第二次大作业

作业任务：

从如下分类问题中选择一个任务完成：
- Fashion-Mnist 商品类别
- Cifar-10 普适物体识别
- 其他数据集（手势、人脸⋯）上的分类或检测或分割等任务

建议使用的神经网络模型：BP、LeNet、AlexNet、VGG-16、VGG-19、ResNet-18、ResNet-50 等；建议使用 Pytorch 等开发平台。由于本次作业需要需要大家自己寻找 GPU/CPU 等计算资源，因此对于电脑算力有限的同学可以适当选择较小的数据集与模型。

作业要求：

- 独立完成。**独立或团队（不超过三人）完成**，严禁互相抄袭；团队要有明确具体的分工（在报告中说明）。
- **任务可选**。即可使用规定的数据集完成规定任务，也可以利用其它数据集完成类似任务。
- **按时提交**。

作业提交：

作业提交内容及形式：
1、PDF 报告一份：含任务介绍、实现方案、实验结果、结果比较分析；（单栏十页以内）
2、Readme 文档一份(markdown)，包括实验结果简介，代码环境，Usage 等即可；
3、源码一份，利用提供的数据集图像，对其进行测试验证。
请在 2022.05.22 之前，将报告和代码打包上传至 Canvas。

附**作业提交要求**：在 Canvas 上传三份文件如下:
　1. 代码文件(zip 格式压缩);
　2. PDF 报告(报告内附有姓名、学号);
　3. Readme 文档(转为 PDF 格式提交)
　命名格式如下: 姓名_学号_代码/报告/Readme

附**评分要求**:
　1. 基础任务实现、实现方案、实验结果等 -10
　2. 实验方案改进，模型优化，参数对比分析等 -10
　3. Readme 文档 -5

附 **Readme 文档参考**:
https://github.com/autonomousvision/giraffe/blob/main/README.md

1. Key Indicators of Heart Disease(★ ★ ★ ★)

According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

As described above, The dataset contains **18 variables (9 booleans, 5 strings and 4 decimals)**. In addition to classical EDA, this dataset can be used to apply a range of machine learning methods, most notably classifier models (**logistic regression, SVM, random forest, etc.)**. You should treat the variable "HeartDisease" as a binary ("Yes" - respondent had heart disease; "No" - respondent had no heart disease). But note that **classes are not balanced**, so the classic model application approach is not advisable. Fixing the weights/undersampling should yield significantly betters results. Can you indicate which variables have a significant effect on the likelihood of heart disease?

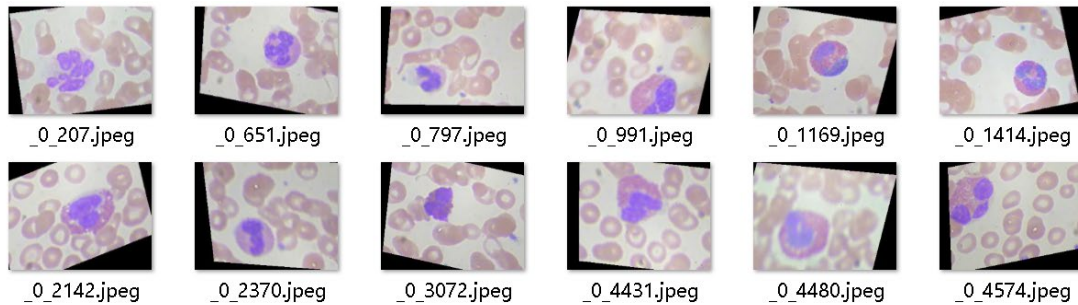| ✓ HeartDisease	# BMI	✓ Smoking	✓ AlcoholDrinking	✓ Stroke	# PhysicalHealth	# MentalHealth
Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	Body Mass Index (BMI)	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week	(Ever told) (you had) a stroke?	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30	Thinking about your mental health, for how many days during the past 30 days was your mental health not good?
true 27.4k 9% / false 292k 91%	12 ... 94.8	true 132k 41% / false 188k 59%	true 21.8k 7% / false 298k 93%	true 12.1k 4% / false 308k 96%	0 ... 30	0 ... 30
No	16.6	Yes	No	No	3.0	30.0
No	20.34	No	No	Yes	0.0	0.0
No	26.58	Yes	No	No	20.0	30.0
No	24.21	No	No	No	0.0	0.0

Download link: https://jbox.sjtu.edu.cn/l/i1XdmV
Password: xuqa

2. Blood Cell Images(★ ★ ★ ★)

The diagnosis of blood-based diseases often involves identifying and characterizing patient blood samples. Automated methods to detect and classify blood cell subtypes have important medical applications.

