Detection of Disease Using Deep Learning, Transfer Learning, & AutoML

The early detection of disease is a crucial step in the medical diagnosis/treatment process. Developing rigorous, reliable, and scalable solutions to this problem has proven to be a difficult task. Machine learning provides a powerful toolbox that can be used to solve these problems with a high degree of accuracy. In this project, I will explore different models for image classification of various diseases.

Initial Plan:
As a starting point, I will use the Kaggle Malaria Images Dataset (https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria). I will build a neural network from scratch to classify whether the cell in the image has malaria or not. Additionally, I note that there is already a model available which has been trained on this dataset (Rajaraman et al. 2018 PeerJ). I will compare my model to this model, as well as use another malaria images dataset (Quinn et al. 2016 arXiv) as a test dataset/to further refine the model.

Building an accurate ML model is a difficult process. There is a nonobvious workflow involved in creating the most appropriate model. One may spend much time in developing the ML model. Automating the process of choosing an appropriate model, then, is an important task which can allow one to focus more on solving problems rather than model building. Tools such as autosklearn and autokeras help automate the model selection process. I will apply these tools to provide alternative models for classification of malaria.

Training a separate model on each dataset without being able to transfer the model's learned knowledge to a new problem would be a very ineffective use of ML. Hence, I will utilize the models trained on the malaria dataset on both a smaller dataset (https://www.kaggle.com/paultimothymooney/breast-histopathology-images) and a larger dataset (https://datasets.simula.no/kvasir/#dataset-details) to see how well the knowledge transfers for binary classification.

Finally, time permitting, I will use whichever approach proved to be the most powerful in the previous work to perform multi-class classification to identify the pathology of datasets related to GI tract health & colorectal caner (https://datasets.simula.no/kvasir/#dataset-details, https://zenodo.org/record/1214456#.XQbpPIhKhPY).

In general, this problem is a supervised learning problem. As exploratory analysis, I intend to perform clustering to see if the classes can be easily identified and grouped without any training. All data identified in this project have been labeled, so I will use the labels of the data as my targets, with the vectorized images as my predictors.

My final deliverable will be deployed as a web service with an API.

My estimate for the resources required for this project are a CPU with 4 cores (8 threads) and 16GB of RAM. I will need approximately 100GB for storage of all relevant data.