How to run the code:

- **01 - Data Preparation and Split**
  - This notebook should be run on Colab
  - It takes the raw dataset from Kaggle, removes the corrupted image files and creates the dataset split into 30 classes by their age cohort and gender
- **02 - Data Preprocessing**
  - This notebook should run on Colab
  - It takes the prepared dataset and runs preprocessing on it, and was also used to test other preprocessing processes with a CNN model
- **03 - CNN Models**
  - This notebook has all the relevant links to be run in Colab, but the dual_model will probably cause a RAM error as it requires more RAM than Colab offers
  - If you download the data, make sure to update the path to the dataset accordingly
  - When run, this notebook performs the training of both CNN-Models, starting with the single output model. This model also is the source for the feature extractor used in the later notebooks
- **04 - Cross Validation Dataset**
  - The first half of the notebook needs to be run locally, as the zip file that needs to be downloaded from kaggle is too big to store on google drive
    - The CSV file needs to be downloaded from NIH, as it is an updated version
    - Alternatively, Kaggle API can be used to download it in Colab directly
  - The data needs to be downloaded beforehand and the first half needs to be run
  - The preprocessing can then be run in Colab once a sample from the dataset is taken
  - Alternatively, this can be skipped by using the preprocessed data for the validation data that can be found in a link below
- **05 - Feature Extraction**
  - Your inputs will be the preprocessed training data, preprocessed cross validation data, and the feature extractor model
  - The output will be the features and labels of the datasets
- **06 - SVM Classification**
  - Your inputs will be the features extracted from the data and their labels (the contents of the Features and Labels folder listed below)
  - Your outputs will be the trained SVM models, their classification results, and various graphs and images used in the final report
- **07 - Evaluate_CNNs**
  - contains code to evaluate the trained models on the validation split of our dataset
  - creates the graphics including the Heatmaps and confusion matrices shown in the report for the CNN models

Raw original data

Grouped by Age and Gender (144 Classes):

https://drive.google.com/file/d/14zgM09m1pKELT2aX3JPUhGtI2MMkN3cK/view?usp=share_link

Bash file to create dataset:

https://drive.google.com/file/d/1EcMuDdsfpoEQypps03EUcBpceWhN4eQ8/view?usp=sharing

Grouped by Age and Gender in 5-Year cohorts:

https://drive.google.com/file/d/1XLVOQ5wF-7LRjxYa5vDfWN6uZN23s4WT/view?usp=share_link

Cross-validation dataset:

https://www.kaggle.com/datasets/nih-chest-xrays/data

https://nihcc.app.box.com/v/ChestXray-NIHCC/file/219760887468 (csv file)

Data after preprocessing

Training dataset:

https://drive.google.com/uc?id=1qBnv1rc2IJlTnFkE4iLecgnIS6-aEOKA

https://drive.google.com/uc?id=1u1bYdXnffUfAM4UeAJIRoH_6Q7C6n0xQ

Cross-validation dataset:

https://drive.google.com/uc?id=1FL6wP5e-BP9MrqLVCT680T9BW1sk3sZ9

Models

CNN Model definitions:

https://drive.google.com/file/d/1ByZhfo-tKMDJDuKvq6qMmO0o-718PqSw/view

Feature extractor:

https://drive.google.com/file/d/1yd3HDakuh_ckFGmzUa_SFAmkKgsa0gkl/view?usp=share_link

Single CNN Model:

https://drive.google.com/file/d/16VUeHJLQAsx-e6cpqbi26HKAwEqFjXWX/view?usp=sharing

Dual CNN Model:

https://drive.google.com/file/d/1nYVUmkYSxivpbRe4qEtIsBYVOI1-jkPH/view?usp=sharing

Models zip file:

https://drive.google.com/file/d/13iLdc7uyaYVvZbWb5viTKkSmebKx2-B5/view?usp=sharing

Features and labels of the data

https://drive.google.com/drive/folders/1Tr7QVX5b3cgZ2IUt-e9yBmJy5zv4QcHD?usp=share_link

Additional files to run the code

https://drive.google.com/file/d/1zt4EeZomGOYiD42NE5nYKzFjYIDQSgle/view?usp=sharing