动窗口对于算法的实验结果的影响不是很大，并且结果没有很大的变化，在 20 次独立的实验中，5 种不同的滑动窗口的分类准确率的值大多都趋于一致。除去少数特殊数据外可以认为对于实验数据而言忽略滑动窗口带来的影响。分析其原因，我认为是**数据量过少**导致的，测井岩性识别数据仅有 285 个数据，而一般的深度学习所用的训练集大多是上万或上百万的数据集。在对于很大数量的数据

集时，利用滑动窗口的多样性来增加模型的复杂度以及训练样本的多样性是必要的，而对于仅有 285 个数据的岩性识别数据，其影响是很小的。

但通过对两个算法的实验结果图可以看出，选取多个不同的滑动窗口在 20 次实验过程中产生的波动数据是最少的，通过以上分析，在以后的实验中均采用滑动窗口大小为$\frac{d}{4},\frac{d}{9}$和$\frac{d}{16}$三种窗口组合的方式进行多力多扫描模块。

2、n_estimators

在深度森林的级联森林模块中共设置了 4 棵 2 种不同的随机森林来构成级联结构——2 棵随机森林算法以及 2 棵完全随机森林算法。而随机森林是一种 boosting 集成算法，其是由多个决策树模型级联而成的，在这其中，每颗随机森林由多少颗子树构成就是一个非常重要的参数，现在我们对这个参数进行讨论。

首先我们探究随机森林和完全随机森林构成的子树个数不同时是否对算法的效果有影响，图 11 展示了当随机森林的子树个数为 1000 而完全随机森林的子树个数为 1500 时与随机森林与完全随机森林都是由 1000 棵子树构成的算法运行 20 次的结果准确率。

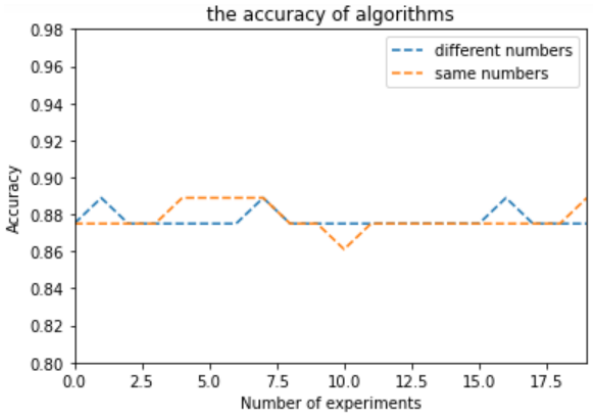

图 11 两类随机森林子树数目对算法分类准确率的影响

通过图片我们可以看出，当随机森林和完全随机森林的子树的个数是否相同对分类准确率的影响不大，因此正如论文中所设定的在后续实验中，设置两类随机森林的子树个数相同。后续中，为了探究合适的子树个数，我们分别设定子树个数分别为 100、500、1000 和 2000。图 12 展示了深度森林分类结果的准确率重复实验 20 次的结果。

通过图片我们可以看出随着子树数目的增加，深度森林算法的准确率是逐渐提升的，当子树数量在 2000 棵时，20 次实验的效果要好于子树数量较少时的情形。在不考虑实验运行时间的前提下，我们认为子树的数量较多时算法的效果更好。

3、n_neighbors

在处理多粒度扫描得到的新的数据的特征数据时，将累加的数据集分别输入到拥有不同参数的 KNN 子算法中。本文在利用多个 KNN 算法进行级联而构成深度学习模块时，设置了不同参数的 KNN 算法以增加模型的复杂度。因 KNN 算法本身相较于随机森林而言处理问题能力相对有限，而随机森林又是多个决策树集成的算法，利用随机森林进行级联本身就有相当的模型复杂度。而 KNN 算法的一个重要的参数就是在确定每个样

