Breast Cancer Classification using Computer Vision

Chandramouli Yalamanchili
DSC680 - T302 Applied Data Science (2215-1)
https://chandu85.github.io/data-science

Proposal Document - https://github.com/chandu85/data-science/blob/main/Project%202%20-%20Breast%20Cancer%20Classification%20using%20Computer%20Vision/Documentation/Project%202_Proposal.pdf

Project details

Project Domain:

I chose to work on image classification or computer vision as my 2nd project, especially how it is currently being used in medical domain. For this project I chose the use case of detecting the breast cancer using CNN model.

Project Abstract:

Breast cancer is one of the most common types of cancer in American women, it is estimated that in the year of 2020, approximately 30% of the new cancer diagnosed women will be breast cancer. Of all the breast cancer, the Invasive Ductal Carcinoma (IDC) is the most common subtype. In the year of 2020, IDC subtype accounted for 85% of total breast cancer cases.

To assign an aggressiveness grade to a whole mount sample, pathologists typically focus on the regions which contain the IDC. As a result, one of the common pre-processing steps for automatic aggressiveness grading is to delineate the exact regions of IDC inside of a whole mount slide. Using the automated process to evaluate each of the patches or the mount samples would be great help in saving time and increasing the accuracy of the diagnosis.

Through this project we will build a Keras based CNN (Convolutional Neural Network) classifier model that would evaluate the patches collected from several whole mount slide images and accurately classify a histology image as benign or malignant.

Week 2 check-in

Any surprises from your domain from these data?

- I have not come across any challenges or surprises with respect to the domain.
- I am still learning more about both the health care domain, breast cancer in specific and the computer vision domain.
- It seems like there is a lot of progress made in both of these domains in last couple of decades. I am trying to review several referrals I have gathered on these topics.
- I am hoping to have my literature review complete by end of next week before I create my project report

The dataset is what you thought it was?

- Yes, data was what I thought it would be, it is pretty big in size due to the number of images it has.
- It has information for 280 patients and the images are split into two folders for each patient, depending on classification of the cell.
- Based on the quick EDA I have performed at this time, it seems like observations for Benign class are higher than the observations for malignant class.
- It took quite a bit in formatting the images to make sure they are all of same size before I use them in modeling.
- So far the dataset is as I expected it to be, but hasn't presented any major concerns in processing till now.

Have you had to adjust your approach or research questions?

- Not really, at least at this time I am still going ahead as per my plan at this time.
- I have only completed EDA at this time on the data, trying to understand the dataset and the images.
- Next week I will be finishing up the EDA and work on modeling.
- But, at this time I am still going ahead with the same plan and with the same research questions I started with.

Is your method working?

- Yes, I would say my initial plan is still working at this time.
- I have planned to use Python notebook and haven't had any issues with it so far. Everything has been going as per the initial plan.
- I am still planning to build CNN model using Keras for this classification problem.

What challenges are you having?

- I haven't run into any challenges yet either with technology or with data. Everything has been going pretty smooth so far for this project.
- I am still worried about modeling part considering the size of the dataset I am working with.
- EDA process has been smooth so far and I have not run into any issues.
- Based on the analysis completed so far, observations for the classes doesn't seem to be balanced, so I might use SMOTE technique to get the balanced training samples even in this project.
- So, at this time everything is going smooth for me, but next week will be interesting as work on classification model.