# 《生物大数据分析》大作业要求及题目

## 要求：

1、以下共有七个题目，33位同学需分为七组，人数分别为5人，5人，5人，5人，5人，4人，4人，每组选定一个题目完成。不可有两组选择同一题目。

2、提交内容包括：代码及readme文档、报告、PPT、PPT汇报录屏。

- 代码及readme文档：需提交项目源代码，将各文件以标准易懂的方式进行命名，并在readme中注明每个文件的功能及调用方法，对环境进行准确的描述，以便进行结果复现。以上文件按逻辑存放在以codes命名的总文件夹下。
- 报告：包含摘要、背景、整体逻辑框架、（描述性统计）、方法、结果、结论和讨论等部分。提交格式为pdf，命名格式为题目编号-报告的title，如：1-使用**方法实现甲状腺分割任务.pdf。
- PPT：符合基本项目汇报要求，最后一页写清每个人的分工。命名格式同上，如：1-使用**方法实现甲状腺分割任务.ppt。
- PPT汇报录屏：使用腾讯会议等录制PPT汇报过程，每位组员都需进行讲解，讲的同时需打开摄像头（录入本人），总时长不超过15分钟。提交MP4格式文件，命名格式同上，如：1-使用**方法实现甲状腺分割任务.mp4。

  将以上四个同级文件（夹）进行压缩打包，压缩包命名格式同上，如：1-使用**方法实现甲状腺分割任务.rar。将压缩包上传至canvas。注意：不符合以上各项要求会酌情扣分。

## 项目题目：

对题目有任何不理解请及时与助教沟通。

### 1、以代谢当量表征的儿童运动强度分类探究

**MET**是Metabolic Equivalent的缩写，中文翻译过来是“代谢当量”。1 MET也被定义为每公斤体重每分钟消耗3.5毫升氧气，大概相当于一个人在安静状态下坐着，没有任何活动时，每分钟氧气消耗量。一个5 METs的活动表示运动时氧气的消耗量是安静状态时的5倍。MET是用于表示各种活动的相对能量代谢水平，也是除了心率和自觉运动强度以外的另一种表示**运动强度**的方法。

本项目将从网络获取并处理过的儿童运动视频（数据请勿外传）按照MET等级分为light、moderate、vigorous三类，分别有519、541、536段视频。

| | | |
|---|---|---|
| **文件大小: 880.9 MB** | **分享内容: 3_class_dataset.rar 链接地址:https://jbox.sjtu.edu.cn/l/k1RqHy** | |
| 到期日: 2022-05-27 | | |
| 来自于: 高强 | | |

**要求：**

设置随机种子，随机划分3:1的训练集和验证集，使用合适的深度学习方法对儿童运动视频进行分类。

## 2、基于胸部X-ray成像的肺炎诊断

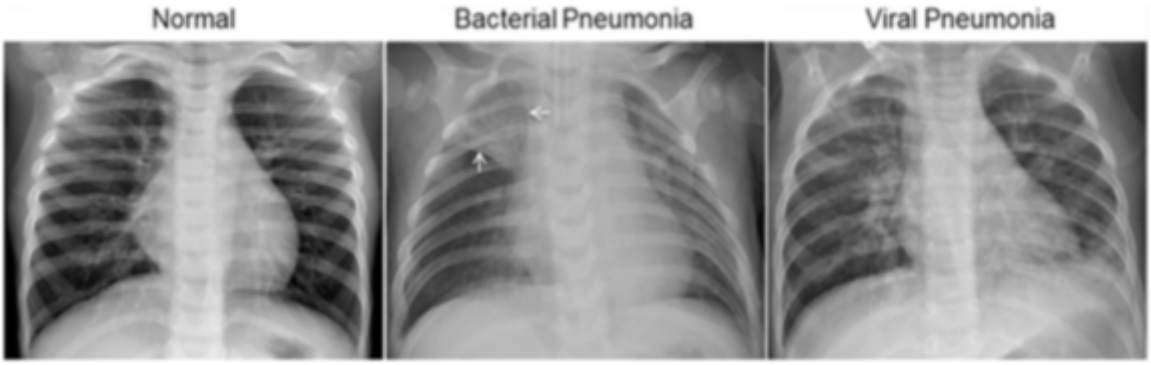


图1、肺炎患者胸部X线检查的说明性实例

正常的胸部X射线（左）描绘了清晰的肺部，图像中没有任何异常混浊的区域。细菌性肺炎（中）通常表现为局灶性肺叶实变，在这种情况下在右上叶（白色箭头）中，而病毒性肺炎（右）在两个肺中表现为更弥漫的"间质"模式。