本所属类别的“邻居”个数，因此我们现在对这个参数进行讨论。

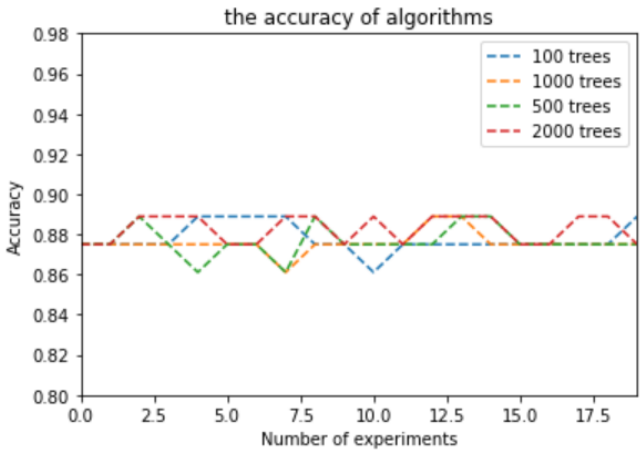

图 12 随机森林子树数目对算法分类准确率的影响

3、n_neighbors

在处理多粒度扫描得到的新的数据的特征数据时，将累加的数据集分别输入到拥有不同参数的 KNN 子算法中。本文在利用多个 KNN 算法进行级联而构成深度学习模块时，设置了不同参数的 KNN 算法以增加模型的复杂度。因 KNN 算法本身相较于随机森林而言处理问题能力相对有限，而随机森林又是多个决策树集成的算法，利用随机森林进行级联本身就有相当的模型复杂度。而 KNN 算法的一个重要的参数就是在确定每个样本所属类别的“邻居”个数，因此我们现在对这个参数进行讨论。

实验过程如下，我们设置以下多个不同参数进行实验：4 层具有相同参数的 KNN 算法；有 2 个参数是相等的 KNN 算法；4 个参数相差不大的 KNN 算法；参数有大有小的 KNN 算法。其中，4 层具有相同参数的 KNN 算法又分为参数较大（k=20）和参数较小（k=3），2 个参数相同的算法即 2 个 KNN 算法的 k=3 其余两个 KNN 算法的 k=20。在实验过程中我们除了邻居个数外，其余参数均相同，滑动窗口大小按照前述讨论结果设定，独立运行 20 次结果如图 13 所示。

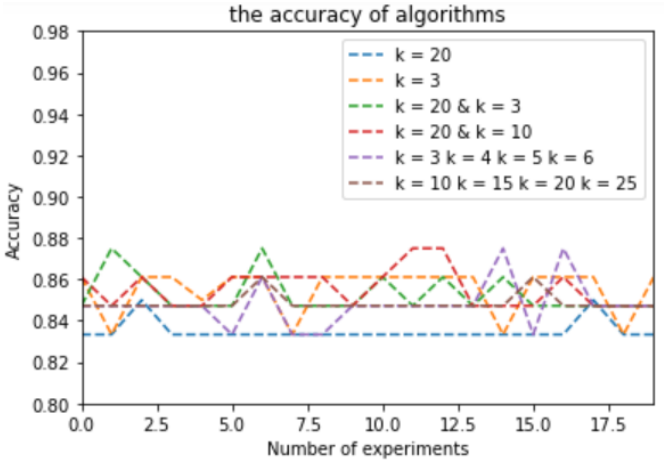

图 13 参数对 KNN 级联算法的影响

通过图片我们可以看出，分类效果最差的是利用 4 个 k 值都很大的 KNN 算法进行级联，其准确率在所

有的实验结果中是最低的。而同样利用 4 个 k 值都很小的 KNN 算法级联时，虽然效果有所提升，但其波动范围较大。因此，如果均使用参数相同的 KNN 进行级联，其效果是较差的，也正如文章中所说的，一个有效的深度学习模型是需要足够的模型复杂度的，而单参数的模型进行级联模型复杂度是不够的。

而对于 4 个 KNN 算法模型参数都不同进行级联而构建的级联模型，其效果也没有 2 种不同的 KNN 算法进行级联好。通过图片可以看出，在所有的实验结果中，效果最好的是 2 个 k=20 的 KNN 算法和 2 个 k=10 的 KNN 算法进行级联。因为测井岩性识别数据数据量不大，因此在所有实验中的样本准确率的变化不大。对于 20 次独立实验，我认为通过多个 KNN 模型进行级联的深度学习模型在处理这个测井岩性识别分类数据时利用两组参数不同的模型效果是最好的。

