Cancer Detection in Histopathological Slides

Nitin Mahajan

Milestone Report 1 (EDA)

OVERVIEW

Cancer is a deadly disease. Researchers and clinicians are trying to find the methods to detect it at early stages. Early diagnosis will play an important role in planning the treatment plan and improvement of the patient's survival rate. Cancer can be benign (localized) or metastatic (spread to distant organs). One of the most important early diagnosis is detection in lymph nodes to find out whether the cancer has metastasized or not. The method to do this is H & E staining of histological slides of lymph nodes taken from biopsies.

GOAL

Currently pathologists manually examine the slides in the laboratory and decide if the patient has metastatic cancer or not. Reading the slides and making a report based on human judgement which can be inconsistent and vary from person to person and from day to day. Therefore, developing a computation model to read the slides would provide and can automate the process to give unbiased results.

DATA SOURCE

The data for this project are downloaded from Kaggle website
https://www.kaggle.com/c/histopathologic-cancer-detection/data

WHAT WE ARE LOOKIING AT (IMAGES)

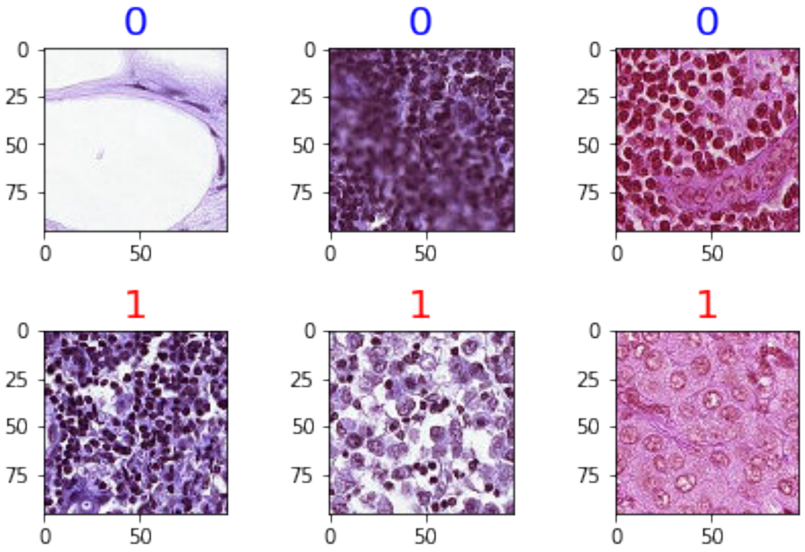
The histopathological images are glass slide microscope images of lymph nodes that are stained with hematoxylin and eosin (H&E). This staining method is one of the most widely used in medical diagnosis and it produces blue, violet and red colors. Dark blue hematoxylin binds to negatively charged substances such as nucleic acids and pink eosin to positively charged substances like amino-acid side chains (most proteins). Typically, nuclei are stained blue, whereas cytoplasm and extracellular parts in various shades of pink.

Lymph nodes are small glands that filter the fluid in the lymphatic system and they are the first place a breast cancer is likely to spread. Histological assessment of lymph node metastases is part of determining the stage of breast cancer. The diagnostic procedure for pathologists is tedious and time-consuming as a large area of tissue has to be examined and small metastases can be easily missed.

EXPLORATORY DATA ANALYSIS

Following data are provided:-

1. Sample_submission.csv - a sample submission file
2. Train_labels.csv – A CSV file with labels of 0 or 1 (0 for cancer not detected and 1 for cancer detected) for corresponding images in training dataset.
3. Train – A directory with 220,025 images. TRAINING DATA SET
4. Test - A directory with 57,458 images. TEST DATA SET
5. The training data set have 130908 and 89117 images with '0' and '1' label, suggesting the data is imbalanced.
6. Below are the representative images for normal and cancer



DATA WRANGLING & SAMPLING OF IMAGES

Since the dataset is very large and neural networks take very long to train on all the datasets, I decided to sample 10,000 images in each class (a total of 20,000 images) to make the dataset balanced and smaller yet containing enough images to train my models. Once sampling is completed, move the images to separate folders to be consistent in training different models' multiple times. As the dataset has already splitted the training and the test data set, there is no need to split data in train and test. However, I will split the training data in validation and training data sets in a ratio of 20/80. data

MACHINE & RESOURCES FOR BUILDING CONVOLUTIONAL NEURAL NETWORK MODELS

We need GPUs for sure to handle the image data and building the CNN models. As motioned earlier, I don't have GPUs and so I subsampled the data set. I will be building the CNN models on my computer which has following configuration.

Laptop: Apple MacBook Pro

Processor: 2.6GHz 6-Core Intel Core i7
Memory: 16GB 2400 MHz DDR4
Modules and libraries: Python and Keras with TensorFlow in backend will be the main language and library
Reading Material: Deep Learning with Python (Francois Chaollet), Udemy Course – AZ Machine Learning and other online resources.