This dataset contains **10,000 augmented images of blood cells** (JPEG) with accompanying cell type labels (CSV). There are approximately **2,500 images for each of 4 different cell types** grouped into 4 different folders (according to cell type). The cell types are Eosinophil, Lymphocyte, Monocyte, and Neutrophil.

| _0_207.jpeg | _0_651.jpeg | _0_797.jpeg | _0_991.jpeg | _0_1169.jpeg | _0_1414.jpeg |
| _0_2142.jpeg | _0_2370.jpeg | _0_3072.jpeg | _0_4431.jpeg | _0_4480.jpeg | _0_4574.jpeg |

Download link: https://jbox.sjtu.edu.cn/l/M18J5w
Password: nwfj

3. Animals-10(★ ★ ★ ★ ★)

The Animals-10 datasets contains Animal pictures of 10 different categories taken from google images. It contains about 28K medium quality animal images belonging to 10 categories: dog, cat, horse, spyder, butterfly, chicken, sheep, cow, squirrel, elephant.

All the images have been collected from "google images" and have been checked by human. There is some erroneous data to simulate real conditions (eg. images taken by users of your app).

The main directory is divided into folders, one for each category. Image count for each category varies from 2K to 5 K units.

Download link: https://jbox.sjtu.edu.cn/l/X1L5Rp
Password: hwqw

4. Garbage Classification(★ ★ ★)

Garbage is the most common thing in our daily life. And how to efficiently separate and recycle waste is also one of the issues that human society has been concerned about.

The Garbage Classification Dataset contains **6 classifications**: **cardboard (393), glass (491), metal (400), paper(584), plastic (472) and trash(127).** The number indicates the number of images per category.

Download link: https://jbox.sjtu.edu.cn/l/e1ujEu

Password: dbbx

5. Fingers(★ ★ ★ ★)

The goal of the project is to build a model able to count fingers as well as distinguish between left and right hand. The dataset contains **21600 images** of left and right hands fingers, including **Training set: 18000 images, Test set: 3600 images**, with the same size, **128 by 128 pixels**. All of the images are centered by the center of mass, Noise pattern on the background.

Note that labels are in 2 last characters of a file name. L/R indicates **left/right hand**; 0,1,2,3,4,5 indicates **number of fingers,** and images of a left hand were generated by flipping images of right hand.
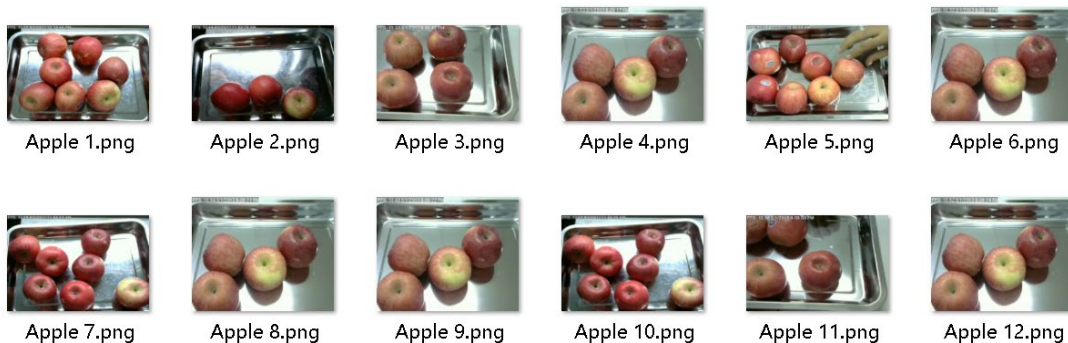

Download link: https://jbox.sjtu.edu.cn/l/M18nXy

Password: bkab

6. Fruit Recognition(★ ★ ★ ★)

The database used in this study is comprising of **44406 fruit images**, with **15 different categories**, which we collected in a period of 6 months. The images where made with in our lab's environment under different scenarios which we mention below. We captured all the images on a clear background **with resolution of 320×258 pixels**. We used HD Logitech web camera to took the pictures. During collecting this database, we created all kind of challenges, which, we have to face in real-world recognition scenarios in supermarket and fruit shops such as light, shadow, sunshine, pose variation, to make our model robust for, it might be necessary to cope with illumination variation, camera capturing artifacts, specular reflection shading and shadows. We tested our model's robustness in all scenarios and it perform quit well.