3.3 本章小结

在第 3 章中，我们利用本文中提及的算法进行了实验。首先利用提出的 2 种算法与官方给出的算法包对测井岩性识别数据进行实验，通过分类准确率与 F1 score 两种刻画方式，接着选取数据集中的 2 个维度绘制算法的分类边界，比较了算法对于测井岩性识别数据的分类能力。在第二节中，我们讨论了算法参数对于算法分类效果的影响。分别讨论了多粒度扫描模块中滑动窗口的大小、随机森林算法中子树的数量和 KNN 算法中邻居的个数 3 种参数对于深度学习模型效果的影响。

4 讨论

在本章中我们讨论本论文所做的工作对应于老师布置作业中的哪些条例要求，以便于在后面的章节对于论文进行总体的总结与探究。老师在课程结束时共布置了 4 个作业，我选择的是大作业题目 3，题目是“利用深度森林方法建立测井岩性识别模型”。现在第 4 章针对论文内容进行梳理，并与布置作业的题目相对应进行结课论文的讨论。

课程作业的第一个方面是“深度森林算法应包含多粒度扫描和级联森林”，并在后续提出了其他的可行方案，比如只实现级联森林算法但将其中的随机森林算法替换为其他分类算法，或者将随机森林中的决策树算法替换为 KNN 算法。在论文的第 2 章，我们重点阐述了 Zhou 等人在 2017 年提出的 gcForest 模型的细节，包括多粒度扫描模块以及级联森林模块实现的细节。我们按照论文中的实现细节以及步骤自实现了深度森林模型，即包含了目前官方开源的代码中没有的多粒度扫描模块。但算法的整体性能没有官方提供的算法效果要好，由于时间与能力有限，在老师的启发下，我将级联森林中的随机森林算法替换为了 KNN 算法。即利用 KNN 算法进行级联构成深度学习模块，并在后续的章节进行算法性能的比较。

课程作业的第二个方面是“分析算法的预测能力，并总结较优的参数”。在论文的第 3 章中，论文的重点工作是算法的实验。在第一小节，我们利用自实现的深度森林算法与利用 KNN 算法级联构成深度学习算法进行测井岩性识别数据的分类。利用算法的准确率以及 F1 score 来衡量算法的性能，利用官方提供的深度森林包进行对照实验，得到深度森林算法的分类效果较利用 KNN 算法级联要好，稳定性也较强。在第二节中，我们的重点工作是测试参数对于算法效果的影响。分别测试了多粒度扫描模块的滑动窗口大小、级联森林模块中深度森林算法两种随机森林的子树的数量 n_estimators，以及 KNN 分类器的邻居数量 n_neighbors。对算法的影响，并通过实验最终确定了对于测井岩性识别数据分类效果最好的参数。

5 结论

我国油气资源对外依存度非常高，致密油气储层非常重要的渗流通道之一就是裂缝，单井裂缝的测井解释对分析裂缝期次、厘清裂缝纵横向发育规律、提高三维裂缝网络建模精度等均具有重要意义。

论文应用深度森林算法以及利用 KNN 算法进行级联构成深度学习模型对测井岩性识别模型进行实验，将实验结果与官方提供的深度森林算法进行对比。并后续针对算法模型进行了参数分析。

我们通过学习 Zhou 等人在 2017 年提出的 gcForest 算法，按照其模型的步骤自实现了深度森林算法。并通过老师的启发，将随机森林算法替换为 KNN 算法，利用多个不同参数的 KNN 算法进行级联而构成深度学习模型。基于所用实例测试的实验结果表明，我们的算法提供了良好的结果，为利用深度学习模型进行测井岩性识别分类问题提供了一种新的思路。

本研究的重点内容分为以下两个方面：

（一）自实现了 gcForest 算法。深度森林算法在官方提供的 Github 等平台上提供的源代码中，并没有给出多粒度扫描模块的代码，而按照完整流程实现代码是针对学习算法思想以及后续对算法过程的改良是非常重要的。在充分的查阅资料与学习后，本研究完整的实现了多粒度扫描模块以及级联森林模块，在后续实验中取得了良好的测试效果。

（二）新的级联方式。在老师所布置题目的启发下，本研究尝试了利用多个不同算法的 KNN 算法进行级联而构成深度学习模型。利用新的级联方式应用到测井岩性识别数据中，针对分类结果进行了分析与研究。为后续利用深度学习模型进行测井岩性识别模型数据进行分类提供了新的思路。