数据集分为 3 个文件夹（train、val、test），并包含每个图像类别（肺炎/正常）的子文件夹。有5,863张X射线图像（JPEG）和2个类别（肺炎/正常）。

| | | |
|---|---|---|
| **文件大小: 2.3 GB** | **分享内容: 第2题数据.zip 链接地址:https://jbox.sjtu.edu.cn/l/O1SnHe** | |
| 到期日: 2022-05-27 | | |
| 来自于: 高强 | | |

**要求：**

使用训练集进行训练，验证集验证，并用测试集进行外部验证。使用课上的至少三类（CNN、RNN、GAN、GNN）深度学习方法。挖掘计算机分类的关注点，给出一定的生物学解释。

## 3、乳腺癌数据挖掘

数据集说明如下：

| | | |
|---|---|---|
| **文件大小: 48.6 KB** | **分享内容: 第3题数据.zip 链接地址:https://jbox.sjtu.edu.cn/l/V1yECI** | |
| 到期日: 2022-05-27 | | |
| 来自于: 高强 | | |

Attribute Information:

(1) ID number
(2) Diagnosis (M = malignant, B = benign)
(3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

**要求：**

使用机器学习、深度学习算法，自我挖掘至少5个有价值的问题进行分析探索。

## 4、甲状腺超声图像结节识别定位

该甲状腺超声图像数据库包含400个病例和470张图像。每个病例都以XML文件的形式呈现，包含对应图像的结节位置信息。

| | | |
|---|---|---|
| **文件大小: 17.2 MB** | **分享内容: 第4题数据.zip 链接地址:https://jbox.sjtu.edu.cn/l/01qq3O** | |
| 到期日: 2022-05-27 | | |
| 来自于: 高强 | | |

**要求：**

将甲状腺结节分割点转化为bounding box（矩形）。设置随机种子，随机划分3:1的训练集和验证集，对其甲状腺结节进行识别定位任务，最终需给出bounding box坐标。

## 5、表情识别

表情识别是指从静态照片或视频序列中选择出表情状态，从而确定对人物的情绪与心理变化。

该数据集包含 35685 个 48x48 像素灰度图像的样本，分为训练数据集和测试数据集。图像根据面部表情(happiness, neutral, sadness, anger, surprise, disgust, fear)分为七类。

| | | |
|---|---|---|
| **文件大小: 65.2 MB** | **分享内容: 第5题数据.zip 链接地址:https://jbox.sjtu.edu.cn/l/01qQtD** | |
| 到期日: 2022-05-27 | | |
| 来自于: 高强 | | |

**要求:**

使用课上的至少三类（CNN、RNN、GAN、GNN）深度学习方法进行表情识别，着重解决类间不均衡的问题。

## 6、脑肿瘤分割检测

脑肿瘤被认为是儿童和成人中的侵袭性疾病之一。脑肿瘤占所有原发性中枢神经系统（CNS）肿瘤的85%至90%。每年约有11，700人被诊断患有脑瘤。患有癌性脑或中枢神经系统肿瘤的人的5年生存率男性约为34%，女性约为36%。脑肿瘤分为：良性肿瘤、恶性肿瘤、垂体肿瘤等。应实施适当的治疗、计划和准确的诊断，以提高患者的预期寿命。检测脑肿瘤的最佳技术是磁共振成像（MRI）。

该数据集共有约800张脑肿瘤图像，分为TRAIN、VAL、TEST。

| | | |
|---|---|---|
| **文件大小: 19.8 MB** | **分享内容: 第6题数据.rar 链接地址:https://jbox.sjtu.edu.cn/l/91c1Z5** | |
| 到期日: 2022-05-27 | | |
| 来自于: 高强 | | |

**要求:**

将注释文件中脑肿瘤标记坐标所围成的区域视为脑肿瘤区域，使用深度学习模型对脑肿瘤区域进行分割。并在外部验证集TEST上进行验证。

## 7、文本情绪检测

从文本中检测情绪是自然语言处理中具有挑战性的问题之一。原因是标记数据集不可用以及问题的多类性质。人类有各种各样的情绪，很难为每种情绪收集足够的记录。

该数据集有13个不同情感的40000条记录。

| | | |
|---|---|---|
| **文件大小: 1.6 MB** | **分享内容: 第7题数据.zip 链接地址:https://jbox.sjtu.edu.cn/l/M1yPfR** | |
| 到期日: 2022-05-27 | | |
| 来自于: 高强 | | |

**要求:**

设置随机种子，随机划分3:1的训练集和验证集。对语言内容进行编码，进而使用自然语言处理实现多分类任务。如果效果不尽人意，可以考虑合并类别。