All of images were stored in RGB color-space at 8 bits per channel. The images were gathered at various day times of the day and in different days for the same category. These features increase the dataset variability and represent more realistic scenario. The Images had large variation in quality and lighting. Illumination is one of those variations in imagery. In fact, illumination can make two images of same fruit less similar than two images of different kind of fruits. We were used our own intelligent weight machine and camera to captured all images. The fruit dataset was collected under relatively unconstrained conditions. There are also images with the room light on and room lights off, moved the camera and intelligent weight machine
near to the windows of our lab than open windows, closed windows, open window curtains, closed curtains. For a real application in a supermarket, it might be necessary to cope with illumination variation, camera capturing artifacts, specular reflection shading and shadows. Below are the few conditions which we were considered during collected dataset.
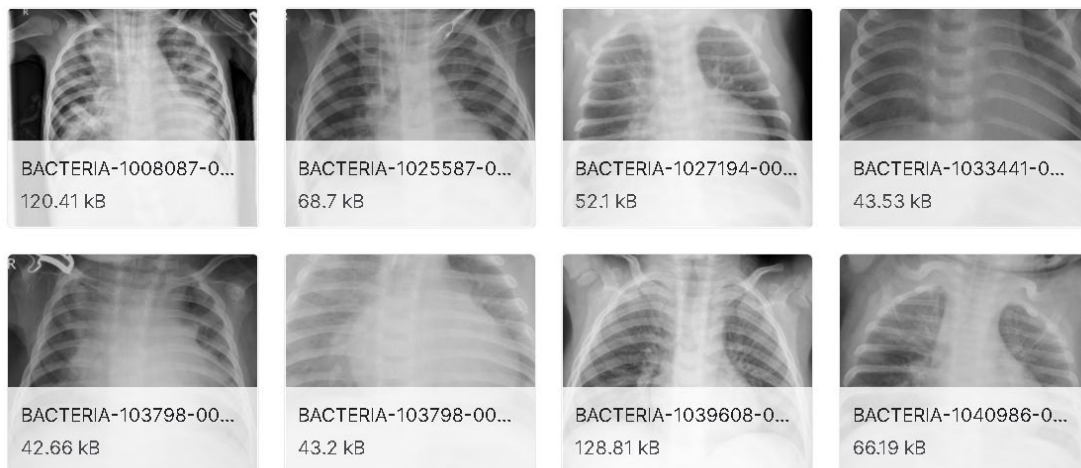

| Apple 1.png | Apple 2.png | Apple 3.png | Apple 4.png | Apple 5.png | Apple 6.png |

| Apple 7.png | Apple 8.png | Apple 9.png | Apple 10.png | Apple 11.png | Apple 12.png |

Download link: https://jbox.sjtu.edu.cn/l/31XP9c
Password: fayq

7. **Chest X-ray Images(★ ★ ★ ★)**

Pneumonia is an infection that inflames the air sacs in one or both lungs. It kills more children younger than 5 years old each year than any other infectious disease, such as HIV infection, malaria, or tuberculosis. Diagnosis is often based on symptoms and physical examination. Chest X-rays may help confirm the diagnosis.

This dataset contains **5,856 validated Chest X-Ray images**. The images are split into a training set and a testing set of independent patients. Images are labeled as (disease:NORMAL/BACTERIA/VIRUS)-(randomized patient ID)-(image number of a patient). For details of the data collection and description, see the referenced paper below. According to the paper, the images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou.

Download link: https://jbox.sjtu.edu.cn/l/i1XxKB
Password: wkbk

8. Other common datasets (★ ★ ★)

Fashion-Mnist	**Classification**	Correctly identify fashion items from a dataset of tens of thousands of grayscale images.
CIFAR-10	**Classification**	classify color images into one of ten classes: airplane, automobile, bird, cat deer, dog, frog, horse, ship, and truck
Caltech101	**Classification**	Pictures of objects belonging to 101 categories. About 40 to 800 images per category.
ImageNet	Classification, Classification with localization	The dataset consist of 150,000 photographs, with 1000 categories collected from flickr and other search engines. Due to memory limitations, We recommend you to use a subset of ImageNet.

Fashion-Mnist 数据集：https://www.kaggle.com/datasets/zalando-research/fashionmnist

Cifar-10 数据集：　http://www.cs.toronto.edu/~kriz/cifar.html