在测试过程中仍然存在以下一些不足，希望在今后可以改进而对这种结合方式进行更好的优化：

（一）实验数据量较小。在一半的深度学习模型中，所进行训练的数据量往往是较大的，比如人脸识别或者 NLP 问题的模型训练中，其训练量达到几十 G。而测井岩性识别数据只有 285 个样本，又将其中 20%的数据用于测试，因此在实验中不能全面的对算法的性能进行评估。期望在后续研究中引入更大的数据量进行训练，来分析算法的分类效果以及参数对算法的影响。

（二）算法改进不完善。虽然完成了自实现完整的深度森林模型，但为了进行更全面的研究，在老师所布置的题目中期望将随机森林中的决策树模型替换为 KNN 算法，由于时间关系，直接将随机森林算法替换为了 KNN，其最后的结果效果不理想。期望在后续的研究中实现更细致的改进，期望得到更好的效果。

参考文献

[1] 孙龙德,邹才能,贾爱林,位云生,朱如凯,吴松涛,郭智.中国致密油气发展特征与方向[J].石油勘探与开发,2019,46(06):1015-1026.

[2] 董少群, 曾联波, 车小花, 等. 人工智能在致密储层裂缝测井识别中的应用[J]. 地球科学, 2022: 1-23.

[3] 董少群，曾联波*，Chaoshui Xu，等. 储层裂缝随机建模方法研究进展[J]. 石油地球物理勘探. 2018, 53(3): 625-641.

[4] 董少群，吕文雅，夏东领，等. 致密砂岩储层多尺度裂缝三维地质建模方法[J].石油与天然气地质, 2020, 41(3): 627-637.

[5] Z-H. Zhou and J. Feng. Deep forest: Toward an alternative to deep neural networks. In *Proceeding of the 26th International Joint Conference on Artificial Intelligence*, pp. 3553-3559, 2017.

[6] 庞明. 新型深度森林模型的研究[D].南京大学,2020.DOI:10.27235/d.cnki.gnjiu.2020.002301.

[7] Shi,G.R.,2008. Superiorities of Support Vector Machine in Fracture Prediction and Gassiness Evaluation. *Petroleum Exploration and Development*, 35(5):588-594. Doi:10.1016/S1876-3804(09)60091-4.

[8] 郑军，刘鸿博，周文，等，2010.阿曼五区块Daleel油田储层裂缝识别方法研究.测井技术，34(3):251-256.

[9] 何胡军，毕建霞，曾大乾，等，2014.基于测井曲线斜率的KNN分类算法常规测井裂缝识别——以普光气田礁滩相储层为例.中外能源，19(1):70-74.

[10] Bhattacharya,S.,Mishra,S.,2018.Applications of Machine Learning for Facies and Fracture Prediction Using Bayesian Network Theory and Random Forest: Case Studies from the Appalachian Basin, USA. Journal of Petroleum Science and Engineering, 170:1005-1017. Doi:10.1016/j.petrol.2018.06.075.

[11] Azizi,H.,Reza,H.,2021.Applied Machine Learning Methods for Detecting Fractured Zones by Using Petrophysical Logs. *Intelligent Control and Automation*, 12(2):44-64. Doi:10.4236/ica.2021.12203.

[12] Dong,S.Q.,Zeng,L.B.,Liu,J.J., et al,2020c.Fracture Identification in Tight Reservoirs by Multiple Kernel Fisher Discriminant Analysis Using Conventional Logs. *Interpretation*, 8(4):P215-P225. Doi:10.1190/int-2020-0048.1.

[13] Nouri-Taleghani,M.,Mahmoudifar,M.,Shokrollahi,A., et al.,2015. Fracture Density Determination using A Novel Hybrid Computational Scheme: A Case Study on An Iranian Marun Oil Field Reservoir. *Journal of Geophysics and Engineering*, 12(2):188-198. Doi:10.1088/1742-2132/12/2/188.

[14] Tian,M.,Li,B.T.,Xu,H., et al., Deep Learning assisted Well Log Inversion for Fracture Identification. Geophysical Prospecting, 69(2):419-433. Doi:10.1111/1365-2478.13054.

[15] Sebastian Raschka, Vahid Mirjalili , et al. Python Machine Learning – Second Edition. EXPERT INSIGHT.