**THỰC HÀNH 5: PHÂN LỚP ẢNH**

**Mục tiêu**: Xây dựng mô hình phân loại cho bài toán phân lớp ảnh.

**Bài toán**: Nhận diện bệnh nhiễm khuẩn phổi từ ảnh X-Quang.

**Input**: Một bức ảnh đầu vào.

**Output**: Một trong 2 nhãn: NORMAL - bình thường và PNEUMONIA - bệnh.

Bộ dữ liệu: **Chest X-ray dataset**.

Công bố khoa học: Kermany, Daniel S., et al. "**Identifying medical diagnoses and treatable diseases by image-based deep learning.**" Cell 172.5 (2018): 1122-1131.

Link: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

**Các bước thực hiện tổng quát:**

Đọc dữ liệu: Các file được tổ chức thành từng bức ảnh có định dạng là .jpeg. Các bức ảnh thuộc nhãn NORMAL hoặc PNEUMONIA sẽ được gom chung thành 1 thư mục.

Code tham khảo đọc dữ liệu:

import glob

import numpy as np

import cv2

IMG\_SIZE = 227

def load\_dataset(path):

X = np.array([])

y = np.array([])

classes = ['NORMAL', 'PNEUMONIA']

for c in classes:

files = glob.glob(path + c + "/\*.jpeg")

for f in files:

print(f)

img = cv2.imread(f)

img = cv2.resize(img, (IMG\_SIZE, IMG\_SIZE))

if X.size == 0:

X = np.array([img])

else:

X = np.vstack([X,[img]])

y = np.append(y, c)

assert(X.size > 0), 'Cannot read file'

return (X,y)

Để load dữ liệu từ tập train, ta sử dụng lệnh:

X\_train, y\_train = load\_dataset("<đường dẫn tới dataset>")

Thực hiện tương tự để đọc dữ liệu từ tập dev và test nhé.

**BÀI TẬP CƠ BẢN**

**Bài 1**: Hãy đọc dữ liệu, và cho biết các thông tin sau:

a) Mục tiêu / tác vụ mà bộ dữ liệu hướng tới là gì?

b) Kích thước của dữ liệu là bao nhiêu trên các tập train, dev và test?

c) Phân bố nhãn theo từng tác vụ trên các tập train, dev và test. Vẽ biểu đồ barplot thể hiện phân bố nhãn trên từng tập dữ liệu.

*Gợi ý*: Sử dụng hàm *barplot()* trong thư viện **seaborn**.

d) Hãy hiển thị khoảng 3 bức hình về ảnh X-Quang phổi lên màn hình.

*Gợi ý*: dùng hàm *imshow()* trong thư viện **matplotlib**.

**Bài 2**: Huấn luyện mô hình bằng phương pháp Logistic Regression.

**Bài 3**: Huấn luyện mô hình bằng phương pháp SVM.

**Bài 4**: Hãy so sánh kết quả thực hiện giữa 2 phương pháp phân lớp trên. Phương pháp nào cho kết quả tốt hơn.

**BÀI TẬP NÂNG CAO**

*Hãy thử xây dựng một ứng dụng đơn giản, với input là một tấm ảnh X-quang, và output ra là nhãn dự của tấm ảnh đó.*