Nicholas Kellogg

DS4002 - Rubric

4/28/2025

Purpose: The purpose of this assignment is to give you hands-on experience training a machine learning model to tackle a real-world health crisis. You will apply skills in image processing, exploratory data analysis (EDA), and convolutional neural networks (CNNs) to distinguish between COVID-19 infected lung x-ray images, Pneumonia-infected images, and Normal images. You'll also practice communicating technical results in a professional format.

Tasks:

- Explore the lung x-ray dataset and summarize key findings through data visualization and basic statistics.
- Preprocess the images appropriately (resizing, scaling, augmentation if necessary).
- Build and train a convolutional neural network (CNN) to classify images into three categories: COVID-19, Pneumonia, Normal.
- Evaluate your model using appropriate metrics (accuracy, confusion matrix, ROC-AUC scores, etc.).
- Summarize your work clearly in a professional report that includes:
    - Overview of methods used
    - Key EDA findings
    - Model architecture and training decisions
    - Model performance and discussion of results

Tips for Success:

- Learn how Central Neural Network models absorb and analyze data, especially when picking a training / test split
- Make sure all of the input images are the same size, clear, feature more or less the same brightness, etc.
- Do background research on how doctors normally identify and diagnose COVID-19, especially in how those differentiate from Normal and Pneumonic scans

How Will I Know if I Succeed?

- Set a quantifiable goal for your model to reach in terms of classification, such as above 80%. If you're having trouble reaching this, try to increase the epochs, change the train/test split, or increase the depth of the model.

Rubric:

Exploratory Data Analysis	**Goal:** Clear visualizations, thoughtful statistics, identifies trends and potential issues. - Focus on key traits such as image brightness, pixel distribution, etc. - Include all references and scripts - Include all outputs, specifically in clear and well thought out graphs. - Execute the eda_project3.py file
Data Preparation and Cleaning	**Goal:** successfully execute scripts needed in order to resize, brighten, and clear images in the COVID19_Images dataset. - Simply run the eda_project3.py file, as it will do this automatically
Model Design and Training + Results	**Goal**: successfully execute the COVIDModelTraining.py file in order to achieve an output of 95% correct analysis. - Use the data output visualization to accurately describe how the model works, including its strengths and weaknesses - Provide clean visualizations and concise, yet meaningful, results that explain the outputs - Discuss the overall performance and next steps or improvements
Repository and Submission	**Goal:** submit a link to Github that contains a repository featuring the following elements: - All data, including precleaned and postcleaned - EDA analysis outputs and scripts - Folders sorting the data into outputs, scripts, and data - Ex: https://github.com/kellogg9/DS4002-project3 - Creation of a README file and an MIT License for your reproduction - The final model script, as well as visualization outputs (such as COVIDModelTraining.py)