第九届“泰迪杯”数据挖掘挑战赛——

B 题：岩石样本的智能识别

一、背景

在油气勘探中，岩石样本识别是一项即基础又重要的环节；在矿产资源勘探中，尤其是在固体金属矿产资源勘探中，岩石样本识别同样发挥着不可估量的作用；岩石样本的识别与分类对于地质分析极为重要。目前岩石样本识别的方法主要有重磁、测井、地震、遥感、电磁、地球化学、手标本及薄片分析方法等方法，而采用图像深度学习的方法建立岩石样本自动识别分类模型是一条新的途径。

现有样本数据是采用工业相机在录井现场对岩屑和岩心样品进行拍照，分别在暗箱内拍摄白光和荧光两种相片，如图 1 和图 2 所示。白光灯下拍摄的相片是用于提取颜色、纹理、粒度等特征识别岩性，荧光灯下拍摄的相片是用于识别含油气性（石油在紫外线照射下具有的发光特征，其中的绿色和黄色部分是含油的，见图 2）。

图 1 白光灯下拍摄的图片

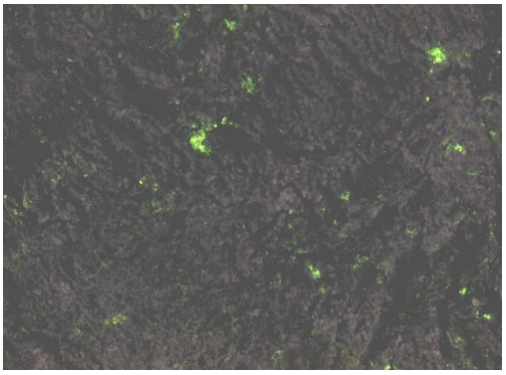

图 2 荧光灯下拍摄的相片

本赛题期待参赛者能够通过图像处理技术和深度学习算法，设计出有效的模型识别出岩石样本的岩性类别及含油气情况，实现岩石样本智能识别分类。

二、要解决的问题

数据集 rock 中，“白光/荧光”标签为“1”的数据是相同白光环境下拍摄的岩石样本图像数据，“白光/荧光”标签为“2”的数据是相同荧光环境下拍摄的岩石样本图像数据，请利用数据集 rock 的数据完成以下问题。

1. 构建岩石样本岩性智能识别模型。请设计合适的机器学习或深度学习算法，针对数据集 rock 实现岩石样本岩性智能识别与分类。

2. 计算岩石含油面积百分含量。石油在紫外线照射下具有发光特征，即荧光灯下拍摄的相片中绿色或黄色部分是含油的，请设计合适的算法计算岩石的含油面积百分含量（绿色和黄色部分的面积占总岩石面积的百分比）。

三、数据说明

本赛题的数据总共由两部分组成。

(1) 数据集 rock：包含白光和荧光下拍摄的图像数据，图像名称为“X-Y.bmp”或“X-Y.jpg”。X 表示样本编号，编号相同表示为同一个岩石样本拍摄的图像，其中编号在 1～321 之间的为“.bmp”格式的图像数据，编号在 322～350 之间的为“.jpg”格式的图像数据（见图 3）；Y 表示拍摄图像时的光照环境（1 表示白光，2 表示荧光）。

(2) 标签表 rock_label.csv：此表格的内容为数据集 rock 中每个样本的岩性类别（见图 4）。

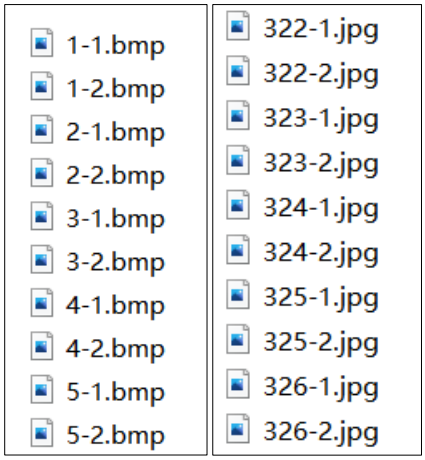

图 3 部分岩石样本图像数据集展示

样本编号	样本类别
13	深灰色泥岩
14	灰黑色泥岩
15	深灰色泥岩
16	浅灰色细砂岩
17	灰色泥质粉砂岩
18	灰黑色泥岩
19	浅灰色细砂岩
20	深灰色泥岩
21	灰黑色泥岩
22	灰色泥质粉砂岩
23	灰色细砂岩

图 4 部分样本标签表展示